Accelerating Emergency Response: An Optimized Ambulance Dispatch Model for Manhattan

Ellie Yang, Naiqi Zhang, Sneha Sathish Kumar

15.C57 Optimization Methods

Abstract

This project examines three optimization strategies to minimise response of ambulance dispatch in Manhattan: a min-sum model, a min-max model, and robust versions of both. The min-sum model seeks to minimize the total response time across all emergency incidents, while the min-max model aims to minimize the maximum response time to ensure equitable service delivery. The robust models address uncertainties in real-world scenarios by incorporating potential variations in station capacities and incident demands. By comparing these approaches, we aim to identify the most effective and resilient deployment strategy for emergency medical services (EMS). Our mathematical models account for practical constraints and potential fluctuations in parameters. This comparative analysis provides valuable insights into the trade-offs between system efficiency, equitable response, and performance under uncertain conditions. Based on the models, we performed sensitivity analysis across stations. We identified the bottleneck stations with non-zero shadow prices and made staffing suggestions to dispatch areas.

1 Introduction

Manhattan's emergency medical response system faces significant challenges, with response times reaching concerning levels that directly impact patient outcomes. Recent data shows that the average EMS response time to life-threatening emergencies in Manhattan has increased to 12 minutes and 26 seconds, a 29% increase from 9 minutes and 40 seconds in 2014. This stands in stark contrast to the national average response time of seven minutes in urban areas. The situation has deteriorated further, with non-life-threatening emergencies now experiencing wait times of up to 28 minutes.

The severity of this problem is magnified by Manhattan's unique characteristics. With 1.6 million residents concentrated in just 22.8 square miles, the borough's emergency response system operates under intense pressure. Traffic congestion has reached critical levels, with average vehicle speeds in Midtown dropping to just 4.8 mph and 6.9 mph in Lower Manhattan. These delays have deadly consequences—medical research indicates that in cases of cardiac arrest, every minute without CPR reduces survival chances by 10%.

The challenge is compounded by increasing demand and resource constraints. The city experiences approximately 5,000 emergency calls daily, despite being equipped to handle only 4,000. This volume, combined with a 17.5% decline in active certified EMS responders from 2019 to 2022, creates a perfect storm that threatens the system's ability to provide timely emergency care. This research addresses this critical public health challenge by exploring mathematical optimization approaches to improve ambulance dispatch efficiency and response times in Manhattan.

2 Data Preprocessing

To perform the optimization model on ambulance waiting time, we primarily utilized the EMS Incident Dispatch Data from the Fire Department of New York City (FDNY). This dataset includes comprehensive information on EMS operations, such as incident timestamps, dispatch areas and incident locations, and response times. Updated annually, it includes data spanning from October 18, 2016, to the most recent updates (October 16, 2024), providing a robust foundation for analyzing trends in emergency medical service (EMS) operations.

To preprocess the data for the optimization model, we first filtered the dataset to include only incidents occurring in Manhattan by selecting rows with ZIP codes between 10001 and 10282. We mapped each dispatch area code, such as M1, to its corresponding zip codes for better geographical alignment. Next, we ensured data quality by retaining rows with valid response and dispatch time indicators. After narrowing down the data, we extracted key features necessary for our analysis. Examples of the included columns are `INCIDENT_DATETIME` to capture temporal information, `INCIDENT_RESPONSE_SECONDS_QY` (call-to-response time), and indicators like `SPECIAL_EVENT_INDICATOR`. Notably, for categorical variables like `INITIAL_SEVERITY_LEVEL_CODE` and `INITIAL_CALL_TYPE`, we focused on their initial states rather than post-evaluation updates, as these initial values align with the inputs that a dispatch strategy would typically rely upon. These features were then used to simulate ambulance deployment strategies and optimize the minimization of patient waiting times.

For our initial optimization model, we grouped the data by these pairs of incident and dispatch areas and extracted the maximum recorded response time (`INCIDENT_RESPONSE_SECONDS_QY`) for the response time between them. For pairs with no recorded response times, indicating rare or nonexistent dispatches between those areas, we assigned a very large placeholder value. This approach ensures that such pairings are effectively deprioritized in the optimization model, reflecting their impracticality in real-world scenarios. To estimate the demand for ambulance services, we used data from the most recent hour for each dispatch area; we counted the number of incidents per zip code within the relevant hour. To estimate ambulance capacity in each dispatch area, we analyzed hourly dispatch data grouped by `YEAR_MONTH_DAY_HOUR` and `INCIDENT_DISPATCH_AREA`. Focusing on 2023–2024, we identified the maximum number of incidents handled by each dispatch area across the hours, representing its peak capacity. The combination of hourly demand at the zip code level and capacity at the dispatch area level allowed us to build a model for optimal ambulance allocation, targeting the minimization of response times and better resource management.

