Optimization Project: Optimal MBTA Bus Stop Selection

Hayden Ratliff (hratliff@mit.edu), Maxime Wolf (maximew@mit.edu)

December 8, 2023

Contents

1 Introduction and Problem Description

In the face of growing urban transportation challenges, our project embarks on a critical mission: to revolutionize the bus stop selection strategy for the Massachusetts Bay Transportation Authority (MBTA). Our primary goal is to optimize profit margins and simultaneously reduce CO_2 emissions, addressing both economic and environmental concerns. By adopting a weight-based approach, we refine the existing bus stop network, balancing these objectives under a set of defined constraints. To that end, we attempted to create a formulation that is flexible, allowing future modelers or city planners to add constraints as they see fit.

This report contains the following sections: Section 2 details our data sources and pre-processing pipeline; Section 3 explains our first modeling approach; Section 4 explains our second model approach; Section 5 unpacks our results; and Section 6 concludes the report.

2 Data and Pre-Processing

We use bus data from the Massachusetts Bay Transportation Authority (MBTA). Specifically, we use the two following files:

1. MBTA Bus Ridership by Time Period, Season, Route/Line, and Stop: this file contains average load, average passengers boarding, and average passengers unloading at each stop for every bus route in the MBTA bus system. These statistics are calculated over several seasons (for example, Fall 2021, Spring 2022, etc.). Also, the load, boarding, and unloading data for each bus route and season is further divided into 11 daily time periods (for example, EARLY_AM, MIDDAY_BASE, LATE_EVENING).

2. PATI Bus Stops: this file contains a few interesting columns for the project: Stop_ID, Stop_name and the coordinates of each stop.

For this project, we used ridership data from Fall 2022, which was the most recent season available when we pulled the data. We combined Fall 2022 ridership information and the bus stops dataset into a master dataset which contained both passenger information and location information for each route, stop, and time period.

Unfortunately, we did notice some discrepancies in the data. Some stop IDs were not consistent across routes, and some stops were missing entirely from the bus stop location dataset. For example, Nubian Station, which serves 16 MBTA bus routes (including Silver Line 4 and Silver Line 5), has two different IDs (64 and 64000) in the ridership dataset, but is entirely missing in the stop locations dataset. For this reason, we decided to focus the project on the 15 key routes in the MBTA network: 1, 15, 22, 23, 28, 32, 39, 57, 66, 71, 73, 77, 111, 116, and 117. These routes are published on the official MBTA rapid transit map, and are subject to higher frequency standards. Despite narrowing our focus to these 15 routes, there were still some stops missing. Instead of dropping missing stops, we manually filled in the coordinates of these stops using Google Maps.

3 First Model

3.1 Overview

We implemented a multi-objective, mixed-integer optimization formulation for this project. Our first model optimizes one route and does not take into account multiple time periods; instead, we take the average load, people boarding, and people unloading the bus across all 11 time periods.

We have the following data:

1. load after stop j: A_j

2. number of people unloading the bus at stop j: B_j

3. number of people boarding the bus at stop j: C_j

4. distance between 2 consecutive stops: $D_{j-1,j}$

We make the following assumptions: if one stop is removed, then the people who would usually board the bus at this stop won't take the bus at all. Furthermore, the people who would usually get off at this stop will get off at the next selected stop. While these assumptions are not ideal representations of reality, they were necessary for modeling purposes.

3.2 Variables and Constraints

To model the problem, we define the variables a_i (load after stop i) and b_i (the number going off the bus at stop i). We also define a binary variable z_i, equal to 1 if the bus stop i is selected. We also use the following constraints:

1. We adapt the load. If the stop is not selected, then the load does not change. If the stop is selected, then the load can change, with some people boarding the bus and some people leaving.

$$a_i = a_{i-1} + z_i(C_i - b_i) \quad \forall i \in 1 \dots n \tag{1}$$

2. We adapt the number of people unloading. This is equal to the number of people that would unload at the stop if no previous stops were removed, plus any people who were not able to unload at previous stops because those stops were removed from the route (when $z_{i-1} = 0$).

