Multi-stage Integer Robust Optimization For Bluebike Allocation

Peijun Xu
Massachusetts Institute of Technology
Operations Research Center
MA, USA
xup@mit.edu

Jiani He
Massachusetts Institute of Technology
Operations Research Center
MA, USA
jianihe@mit.edu

August 29, 2025

Contents

1 Problem Overview

Bluebike is one of the leading bike-sharing firms in Boston, with over 3,000+ bikes and more than 300+ stations across Boston. However, it might be challenging for the bike service company to serve customers due to the imbalanced distribution of bikes among different stations and uncertainty in demand. It is common to have surplus bikes placed in one station and scarce bikes in another nearby station. Therefore, we want to propose a method to design a reliable Bluebike distribution system, mainly in Cambridge. Also, because we are optimizing the number of bikes transferred, the nature of the problem involves integer decision variables. We touched on integer robust optimization in the class but did not discuss extensively how to deal with robust optimization problems with integer decision variables. We want to take this opportunity to explore efficient methods to tackle a multi-stage integer robust optimization problem and see if these techniques achieve similar effects as their applications on non-integer problems discussed in the class. Based on the empirical results, we would recommend the best model to solve Integer Robust Optimization, considering the trade-off between the computation time and total cost.

To save on computation time and achieve a realistic possible route, all 7 Cambridge stations, all of which are the top 10 most popular stations, are chosen as the region of interest, using actual data from the Bluebike system in 2018-2019. We further aggregate the trip data into a daily level and formulate the problem with daily-level decision variables. For example, although Bluebike staff transfers the bikes from station i to station j several times a day in real life, we assume that they would only make the transfer once a day.

2 Data

2.1 Data Source

The data we use for modeling comes from the official Bluebike website, which provides the following data:

- Trip History Data: We use it to predict the number of rides starting from the station (demand) and the number of rides stopping at the station (supply). The uncertainty in (demand-supply) for each station is estimated based on the error made by the prediction model on the historical data.
- Station Data: It provides the number of docks at each station. We also estimate the transfer cost and the correlated uncertainty based on the distances between stations calculated using the coordinates. The 7 Cambridge stations we pick are: "Ames St at Main St", "Central Square at Mass Ave / Essex St", "Kendall T", "MIT Pacific St at Purrington St", "MIT Stata Center at Vassar St / Main St", "MIT Vassar St", "MIT at Mass Ave / Amherst St"
- Membership and Ridership: We estimate the opportunity cost of unmet demand based on how Bluebike charges the customers.
- Hiring information: We estimate the cost of "emergent transfer" if the number of bikes at the station exceeds its capacity or approaches 0 based on the salary offered to the staff.

2.2 Cost Estimation

2.2.1 Transfer Cost

Bluebike hires bike technicians to move the bikes between stations. The hourly salary for a bike technician is \$12; we assume he can move 30 bikes at one time from station i to station j. We also assume that he carries bikes from station i to station j only twice an hour. Therefore the average cost to move one bike through bike technicians would be 12/(30*2)=\$0.2 per bike.

There are three types of transfer, and each incurs a different cost.

- Internal Transfer ($Cost\_I_{i,j}$): Transfer between Cambridge stations. The cost will vary by the distance of the transportation
- External Transfer($Cost\_O_i$): Transfer between one Cambridge station and a station outside Cambridge. We set the external transfer cost as a fixed value, 2 times the max internal transfer cost.

- Emergent Transfer: The emergent transfer is not a part of the decision variables; it occurs only when the number of bikes at stations reaches the full capacity. We set the cost of emergent transfer higher than the cost of external transfer.

2.2.2 Mismanagement Cost

Mismanagement cost occurs in two situations.

- $Unmet\_Cost$: If the decisions made are unable to meet all the demand, there would be the opportunity cost and the cost of the emergent transfer.
- Sur_Cost: If the decisions made cause the number of bikes to exceed the station's capacity, there would be a cost of the emergent transfer.