3 Methods

We explored different optimization approaches for ambulance dispatch in Manhattan: a min-sum model, a min-max model, and robust versions of both. Each model aims to improve emergency response times while addressing different aspects of system performance and reliability. We also performed sensitivity analysis to arrive at useful insights.

3.1 Basic Models

Our primary objective is to reduce emergency response times effectively. This can be approached in two different ways: first, by minimizing the total response time across all incidents, which aligns with the "min-sum" model; and second, by focusing on minimizing the maximum response time for any single incident.

3.1.1 Min-Sum Model

The min-sum model focuses on minimizing the total response time across all incidents, aiming to achieve better system-wide efficiency.

$$\min \sum_i \sum_j S_{ij} \cdot wait_{ij}$$

such that

$$\sum_j S_{ij} \leq C_i \quad \forall i \quad \text{(station capacity)}$$

$$\sum_i S_{ij} \geq d_j \quad \forall j \quad \text{(demand satisfaction)}$$

$$S_{ij} \geq 0 \quad \text{and integer} \quad \forall i, j$$

Variable Definitions

- S_{ij} is the number of ambulances dispatched from station i to location j
- $wait_{ij}$ is the response time from station i to location j
- C_i represents the capacity (number of ambulances) at station i
- d_j represents the number of incidents at location j

Rationale: This model optimizes overall system efficiency by minimizing the sum of all response times. However, it may lead to service inequities, as it does not explicitly address worst-case scenarios.

3.1.2 Min-Max Model

The min-max model focuses on minimizing the total response time across all incidents, aiming to consider worst-case scenarios.

$$minmaxz_{ij}wait_{ij}$$

such that

$$\sum_j S_{ij} \leq C_i, \quad \forall i \quad \text{(station capacity)}$$

$$\sum_i S_{ij} \geq d_j, \quad \forall j \quad \text{(demand satisfaction)}$$

$$S_{ij} \geq 0 \quad \text{and integer} \quad \forall i, j$$

$$z_{ij} \in \{0, 1\} \quad \forall i, j$$

Variable Definitions

- S_{ij}: Integer, the number of ambulances sent from station i to incident j.
- z_{ij}: Binary, indicates whether at least one ambulance is sent from i to j.
  - $z_{ij} = 1$ if $S_{ij} \geq 1$
  - $z_{ij} = 0$ if $S_{ij} = 0$

And we do the linearization:

$$\min t$$

such that

$$t \geq z_{ij} \cdot \text{wait}_{ij} \quad \forall i, j$$

$$\sum_j S_{ij} \leq C_i \quad \forall i$$

$$\sum_i S_{ij} \geq d_j \quad \forall j$$

$$S_{ij} \leq z_{ij} \cdot M \quad \forall i, j$$

$$S_{ij} \geq z_{ij} \quad \forall i, j$$

$$S_{ij} \geq 0 \text{ and integer}, \quad z_{ij} \in \{0, 1\} \quad \forall i, j$$

Variable Definitions

- S_{ij} and z_{ij} has the same definition as above
- t is the auxiliary variable representing the maximum response time

Rationale: This model ensures equitable service distribution by minimizing the worst-case response time, but it may result in slightly higher average response times compared to the min-sum model.

3.2 Robust Optimization Models

The robust optimization models for ambulance dispatch in Manhattan address real-world uncertainties by incorporating variations in station capacities and incident demands. These models use uncertainty sets with budgets C and D to control deviations from nominal values, allowing for adjustable levels of conservatism. The robust min-sum model focuses on overall system efficiency while accounting for potential fluctuations. The robust min-max model emphasizes service equity by minimizing the maximum response time with added protection against uncertainties. By balancing optimization for likely scenarios with preparation for deviations, these models aim to achieve solutions that perform well under both normal and adverse conditions. This approach leads to more reliable emergency response systems in Manhattan's dynamic urban environment, ensuring timely access to critical care across all communities.

Uncertainty Sets

Capacity Uncertainty Set:

$$\mathcal{U}_C = \left\{ C_i : C_i = \bar{C}_i - \delta_i^C, |\delta_i^C| \leq \Delta C_i, \sum_i \frac{|\delta_i^C|}{\Delta C_i} \leq \Gamma_C \right\}$$

Demand Uncertainty Set:

$$\mathcal{U}_D = \left\{ d_j : d_j = \bar{d}_j + \delta_j^D, |\delta_j^D| \leq \Delta d_j, \sum_j \frac{|\delta_j^D|}{\Delta d_j} \leq \Gamma_D \right\}$$

Where:

- $\bar{C}_i$ and $\bar{d}_j$ represent nominal capacity and demand values
- δ_i^C and δ_j^D are deviations from nominal values
- ΔC_i and Δd_j are maximum possible deviations.
- Γ_C and Γ_D are uncertainty budgets ($0 \leq \Gamma_C \leq |I|$ and $0 \leq \Gamma_D \leq |J|$).

