Optimization Group Project Report

Zachary Bell, Mattia Ravasio

email: zbell@mit.edu, mravasio@mit.edu

Problem Statement

The Massachusetts Bay Transportation Authority (MBTA) is a public agency responsible for operating the Subway service in the Boston urban area. The subway system faces a demand that is impacted by seasonality, weekly, and daily trends and in order to run a high level service the MBTA needs to schedule the number of trains that run every day in advance. In order to tackle the issue, the agency collects data on user ridership during given time windows.

In this project, we plan on minimizing the cost of serving customers by modeling the number of trains that need to run for each line in each direction and in every time window. We want to specify that the green and the red line have their own sub-lines (Green B, Green C, Green D, Green E, Red Ashmont and Red Braintree) and thus we will have to model the fact that some stations are served by more than one sub-line while some are not. We also investigated uncertainty in the passenger demand to create more flexibility in our problem.

Data source and preprocessing

The data source that we are going to use is the **MBTA Rail Ridership by Time Period, Season, Route/Line, and Stop,** which can be found at the following link. For a given subway station, the data aggregates the average flow of people for given time window (details in Appendix). In the data, there are 3 day types, each with a given number of time windows: weekdays have 9 different time windows, and Saturday and Sunday both have only one. We will focus on optimizing the schedule for the weekdays. Since the MBTA does not disclose any information about the duration of the time windows nor the arrival time of customer at different stations, we decided not to optimize the actual arrival time of trains at different stations. There are three seasons of data (Fall 2017, Fall 2018, Fall 2019), and we will only optimize on the 2019 data. The average flow of people is the best proxy available for the demand of customers in the subway service.

Our first step was to preprocess the data. For the different lines, we selected the average flow for the stations on a given direction and then pivoted the dataset in order to have the time windows as columns, the stop names as rows, and the average flows as values. We ordered them according to intuition: "VERY_EARLY_MORNING", "EARLY_AM", "AM_PEAK", "MIDDAY_SCHOOL", "MIDDAY_BASE", "PM_PEAK", "EVENING", "LATE_EVENING", and "NIGHT". We constructed two pivoted matrices for each line, one representing each direction, and the last step is to stack the two matrices along a third dimension to obtain a tensor. The final output is a three-dimensional tensor for a given line, and their dimensions are reported in the following table:

Line	Dimension (Directions, Stations, Time windows)
Blue	(2, 12, 9)
Green	(2, 61, 9)
Red	(2, 22, 9)
Orange	(2, 20, 9)

Lastly, we created an array with the names of each station corresponding to the sub-lines of the Green and Red line. This is used to create a binary matrix of size (l, i), where l is the number of sub-lines and i is the number of stations. We will use this matrix in the constraints of the formulation.

Deterministic formulations

While exploring the data, we noticed that at stations where different lines meet, the average flow is calculated differently for each single line. When different color lines intersect, the data lines up: Park Street station, the Red and Green line intersection, has different average flows for the Red and the Green line, but they are both consistent with the other average flows calculations occurring in each line. However, when sub-lines within the same color line merge or split, the average flow data does not align from station to station: for example, Park Street is a station on all 4 Green sub-lines, and its average flow is not split up by sub-line. These properties of the data allow us to run different individual models for each color lines but constrain us to one single model for the sub-lines running on the same line. Due to characteristics of each line, we end up having two different macro-formulations: one for the Blue and Orange line, that don't have any sub-lines, and one for the Green and Red lines, with 4 and 2 sub-lines respectively.

As explained in the previous section, the data is divided in different time windows. For every line, we are acting under the assumption that in a given direction, a train will only be able to travel in one direction and that it will travel the full length of the

line. We do not have information of the duration of each time period, so we were not able to determine how far trains actually travel in a given time period. While we are aware that trains would keep cycling along one line by changing directions every time they arrive at a final station, but we cannot model this component without any additional temporal information.