$$b_i = B_i + (1 - z_{i-1})b_{i-1} \quad \forall i \in 1 \dots n \tag{2}$$

3. Capacity constraint for the bus:

$$a_i \leq 40 \quad \forall i \in 1 \dots n \tag{3}$$

4. The load and number of people unloading at the route's first stop is the same as the load from the data:

$$a_1 = A_1 \tag{4}$$

$$b_1 = B_1 \tag{5}$$

5. The load for the last stop must be close to 0. To make solving the problem easier for Gurobi, instead of setting it to 0 exactly, we impose that it must be smaller than 1 (i.e.) on average, there is at most one people after the last stop.

$$a_n \leq 1 \tag{6}$$

6. Finally, we never remove the first and last stop of the route:

$$z_1 = 1 \tag{7}$$

$$z_n = 1 \tag{8}$$

3.3 Objective

Our objective function aims to balance profit maximization and CO_2 emissions minimization. Profit depends on the number of people boarding (because they buy tickets) and CO_2 depends on the load of the bus and the number of stops. Indeed, the more stops, the longer the trip, and the weight transported by the bus (i.e. the load) also increases emissions. We measure the trade-off between these 2 quantities with the parameter λ.

$$\max \left(\lambda \sum_{i=1}^{n} z_i C_i - (1 - \lambda) \left(\sum_{i=1}^{n-1} a_i \cdot d_{i,i+1} + \sum_{i=1}^{n} z_i \right) \right) \tag{9}$$

4 Second Model: With Time Periods

4.1 Overview

Our second model is much more complex because it takes into account all 11 time periods. Because the model takes into account all time periods, it can select different combinations of bus stops for each time period. For example, some stops might be selected regardless of the time of the day, while other stops might only be selected during the AM or PM rush hours.

4.2 Additional Constraints

Our second model adopts similar constraints to the first model, with one extension: to make the model realistic, we require every stop to be selected for several consecutive time periods, if it is selected at all. We made this decision because we believe it is not realistic to change the bus stop selection every 2 hours.

To implement this requirement, we form 4 groups of consecutive time periods. Since there are 11 time periods, 3 of the groups will have 3 time periods, and one will have 2 time periods. We can then use these groups to require that stops are selected in consecutive time periods: for each time period in a group, the selection decision for each bus stop must be the same. As an example, we could imagine that the optimal bus stops selection in the afternoon or in the evening would be different, as they correspond to different "patterns" (people go to work in the morning, and go back home in the evening). Thus, our model finds the optimal time periods groups to form, in order to maximize the objective function described in the previous section while also forcing stops to be selected for several consecutive time periods.

Considering time periods makes the formulation significantly more complex because we need to add lots of variables and constraints. The full formulation can be found in the Appendix.

Below is an example of the time period groups that this model finds. In this matrix y, which has dimension 11×4, $y_{i,j} = 1$ if time period i is assigned to group j.

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

This matrix can be interpreted as:

- t_{11}, t_1, and t_2 are in the same group: night time.
- t_5, t_6, and t_7 are in the same group: late morning/early afternoon.
- t_8, t_9, and t_{10} are in the same group: afternoon/evening.
- t_3 and t_4 are in the same group: early morning.

5 Results

5.1 Overview

We used our first and second model approaches to optimize bus stop selection for Boston's 15 key bus routes. This section details the results we uncovered from applying these models and variations to these models, as well as some commentary on the accuracy, efficacy, and implications of these optimized routes.

Subsection 5.2 details the results from our first model approach, focusing in particular on Route 1 (which runs from Harvard, through MIT and Back Bay, to Nubian Station) with several values of λ. Subsection 5.3 explores the impact of adding additional constraints that limit the removal of stops. Finally, Subsection 5.4 details the results from our second model approach.