2.3 Demand Estimation

While demand for bikes for stations depends on factors such as the time of day and weather, holiday effects, our focus is not on the demand prediction. Therefore, to predict a daily net change(demand-supply) from station i to station j at time t, we will build a time-series prediction model through the Auto-Regressive Integrated Moving Average(ARIMA) with cross validation to select the hyperparameters and set the seasonality = 7 to capture the weekly seasonality. We use the daily net change of each Cambridge station in 2019/01/01 - 2019/09/30 to predict the daily net change of each Cambridge station in 2019/10/01 - 2019/12/31.

2.4 Uncertainty

We adapt the formulation in HW4 to generate the P matrix based on the ARIMA model's performance on 2018's data and the distance between each station and use budget uncertainty set for z.

$$P_{i,j} = \begin{cases} \frac{rmse_i}{2}, & \text{if } i = j \\ e^{-\frac{dist_{i,j}}{R_D}} \times \frac{rmse_i + rmse_j}{24}, & \text{if dist}_{i,j} \leq \text{R}_D\text{, i} \neq j \\ 0, & \text{otherwise} \end{cases}$$

$$Z = \begin{cases} |z|_\infty \leq \rho \\ |z|_1 \leq \Gamma \end{cases}$$

We chose $\rho = 1$. To protect against 95% uncertainty, $\Gamma = \sqrt{2ln(1/\epsilon) * L} = \sqrt{2ln(20) * 7} = 6.476$

2.5 Additional Declaration

The primary purpose of this project is not to give exact recommendations on management decisions but to offer a formulation to solve the problem with a similar structure so that the issues can be solved efficiently once better data are obtained. Therefore, our estimation aims to capture the relative influence in real life rather than give an accurate estimation of the actual data.

3 Models

3.1 Formulations

*Detailed description of each variable is included in the appendix.

$$\begin{array}{lrl} \min_{X,Y} & \sum_{i,j,t} X_{i,j,t} Cost\_I_{i,j} + \sum_{i,t} |Y_{i,t}| Cost\_O_i & \\ \text{s.t.} & b_{i,t} \leq b_{i,t-1} + trans_{i,t-1} - (d_{i,t-1} + (Pz)_{i,t-1}) & \forall t, i \neq 1 \forall z \in Z \\ & trans_{i,t} = Y_{i,t} + \sum_j X_{j,i,t} - \sum_j X_{i,j,t} & \forall i, t \\ & b_{i,t} \geq 0 & \forall i, t \\ & b_{i,t} \leq C_i & \forall i, t \\ & b_{i,1} = 0 & \forall i \\ & X_{i,j,t} \geq 0 & \forall i, j, t \\ & Z = \{z | \, |z|_\infty \leq \rho, |z|_1 \leq \Gamma\} & \end{array} \tag{1}$$

The objective is to minimize the total transportation cost to meet all predicted demands under uncertainty through technicians' internal and external transfers with certain constraints.

There are several constraints that we will consider:

- The bikes available at station i at time t, $b_{i,t}$, should be no bigger than the bikes available at station i at time t-1, $b_{i,t-1}$, $-$ (departed bikes - arrived bikes) at station i at t-1, $d_{i,t-1}$, $-$ uncertainty at station i at t-1, $Pz_{i,t-1}$, $+$ net transfer at station i during time t-1, $trans_{i,t-1}$
- The number of available bicycles at station i at time t, $B_{i,t}$, is no smaller than 0 and no larger than the total number of docks at station i, C_i.
- The uncertain variable z is in the budget uncertainty set with bounded support [-1,1] and guarantee the solution to be feasible for 95% of uncertain outcomes.

3.2 Strategies

As mentioned at the beginning, the nature of this problem is a multi-stage integer optimization problem, with continuous parameters like P, Γ and ρ in the constraints. Therefore, it is computationally expensive to solve the vanilla robust optimization problems, especially if we want to expand the scope to the entire Boston area (400+ stations) and solve the problem on a hourly basis. To make the problem more tractable, we consider the following modeling strategies:

- Shrinking Horizon: Instead of solving the problem for the entire time span, we shrink the problem size by optimizing 7 days ahead and resolve the problem in a rolling basis.
- Network Flow: We round up the continuous parameters in the constraints so that the problem becomes network flow optimization, which can be solved much faster.