The initial formulation will not incorporate robust components. We will first describe the formulation for the Blue and Orange line:

$$
\begin{aligned}
\min_{x,u,s,r} \quad & \sum_{d=1}^{2}\sum_{t=1}^{10}(Cost_l \cdot x_{dt} + 0.95 \cdot Cost_l \cdot s_{dt} + \sum_{i=1}^{I} q \cdot u_{dti}) & \\
\text{s.t.} \quad & u_{dti} + Capacity_l \cdot (x_{dt} + s_{dt}) \geq Avg\ flow_{dti} + u_{d,t-1,i} & \forall d, i, t = 2, ...T \\
& u_{d1i} + Capacity_l \cdot (x_{d1} + s_{d1}) \geq Avg\ flow_{d1i} & \forall d, i \\
& \sum_t x_{1t} + \sum_t x_{2t} \leq Total\ Number\ of\ Trains & \\
& s_{d1} = 0 & \forall d \\
& s_{dt} + x_{dt} \geq 1 & \forall d, t \\
& r_{1t} = x_{2,t-1} + s_{2,t-1} - s_{1t} + r_{1,t-1} & \forall d, t \\
& r_{2t} = x_{1,t-1} + s_{1,t-1} - s_{2t} + r_{2,t-1} & \forall d, t = 2, ...9 \\
& s_{dt} \leq r_{d,t-1} & \forall d, t = 2, ...9 \\
& r_{d9} \geq x_{d1} & \forall d \\
& r_{11} = x_{21} & \\
& r_{21} = x_{11} & \\
& x \in \mathbb{N}^{2\times 9}, u \in \mathbb{N}^{2\times 9\times I}, r \in \mathbb{N}^{2\times 9}, s \in \mathbb{N}^{2\times 9} &
\end{aligned}
\tag{1}
$$

Variables and Coefficients:

- d indicates the direction and thus has values 1 and 2,
- t inidcates the time windows and thus has values from 1 to 10
- i indicates the single stations and thus has values from 1 to 12 for the Blue line, 1 to 20 for the Orange line.
- x_{dt} is a decision variable that indicates the number of new trains that run on a given line direction during a given time window. With “new trains” we refer to trains that have to be taken out of the depot to be used. This decision variable can only assume non-negative integer values.
- s_{dt} is a decision variable that represents the number of trains already running in previous time windows that move across a given line direction during that time window. These trains can then be reused on the opposite direction in the following time windows without having to store them in the depot. Again, this decision variable can only assume non-negative integer values.
- r_{dt} is a decision variable that helps us keeping track of the amount of trains that are “resting” at some final station, meaning trains that were either x or s and are not moving in that particular time window. Again, this decision variable can only assume non-negative integer values.
- u_{dti} is a decision variable that models the amount of unmet demand for a line at a given station, during a given time window and on a given direction. Again, this decision variable can only assume non-negative integer values.
- q will be the cost of unmet demand and we will be using $5 as it factors in the price of a single ride ticket ($2.4) and the cost of the inconvenice caused to the user.
- $Cost_l$ is the cost of running a train for a speicifc line, the estimations of this coefficient are from available data online.
- $Capacity_l$ is the capacity of a train belonging to line l and the estimation was based on reported capacities.
- $Avg\ Flow_{dti}$ is the average flow of customers for a given line in station i, on direction d and in time window t. This will be our proxy for the demand of customers.

Objective function and constraints:

- The objective function is (1) the cost of running a new train (summed over all directions and time windows) plus (2) the cost of running a train that was already moving plus (3) the cost of unmet demand (summed over all stations). We assume that the cost of a train s will be slightly less than the cost of a x train since the train is already up and running.
- Train capacity constraints (1,2): For every direction, time window, and station, the sum of the average flow + the unmet demand from the previous time window must be less or equal to the sum of the unmet demand time the capacity of a train of

that line multiplied by the total number of trains running. This is modeled in two constraints to account for the first time window.

- Train quantity constraint (3): The solution cannot include more trains than are available. The incorporation of this constraint allows us to observe how our models behave differently under a limited trains fleet. We can see this constraint as the number of trains we have available at the depot.
- Running train constraints (4,5): The amount of trains that running before the first time window will be 0. Additionally, there must be at least one train running in every time window and direction. This second constraint ensures that we are running a train even when there are time windows with very low demand. We don't want the optimal solution to include lines running 0 trains because that will cause consumer complaints.
- Resting train constraints (6,7): For a given direction and time window excluding the first, r will be equal to the sum of the trains coming from the opposite direction in the previous time window minus the number of already moving train that will move in that time window along that direction plus r in the previous time window. r helps us keep count of the number trains that had moved in previous time windows but will be staying still at a starting station in the current time window. r in the first time window will be only equal to the number of new trains coming from the opposite directions (as shown in constraints 10,11)
- The eigth constraint states that the number of s starting along a given direction must be less or eqaul than the number of r in the previous time window.
- The ninth constraint formulates the fact the the number of trains that are not moving at a final station at the end of the day will be greater or equal than the number of new trains moving from that station during the next morning in the first time window. This allows for a schedule that can be used in the same way for every day in the season.