5.2 Results for the First Model, Varying λ

For Route 1, we obtain the following results from our first model approach, which does not consider time periods:

Lambda	Profit	CO_2	Number of stops	CO_2 reduction (%)	Profit decrease (%)
0.00	16.54	5.57	5.00	77.70	49.70
0.05	17.28	5.59	5.00	77.60	47.50
0.10	17.28	5.59	5.00	77.60	47.50
0.15	17.28	5.59	5.00	77.60	47.50
0.20	17.28	5.59	5.00	77.60	47.50
0.25	17.28	5.59	5.00	77.60	47.50
0.30	20.73	6.98	6.00	72.00	37.00
0.35	22.61	7.96	7.00	68.10	31.30
0.40	22.61	7.96	7.00	68.10	31.30
0.45	26.91	11.00	10.00	55.90	18.20
0.50	26.91	11.00	10.00	55.90	18.20
0.55	26.91	11.00	10.00	55.90	18.20
0.60	30.61	15.94	15.00	36.10	7.00
0.65	30.61	15.94	15.00	36.10	7.00
0.70	30.61	15.94	15.00	36.10	7.00
0.75	30.61	15.94	15.00	36.10	7.00
0.80	30.91	16.96	16.00	32.00	6.00
0.85	30.91	16.96	16.00	32.00	6.00
0.90	31.39	20.99	20.00	15.90	4.60
0.95	31.48	21.99	21.00	11.90	4.30
1.00	31.48	21.99	21.00	11.90	4.30

As expected, when λ increases, the profit, the CO_2 emissions, and the number of stops increase as well, since larger values of λ give more weight to the profit in the objective function. We also computed the CO_2 emission reduction and the profit decrease we can achieve for each λ, relative to the status quo (all of Route 1's 24 stops are selected). Again, these results make sense. Overall, our model successfully balances profit and CO_2 emissions reduction: by selecting 16 stops, we can reduce CO_2 emissions by 32%, by losing only 6% of the current revenue. A graphical representation of the trade-off between profit and CO_2 emissions can be found in the pareto-optimal frontier plot in Figure 1.

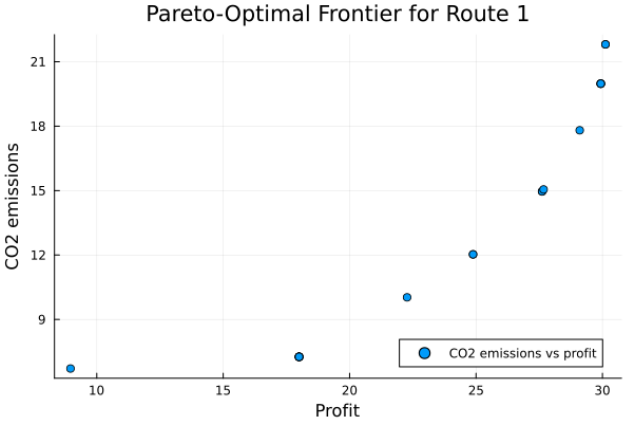


Figure 1: CO_2 emissions and profit by changing λ from 0 to 1

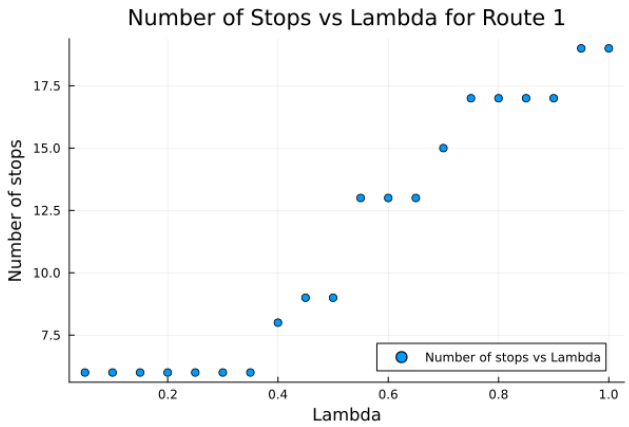


Figure 2: Number of stops when λ changes from 0 to 1

We also see in Figure 2 that when λ increases, the number of stops increases as a stepwise function, visualizing the values from the table.