Through experiments, we would like to compare the effectiveness of these strategies when combined with nominal model and robust model.

4 Results

We measure the models' performance by its **total cost**, which is the sum of **transfer cost** and **mismanagement cost**. **Transfer cost** captures how the model efficiently allocates the bikes while **mismanagement cost** indicates the models' ability to incorporate uncertainty.

Table 1: Key Performance Metrics (*gap to optimality $\leq 1\%$)

FH	Horizon	Model	Formulation	Trans Cost	Mismgnt Cost	Total Cost	Time
Yes	Full Horizon	Nominal	Integer Opt	1616.50	85320.0	86936.50	5.62
		Robust	Integer Opt	6189.59	80410.0	86599.59	340.19*
		Robust	Network Flow	5805.03	**72850.0**	**78655.03**	7.81
	7 days ahead	Nominal	Integer Opt	2058.09	113860.0	115918.09	2.07
		Robust	Integer Opt	6936.70	100640.0	107576.70	21.9*
		Robust	Network Flow	5650.90	105230.0	110880.90	0.85
No	Full Horizon	Nominal	Integer Opt	1601.34	90440.0	92041.59	**0.07**
		Robust	Integer Opt	6331.21	100990.0	107321.21	14.71*
		Robust	Network Flow	5651.75	95880.0	101531.75	0.10
	7 days ahead	Nominal	Integer Opt	**1540.59**	116010.0	117550.59	0.77
		Robust	Integer Opt	5404.50	122670.0	128074.50	3.42*
		Robust	Network Flow	4851.07	122330.0	127181.07	0.13

Note:
1. The reason that mismanagement cost is much higher than transfer cost is that we set mismanagement cost much higher than transfer cost so that the model would try to avoid as much mismanagement as possible. However, due to the simplified forecasting, decision process and uncertainty estimation, mismanagement is unavoidable, leading to high mismanagement cost. Although our result may be very different from the real value, its purpose is to compare the effectiveness of different model formulations.
2. The computation time is the total time spent solving the problems in all the stages. Although we only

need to solve one stage per day in real-world application, we think it's worth comparing the results if we need to optimize the problem online with a timestep in minutes or even seconds.

5 Conclusions and Future Work

- Conclusions
  - We can see that shrinking the horizon and applying network flow formulation reduce the Computation time. However, applying shrinking horizon would worsen a model's performance since the model is unable to consider the uncertainty in the far future with a short 7-day horizon. The network flow formulation doesn't seem to detriment the model's performance and because it is able to solve to optimality, it yields the best empirical results (minimal transfer cost plus mismanagement cost).
  - Considering the trade-off between the performance and computation time, we think the best strategy to solve the multi-stage integer optimization problem is the following:
    1. Apply Folding Horizon.
    2. Formulate the problem as network flow.
    3. Choose the appropriate horizon to make the problem tractable while considering enough uncertainty in the future.
  - From the empirical results, we find the nominal solution always results in lower transfer cost than robust solution but the mismanagement cost of robust solution is usually higher. This result is in consistency with the idea of robust optimization which is that by sacrificing some optimality, the solution is protected against a lot of uncertainty. In this problem, the robust solution is able to satisfy more demand, leading to less opportunity cost.
- Future Work
  - Apply the cutting plane method to get the near-optimal solutions for the current intractable methods such as solving network flow problem using folding horizon.
  - Build a better prediction model by considering the weather effect and holiday effect through Prophet Model. We could also consider lasso regularization for robustness.
  - Improve our optimization model by specifying the different opportunity costs for the subscribers (annual or monthly members) versus the causal customers (single trip or day pass user) since we have the user type for each ride in the Trip History Data.
  - Expand the scope to the entire Bluebike stations in Boston area and solve the problem with hourly intervals.

6 Acknowledgements

We are grateful to Dimitris Bertsimas and Dick Den Hertog for their inspirations. We also would like to express our special thanks of gratitude to Berk Ozturk and Adam Cheol Woo Kim for their critical comments on the project.