Let's now focus on the formulation for the Red and the Green line. The formulation is similar but uses one more auxiliray binary matrix and different dimensions for the decision variables:

$$
\begin{aligned}
\min_{x,u.s.r} \quad & \sum_{d=1}^{2}\sum_{t=1}^{10}(\sum_{g=1}^{G} Cost_l \cdot x_{dtg} + 0.95 \cdot Cost_l \cdot s_{dtg} + \sum_{i=1}^{I} q \cdot u_{dti}) \\
\text{s.t.} \quad & u_{dti} + Capacity_l \cdot \sum_{g=1}^{G}((x_{dtg} + s_{dtg}) \cdot z_{gi}) \geq Avg\ flow_{dti} + u_{d,t-1,i} && \forall d,i,t = 2...T \\
& u_{d1i} + Capacity_l \cdot \sum_{g=1}^{G}((x_{d1g} + s_{d1g}) \cdot z_{gi}) \geq Avg\ flow_{d1i} && \forall d,i \\
& \sum_t x_{1t} + \sum_t x_{2t} \leq Total\ Number\ of\ Trains \\
& s_{d1g} = 0 && \forall d,g \\
& s_{dtg} + x_{dtg} \geq 1 && \forall d,t,g \\
& r_{1tg} = x_{2,t-1,g} + s_{2,t-1,g} - s_{1tg} + r_{1,t-1,g} && \forall d,t,g \\
& r_{2tg} = x_{1,t-1,g} + s_{1,t-1,g} - s_{2tg} + r_{2,t-1,g} && \forall d,g,t = 2,...9 \\
& s_{dtg} \leq r_{d,t-1,g} && \forall d,g,t = 2,...9 \\
& r_{d9g} \geq x_{d1g} && \forall d,g \\
& r_{11g} = x_{21g} && \forall g \\
& r_{21g} = x_{11g} && \forall g \\
& x \in \mathbb{N}^{2\times 9\times G}, u \in \mathbb{N}^{2\times 9\times I}, r \in \mathbb{N}^{2\times 9\times G}, s \in \mathbb{N}^{2\times 9\times G}
\end{aligned}
\tag{2}
$$

- g will be the number of sub-lines per line. The Green line has four sub-lines and the Red line has two sub-lines.
- z_{gi} is a binary auxiliary matrix that equals 1 if the sub-line g serves the station i. The matrix will have size (4,61) for the Green line and (2,22) for the Red line. To create this matrix, we used an array with all the names of the stations served by a given sub-line and ran two nested for-loops to populate the matrix with a 1 if the name of a given station is present in a given array (and 0 otherwise).

Robust formulations

It's clear that assuming the average flow to be a deterministic variable is not representative of reality. For this reason we introduced uncertainty in the passegners' demand (a.k.a. the average flow). We decided the most fitting uncertainty set was the budget uncertainty set:

$$\mathcal{U} = \{Avg\ Flow_{dti}, \gamma \in \mathbb{N}^{D\times T\times I} \mid \mu_{dti} - \delta_{dti}\gamma_{dti} \leq Avg\ Flow_{dti} \leq \mu_{dti} + \delta_{dti}\gamma_{dti},\ \forall d,t,i; \|\gamma\|_1 \leq \Gamma\}$$

We use the $Avg\ Flow_{dti}$ from the deterministic formulation to represent μ_{dti} , but we do not have any data for δ_{dti}. In order to create realistic artifical data, we generated a tensor of the same shape of the $Avg\ Flow$ variable and filled it with random values coming from a uniform distribution between 0.02 and 0.06. We then multiplied the newly created tensor with $Avg\ Flow$ in an element-wise fashion and type-casted the result to be of integer type, allowing us to obtain something similar to a standard deviation.