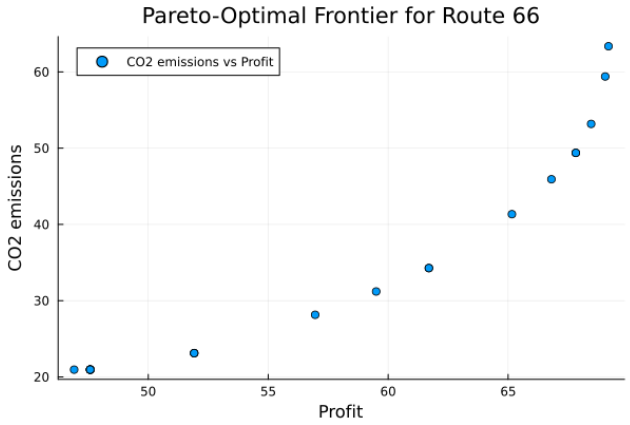


Figure 3: CO_2 emissions and profit by changing λ from 0 to 1

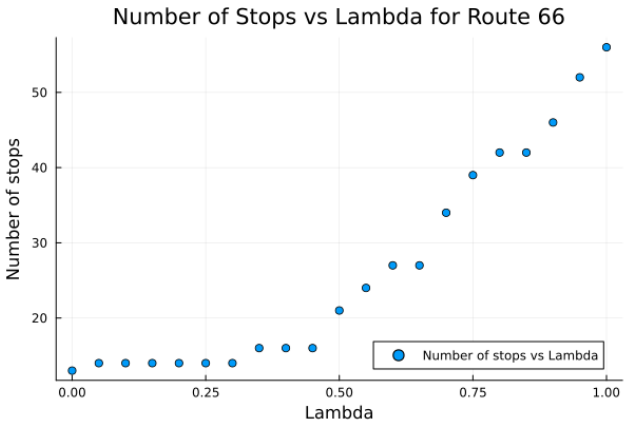
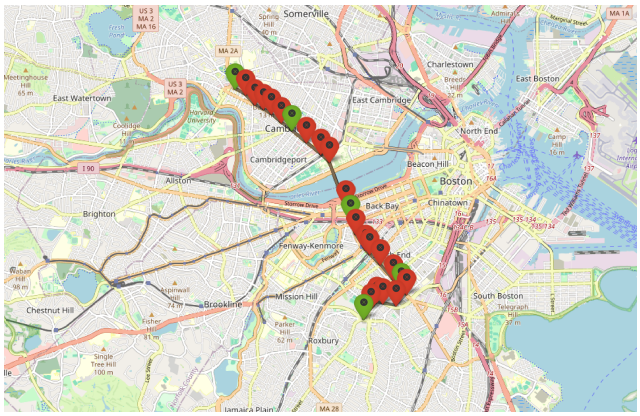


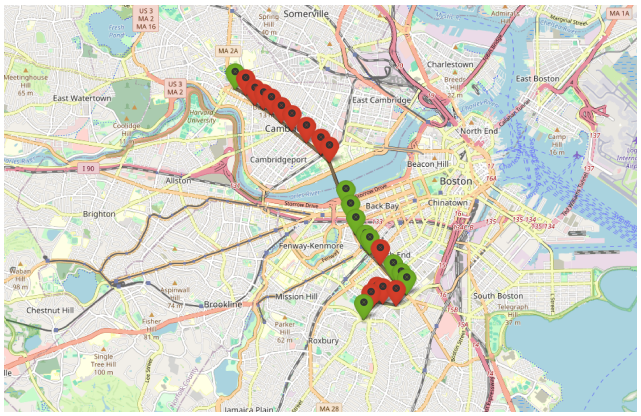
Figure 4: Number of stops when λ changes from 0 to 1

Figure 3 and Figure 4 display the pareto-optimal frontier and the number of stops selected for different values of λ for a different route, Route 66. These plots are interesting because Route 66 has 63 total stops, nearly three times as many as Route 1. This higher volume of stops results in higher granularity in both the pareto-optimal frontier the number of stops plot.