7 Appendix

7.1 Descriptions

7.1.1 Parameters

We first defined the following parameters to incorporate into our model $\forall i, j \in \{1, ..., 7\}, \forall t \in \{1, ..., 92\}$.

- $Cost\_O_i$: the cost to externally transfer one bike between one Boston station and one Cambridge station i
- $Cost\_I_{i,j}$: the cost to internally transfer one bike from Cambridge station i to Cambridge station j
- $dist_{i,j}$: the Manhattan distance from Cambridge station i to Cambridge station j
- $d_{i,t}$: the predicted daily net change at Cambridge station i at time t
- C_i: the available docks at Cambridge station i
- $Pz_{i,t}$: (uncertainty) the daily net change (demand-supply) perturbations at Cambridge station i at time t.
  - P: the correlation for z, calculated based on RMSE from the prediction model. R_D is defined as 1 through several experiments.

$$P_{i,j} = \begin{cases} e^{-\frac{dist_{i,j}}{R_D}} \times \frac{rmse_i + rmse_j}{14}, & \text{if dist}_{i,j} \leq \text{R}_D \\ 0, & \text{otherwise} \end{cases}$$

  - z: the uncertain variable in the budget uncertainty set

7.1.2 Decision Variables

Next, we defined our decision variables $\forall i, j \in \{1, ..., 7\}, \forall t \in \{1, ..., 92\}$. These variables define the choices the Bluebike must make, specifically whether the technician will visit a station or not, and how many bikes they will pick up or drop off.

- $X_{i,j,t}$: the number of the bikes transferred internally from station i to station j within Cambridge at time t
- $Y_{i,t}$: the number of the bikes transferred externally between one Boston station and Cambridge station i at time t
- $trans_{i,t}$: (auxiliary) the number of the total bikes transferred into station i at time t
- $b_{i,t}$: (auxiliary) the number of bikes placed at station i at time t

7.2 Robust Counterpart

Here's the robust counterpart of the uncertainty constraint:

$$b_{i,j} - b_{i,t-1} - trans_{i,t-1} + \rho|y|_1 + \Gamma|P^T - y|_\infty \leq -d_{i,t-1} \quad (2)$$

7.3 Total Cost

Here are the calculations for the total cost:

$$\begin{aligned} &Unmet_t = \textstyle\sum_i max(0, true\_d_{i,t-1} - b_{i,t-1} - trans_{i,t-1}) \\ &Surplus_t = \textstyle\sum_i max(0, C_i - (b_{i,t-1} + trans_{i,t-1} - true\_d_{i,t-1})) \\ &Total\_Unmet\_Cost = \textstyle\sum_t Unmet_t * Unmet\_Cost \\ &Total\_Sur\_Cost = \textstyle\sum_t Surplus_t * Sur\_Cost \\ &Trans\_Cost = \textstyle\sum_{i,j,t} X_{i,j,t} Cost\_I_{i,j} + \sum_{i,t} |Y_{i,t}| Cost\_O_i \\ &Total\_Cost = Total\_Unmet\_Cost + Total\_Sur\_Cost + Trans\_Cost \end{aligned} \quad (3)$$

7.4 References

- Postek, K., amp; Hertog, D. D. (2016). Multistage adjustable robust mixed-integer optimization via iterative splitting of the uncertainty set. INFORMS Journal on Computing, 28(3), 553-574. doi:10.1287/ijoc.2016.0696
- Bertsimas, D., Zhen, J., Subramanyam, A., Romeijnders, W., Doulabi, H., Lim, Y., amp; Georghiou, A. (2016, June 24). Multistage robust mixed-integer optimization with adaptive partitions. Retrieved May 14, 2021, from https://pubsonline.informs.org/doi/abs/10.1287/opre.2016.1515
- Motivate International, I. (n.d.). Bluebikes System Data. Retrieved May 14, 2021, from https://www.bluebikes.com/system-data
- Motivate International, I. (n.d.). Bluebikes Pricing. Retrieved May 14, 2021, from https://www.bluebikes.com/pricing
- Average cycling speed for new and experienced cyclists. (n.d.). Retrieved May 14, 2021, from https://www.road-bike.co.uk/articles/average-speed.php