Before explaining the reformulation we used, we must specify a few adjustments needed: the average flow is an integer value, but this would not allow us to use reformulations due to the lack of strong duality. For this reason, we decided to relax the integrality of the average flow of passengers. This relaxation should not impact the problem in a substantial way. The actual uncertanty set we use is the following:

$$\mathcal{U} = \{Avg\ Flow_{dti}, \gamma \in \mathbb{R}^{D\times T\times I} \mid \mu_{dti} - \delta_{dti}\gamma_{dti} \leq Avg\ Flow_{dti} \leq \mu_{dti} + \delta_{dti}\gamma_{dti},\ \forall d,t,i; \|\gamma\|_1 \leq \Gamma\}$$

Now, we will derive the robust counterpart for this problem. We will only show the sub-problem to keep the derivation comprehensive. For every direction, time window, and train station:

$$\begin{aligned} \max_{AvgFlow,\gamma} \quad & Avg\ Flow_{dti} \\ \text{s.t.} \quad & \mu_{dti} - \delta_{dti}\gamma_{dti} \leq Avg\ Flow_{dti} \\ & Avg\ Flow_{dti} \leq \mu_{dti} + \delta_{dti}\gamma_{dti} \\ & \|\gamma\|_1 \leq \Gamma \\ & Avg\ Flow_{dti}, \gamma_{dti} \in \mathbb{R} \end{aligned} \tag{3}$$

The linearized version will be the following:

$$\begin{aligned} \max_{AvgFlow,\gamma,f} \quad & Avg\ Flow_{dti} \\ \text{s.t.} \quad & \mu_{dti} - \delta_{dti}\gamma_{dti} \leq Avg\ Flow_{dti} \\ & Avg\ Flow_{dti} \leq \mu_{dti} + \delta_{dti}\gamma_{dti} \\ & \sum_{d,t,i} f_{dti} \leq \Gamma \\ & f_{dti} \geq \gamma_{dti} \\ & f_{dti} \geq -\gamma_{dti} \\ & Avg\ Flow_{dti}, \gamma_{dti}, f_{dti} \in \mathbb{R} \end{aligned} \tag{4}$$

The dual of this problem will be:

$$\begin{aligned} \min_{\alpha,\beta,l,\lambda,\phi} \quad & -\mu_{dti}\alpha_{dti} + \mu_{dti}\beta_{dti} + \Gamma l \\ \text{s.t.} \quad & -\alpha_{dti} + \beta_{dti} \geq 1 \\ & -\delta_{dti}\alpha_{dti} - \delta_{dti}\beta_{dti} + \lambda_{dti} - \phi_{dti} \geq 0 \\ & l - \lambda_{dti} - \phi_{dti} \geq 0 \\ & \alpha, \beta, l, \lambda, \phi \geq 0 \end{aligned} \tag{5}$$

The final robust reformulation:

$$\min_{x,u,s,r,\alpha,\beta,l,\lambda,\phi} \quad \sum_{d=1}^{2}\sum_{t=1}^{10}(\sum_{g=1}^{G} Cost_l \cdot x_{dtg} + 0.95 \cdot Cost_l \cdot s_{dtg} + \sum_{i=1}^{I} q \cdot u_{dti})$$

$$\begin{aligned}
\text{s.t.} \quad & u_{dti} + Capacity_l \cdot \sum_{g=1}^{G}((x_{dtg} + s_{dtg}) \cdot z_{gi}) \geq -\mu_{dti}\alpha_{dti} + \mu_{dti}\beta_{dti} + \Gamma l + u_{d,t-1,i} && \forall d, i, t = 2...T \\
& u_{d1i} + Capacity_l \cdot \sum_{g=1}^{G}((x_{d1g} + s_{d1g}) \cdot z_{gi}) \geq -\mu_{dti}\alpha_{dti} + \mu_{d1i}\beta_{d1i} + \Gamma l && \forall d, i \\
& -\alpha_{dti} + \beta_{dti} \geq 1 && \forall d, t, i \\
& -\delta_{dti}\alpha_{dti} - \delta_{dti}\beta_{dti} + \lambda_{dti} - \phi_{dti} \geq 0 && \forall d, t, i \\
& l - \lambda_{dti} - \phi_{dti} \geq 0 && \forall d, t, i \\
& \sum_t x_{1t} + \sum_t x_{2,t} \leq Total\ Number\ of\ Trains && (6) \\
& s_{d1g} = 0 && \forall d, g \\
& s_{dtg} + x_{dtg} \geq 1 && \forall d, t, g \\
& r_{1tg} = x_{2,t-1,g} + s_{2,t-1,g} - s_{1tg} + r_{1,t-1,g} && \forall d, t, g \\
& r_{2tg} = x_{1,t-1,g} + s_{1,t-1,g} - s_{2tg} + r_{2,t-1,g} && \forall d, g, t = 2, ...9 \\
& s_{dtg} \leq r_{d,t-1,g} && \forall d, g, t = 2, ...9 \\
& r_{d9g} \geq x_{d1g} && \forall d, g \\
& r_{11g} = x_{21g} && \forall g \\
& r_{21g} = x_{11g} && \forall g \\
& \alpha, \beta, l, \lambda, \phi \geq 0, x \in \mathbb{N}^{2\times 9\times G}, u \in \mathbb{N}^{2\times 9\times I}, r \in \mathbb{N}^{2\times 9\times G}, s \in \mathbb{N}^{2\times 9\times G}
\end{aligned}$$