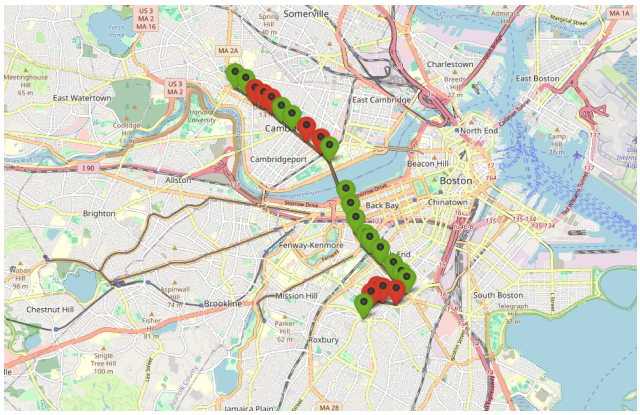
To visualize the results of our model, we built an interactive mapping tool using R's leaflet package. These maps help us understand how modeling choices and varying λ impact the final stops chosen. To illustrate this point, the maps in Figure 5 showcase how the stops selected for Route 1 vary for $\lambda = 0.1, 0.5, 0.75, 1$. When $\lambda = 0.1$, the model only selects 5 stops: the two endpoints, a stop next to Central Square's Red Line stop, a stop in Back Bay, and a stop next to the Boston Medical Center. As λ increases, more stops are selected, with stops that serve the highest volumes of people being selected added first.



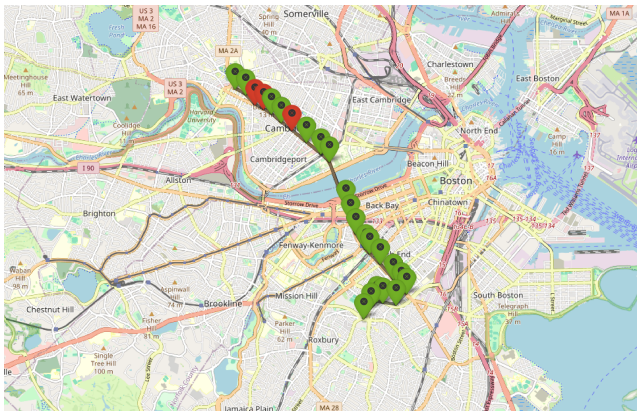
(a) Route 1, $\lambda = 0.1$



(b) Route 1, $\lambda = 0.5$



(c) Route 1, $\lambda = 0.75$



(d) Route 1, $\lambda = 1$

Figure 5: Selected bus stops for Route 1, without time periods, for different λ

5.3 Adding Constraints

We implemented one variant of our first model with the following constraint: we cannot remove two consecutive stops along the route. We visualize the impact of this variation in Figure 8.

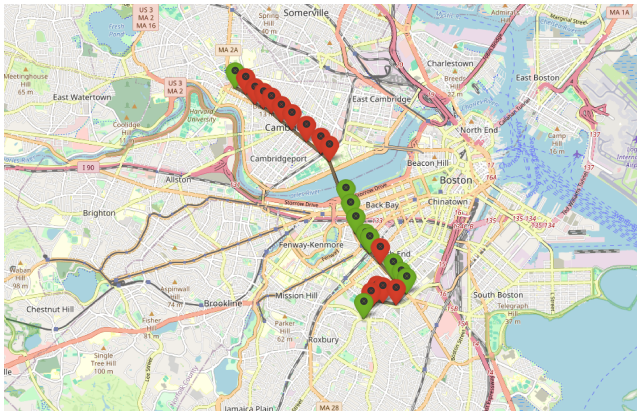


Figure 6: Selected stops for Route 1

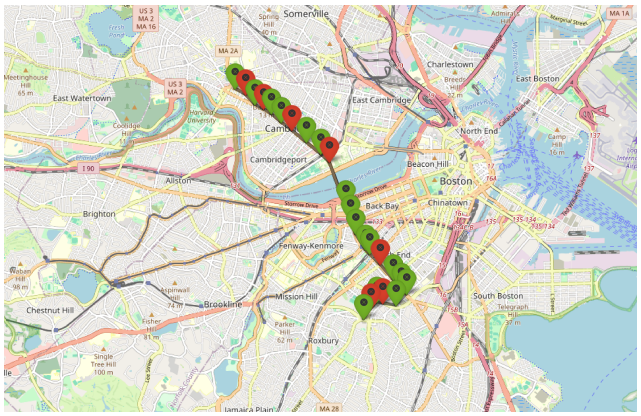


Figure 7: Selected stops for Route 1, without removing consecutive stops

Figure 8: Route 1, $\lambda = 0.5$, with and without the "consecutive stops" constraint

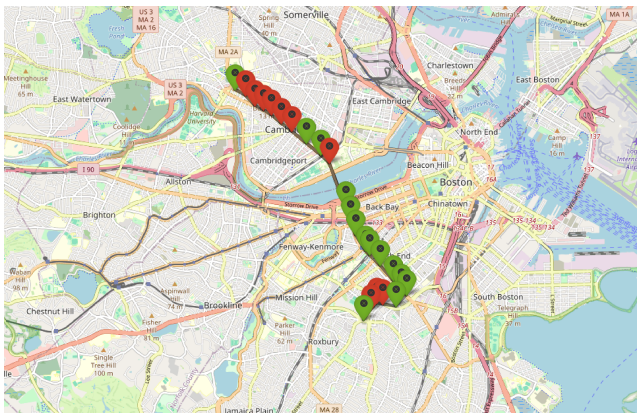
Clearly, prohibiting consecutive stops from being dropped changes which stops are selected. This change also has a large impact on the emissions savings and profit. Without the constraint, we obtained a 55.9% CO_2 emissions decrease, and 18.2% profit decrease with $\lambda = 0.5$. With the constraint, we retain more profit, but nearly halve our CO_2 savings, with a 28.4% CO_2 emissions decrease and 11.5% profit decrease.

These results are expected: we slightly reduce the quality of our solution when adding the constraint, but allow the route to better serve the population. In practice, this is a modeling trade-off that depends on the context. Indeed, we could imagine adding many more constraints to the formulation: for example, one could prefer selecting stops in the center of Boston, because the population density is higher there.

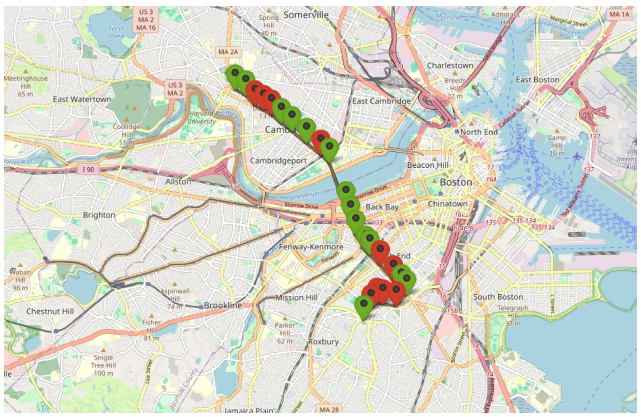
5.4 Results for the Second Model: Considering Time Periods

We also obtained results for our second model approach, which takes time periods into account. Due to the complexity of the model, we focused on Route 1 for two different values of λ: 0.5 and 0.75.