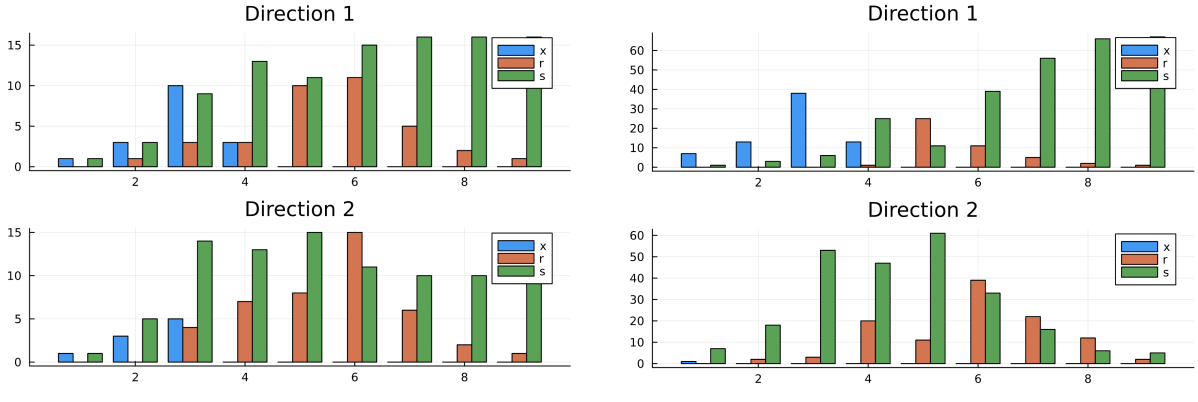
The provided robust formulation models the Green and Red lines. The robust formulation for the Blue and the Orange lines can be found in the appendix, with the only difference being their lack of sub-lines.

Results

The code developed for this project can be found at the following repository. Unfortunately, we could not locate any information about an MBTA baseline. Therefore, we will not concentrate our solution's cost reduction, but rather the understanding of key aspects of the decisions made by our model.

The first thing that we noticed is that our models tend to favor the usage of new trains (the x in the formulations) in the first time windows, and then try to use trains already moving (the s in the fomrulation). This is a behavior that we would expect since there is an additional cost in moving a new train out of the depot and turning it on. In addition, in those time windows that are niether the first nor the last, the models start to balance the number of s and the number of trains that are "resting" at one endopint of the line (r in the formulation), and we can usually see a peak of r. This peak is probably due to the lower overall demand in the middle of the working day, and thus the MBTA would profit by reducing the number of running trains during those time windows. In the last time windows, we see that the number of s starts surging again and r decreases.

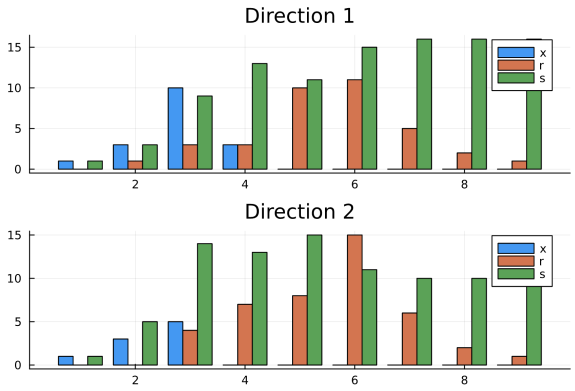
Different lines also have differing behaviors. For example, the Orange line has symmetric demand across the two different directions while the directional demand on the Blue line flips based on time of day (demand for direction 2 in the morning, direction 1 at night). These trends can be observed in the following image:



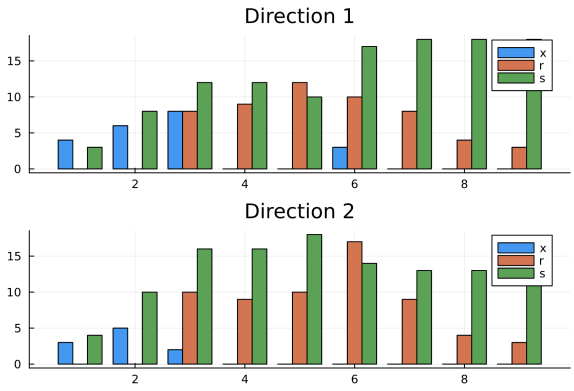
Orange Line Blue Line

The orange line has similar distributions for the three variables across the two directions, while the Blue line has more new trains in direction one, a peak of s in the middle of the day in directions 2 and then a new peak of s by the end of the day in direction 1. The plots were generated using the deterministic formulations with a very large amount of trains at disposal.

With the robust formulations, we do notice a change in strategic decisions from the model. Using the orange line as an example, the robust model (on the right) uses new trains more sparingly. Also, for the robust formulations, despite changes in the distribution, the general trends observed remain. The following plots were realized using a large number of available trains.

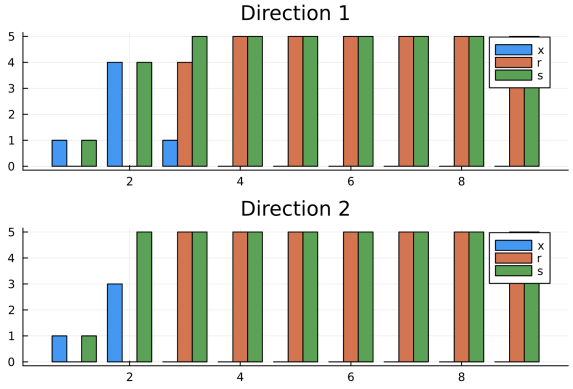


Orange Line with deterministic formulation

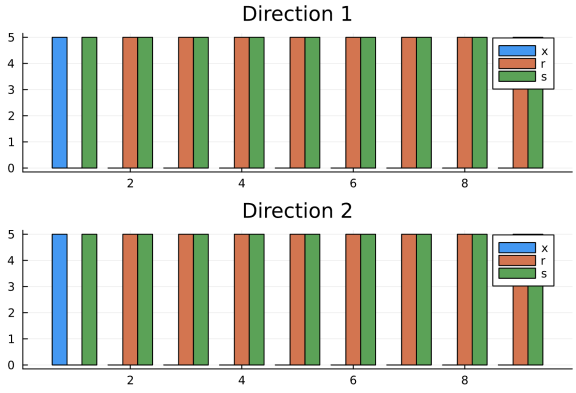


Orange Line with robust formulation

We also wanted to understand the behavior of the model with a limited fleet. As seen in the deterministic plots, new trains enter at different time windows, and once they are all running, a perfect balance between r and s is achieved. In the robust formulations, the model suggest using available new trains in the first time window, leading to a perfect balance between r and s :



Orange Line with deterministic formulation with low number of trains at disposal



Orange Line with robust formulation with low number of trains at disposal

Appendix

- Average flow calculation:

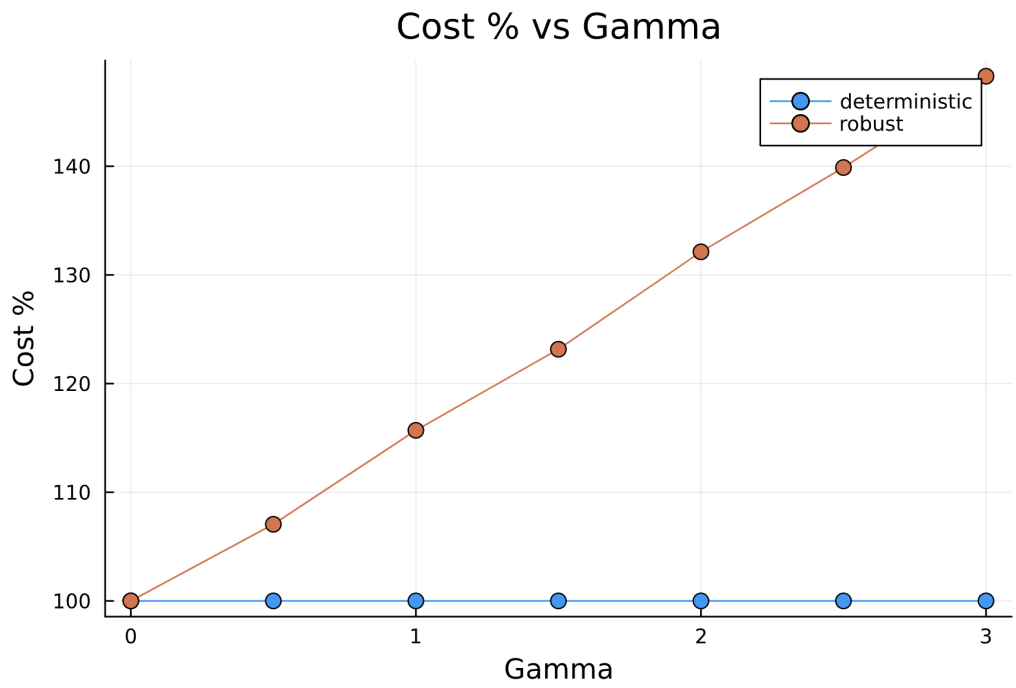
$$Avg\ Flow_{i,t,l} = \begin{cases} Avg\ On_{i,t,l} - Avg\ Off_{i,t,l} + Avg\ Flow_{i-1,t,l} & if\ i \neq 0 \\ Avg\ On_{0,t,l} & if i = 0 \end{cases}$$

where l is a line direction, t is a time window, i is a given station, $Avg\ On$ is the average number of people getting on the train at that given station and $Avg\ Off$ is the average number of people getting off.

- Robust counterpart for the Blue and Orange lines:

$$
\begin{aligned}
\min_{x,u,s,r,\alpha,\beta,l,\lambda,\phi} \quad & \sum_{d=1}^{2}\sum_{t=1}^{10}(x_{dt} + 0.95 \cdot Cost_l \cdot s_{dt} + \sum_{i=1}^{I} q \cdot u_{dti}) \\
\text{s.t.} \quad & u_{dti} + Capacity_l \cdot (x_{dt} + s_{dt}) \geq -\mu_{dti}\alpha_{dti} + \mu_{dti}\beta_{dti} + \Gamma l + u_{d,t-1,i} && \forall d, i, t = 2...T \\
& u_{d1i} + Capacity_l \cdot (x_{d1} + s_{d1}) \geq -\mu_{dti}\alpha_{dti} + \mu_{d1i}\beta_{d1i} + \Gamma l && \forall d, i \\
& -\alpha_{dti} + \beta_{dti} \geq 1 && \forall d, t, i \\
& -\delta_{dti}\alpha_{dti} - \delta_{dti}\beta_{dti} + \lambda_{dti} - \phi_{dti} \geq 0 && \forall d, t, i \\
& l - \lambda_{dti} - \phi_{dti} \geq 0 && \forall d, t, i \\
& u_{d9i} = 0 && \forall d, i \\
& s_{d1} = 0 && \forall d \qquad (7) \\
& s_{dt} + x_{dt} \geq 1 && \forall d, t \\
& r_{1t} = x_{2,t-1,} + s_{2,t-1} - s_{1t} + r_{1,t-1} && \forall d, t \\
& r_{2t} = x_{1,t-1} + s_{1,t-1} - s_{2t} + r_{2,t-1} && \forall d, t = 2, ...9 \\
& s_{dt} \leq r_{d,t-1,} && \forall d, g, t = 2, ...9 \\
& r_{d9} \geq x_{d1} && \forall d \\
& r_{11} = x_{21} \\
& r_{21} = x_{11} \\
& \alpha, \beta, l, \lambda, \phi \geq 0, x \in \mathbb{N}^{2\times 9}, u \in \mathbb{N}^{2\times 9\times I}, r \in \mathbb{N}^{2\times 9}, s \in \mathbb{N}^{2\times 9}
\end{aligned}
$$

- Plotting of the objective value of the robust model and the deterministic one:



The y-axis represents the percentage of the cost of the deterministic formulation. The x-axis contains the Gamma (amount of uncertanty) that was used in the formulation. This plot was is for the Blue line model without any constraints on the total number of trains.