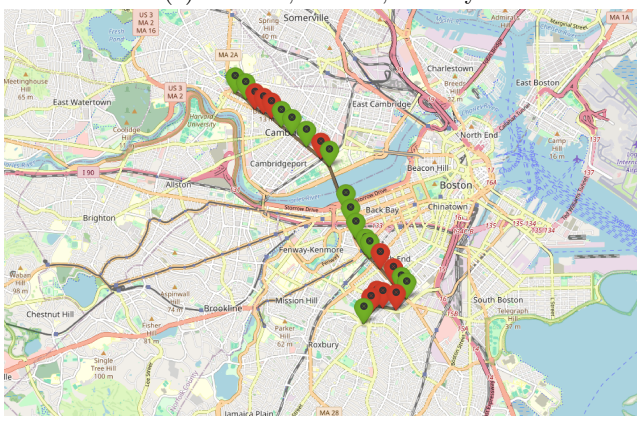
In Figure 9, we show the selected stops for 4 different time periods:



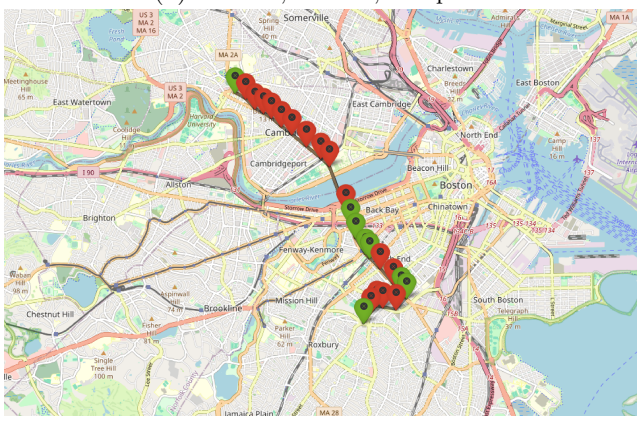
(a) Route 1, $\lambda = 0.5$, Midday



(b) Route 1, $\lambda = 0.5$, PM peak



(c) Route 1, $\lambda = 0.5$, Evening



(d) Route 1, $\lambda = 0.5$, Night

Figure 9: Selected bus stops for Route 1 with $\lambda = 0.5$ and different time periods

These maps showcase how selected stops change based on time period, but that our grouping constraint is in effect. We can see that the model selected PM peak and Evening in the same group because the selected stops for these two time periods are identical (see (b) and (c)).

Despite the grouping constraint, we can see that the model adapts to the different demand patterns observed at different times of the day. For example, fewer stops are selected during the Midday time period versus PM peak and Evening, and the Night time period sees an even larger reduction in stops. Moreover, during the Midday time period, more stops are selected in the center of Boston than during PM peak and Evening. This is also an intuitive finding: during the Midday time period, many people are at work and may tend to stay in the city center during lunch break. During PM peak, these same people leave the city center, and the model addresses this demand by selecting stops more evenly along the route.

6 Conclusion and Impact

Our project aimed to optimize bus stop selection for the Massachusetts Bay Transportation Authority (MBTA). By implementing a multi-objective, mixed-integer optimization approach, we achieved significant progress in enhancing urban transportation efficiency. Our weight-based approach to refine the MBTA bus stop network successfully tackled the dual objectives of maximizing profit and minimizing CO_2 emissions. Our two models, both the simpler initial model and the more complicated time periods model, have demonstrated that public transit planners can achieve a significant environmental impact without sacrificing economic benefits and ridership too much.

Through our analysis, we observed that adjusting the number of stops based on the time of day and specific route requirements can lead to a more efficient allocation of resources. For instance, our results showed that selecting 16 stops along Route 1 could reduce CO_2 emissions by 32% while only incurring a 6% loss in revenue compared to today's route. Additionally, the inclusion of constraints like avoiding the removal of consecutive stops ensures that the transportation network continues to serve the population effectively, balancing optimization with practical utility. To that end, several potential useful (and challenging) extensions to this project exist, including optimizing all routes simultaneously. We believe that our formulation flexible enough that it could allow future modelers or city planners to easily incorporate new constraints depending on the bus network or the city.

Appendix A: Full Formulation of the 2nd Model

A.1 Overview

As before, we define the variables $a_{i,t}$ (load after stop i at time t) and $b_{i,t}$ (the number going off the bus at stop i, at time t). We also define a binary variable $z_{i,t}$, equal to 1 if the bus stop i is selected at time t. Note that t here denotes one of the 11 time periods of a day. We have two additional binary variables: $y_{t,j}$ (if time t is in group j) and $y_{t,j}$ (it will be equal to 1 when $\sum_{t \in G_j} y_{tj}$ for every group G_j). As a reminder, we fix the number of groups and time periods in each group: G_1, G_2, G_3 have 3 time periods each, G_4 has 2 time periods.

A.2 Constraints

We have the same constraints as in our first model, but for every time period t:

$$a_{i,t} = a_{i-1,t} + z_{i,t}(C_{i,t} - b_{i,t}) \quad \forall i \in 1 \dots n \tag{10}$$

$$b_{i,t} = B_{i,t} + (1 - z_{i-1,t})b_{i-1,t} \quad \forall i \in 1 \dots n \tag{11}$$

$$a_{i,t} \leq 40 \quad \forall i \in 1 \dots n \tag{12}$$

$$a_{1,t} = A_{1,t} \tag{13}$$

$$b_{1,t} = B_{1,t} \tag{14}$$

$$a_{n,t} \leq 1 \tag{15}$$

$$z_{1,t} = 1 \tag{16}$$

$$z_{n,t} = 1 \tag{17}$$

Secondly, we need new constraints to form the groups of time periods:

1. Number of time periods in each group:

$$\sum_{t=1}^{11} y_{t,j} = 3 \quad \forall j \in \{1, 2, 3\} \tag{18}$$

$$\sum_{t=1}^{11} y_{t,4} = 2 \tag{19}$$

2. Each time period must be in exactly one group:

$$\sum_{j=1}^{4} y_{t,j} = 1 \quad \forall t \in \{1 \dots 11\} \tag{20}$$

3. Now, we make sure that s verifies the constraint given earlier. Here, a mod b denotes the remainder when dividing a by b. It is necessary to model that times 11 and 1 are consecutive (because it goes to the next day). For one group j, if one of 3 (or 2, depending on the group) consecutive $y_{t,j}$ is 0, then we want $z_{t,j}$ to be 0:

$$3s_{t,j} \leq y_{t,j} + y_{(t \bmod 11)+1,j} + y_{(t \bmod 11)+2,j} \quad \forall j \in \{1, 2, 3\} \tag{21}$$

$$2s_{t,4} \leq y_{t,4} + y_{(t \bmod 11)+1,4} \tag{22}$$

4. Also, we need to impose:

$$\sum_{t=1}^{11} s_{t,j} = 1 \quad \forall j \in \{1, 2, 3, 4\} \tag{23}$$

5. Finally, we use big M linking constraints to" model the following: if times t and k are in the same group, then $\forall i \in 1 \dots n, \quad z_{i,t} = z_{i,k}$

$$z_{i,t} \geq z_{i,k} + M(2 - y_{t,j} - y_{k,j}) \quad \forall i \in 1 \dots n, \quad \forall t, k \in 1 \dots 11, \quad \forall j \in 1 \dots 4 \tag{24}$$

$$z_{i,t} \leq z_{i,k} - M(2 - y_{t,j} - y_{k,j}) \quad \forall i \in 1 \dots n, \quad \forall t, k \in 1 \dots 11, \quad \forall j \in 1 \dots 4 \tag{25}$$

A.3 Objective

Lastly, we adopt a nearly identical objective to our first model:

$$\max \left(\lambda \sum_{i=1}^{n} \sum_{t=1}^{11} z_{i,t} C_i - (1 - \lambda) \left(\sum_{i=1}^{n-1} \sum_{t=1}^{11} a_{i,t} \cdot d_{i,i+1} + \sum_{i=1}^{n} \sum_{t=1}^{11} z_{i,t} \right) \right) \tag{26}$$