3.2.1 Robust Min-Sum Model

$$\min \sum_i \sum_j S_{ij} \cdot wait_{ij}$$

such that:

$$\sum_j S_{ij} \leq \bar{C}_i - \Gamma_C \Delta C_i \quad \forall i$$

$$\sum_i S_{ij} \geq \bar{d}_j + \Gamma_D \Delta d_j \quad \forall j$$

$$S_{ij} \geq 0 \text{ and integer } \forall i, j$$

3.2.2 Robust Min-Max Model

$$\min t$$

such that

$$z_{ij} \cdot \text{wait}_{ij} \leq t \quad \forall i, j$$

$$\sum_j S_{ij} \leq \bar{C}_i - \Gamma_C \Delta C_i \quad \forall i$$

$$\sum_i S_{ij} \geq \bar{d}_j + \Gamma_D \Delta d_j \quad \forall j$$

$$S_{ij} \leq z_{ij} \cdot M \quad \forall i, j$$

$$S_{ij} \geq z_{ij} \quad \forall i, j$$

$$S_{ij} \geq 0 \text{ and integer } \forall i, j$$

$$z_{ij} \in \{0, 1\} \quad \forall i, j$$

$$t \geq 0$$

3.3 Sensitivity Analysis

To propose staffing and ambulance allocations to stations, we conducted sensitivity analysis on both the min-sum and min-max models by examining the shadow prices (dual values) of the capacity constraints. These shadow prices reveal the marginal effect of increasing the capacity of ambulances at each station. Specifically, for stations where the shadow price is nonzero, adding one ambulance can significantly decrease the waiting time, as the shadow price quantifies the reduction in waiting time per additional ambulance.
To make informed suggestions, we analyzed stations with non-zero shadow prices by plotting the relationship between the number of ambulances (x-axis) and waiting time (y-axis). This plot helps identify a "knee point," the point of diminishing returns, where additional ambulances still reduce waiting time but at a slower rate. Using this analysis, we recommend the levels of staffing for each station to balance resource allocation and operational efficiency.

4 Results

4.1 Basic Model Key Findings

In this section, we will discuss the key findings of our basic model. Table 1 shows the number of ambulances dispatched from each station using Min-Sum and Min-Max models. For both models, station M1 and M4 have met the full capacity, which guides us to conduct sensitivity analysis in Section 3.3.

Station Code	**Station Capacity**	**# of Ambulance Dispatched**	
		Min-Sum Model	**Min-Max Model**
M1	5	5	5
M2	26	1	12
M3	28	2	15
M4	9	9	9
M5	16	6	0
M6	8	0	0
M7	20	0	2
M8	11	9	1
M9	16	5	1

Table 1: # of dispatches by station for Min-Sum and Min-Max Models

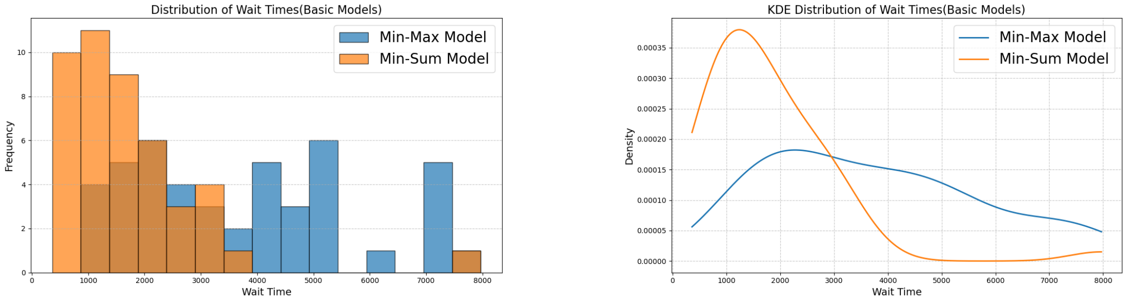
Table 2 compares the wait times for the Min-Sum and Min-Max models. The total wait time for the Min-Sum model is 80,059 seconds, whereas for the Min-Max model, it is 169,691 seconds. This highlights that the Min-Sum model prioritizes minimizing the overall wait time across all incidents. Interestingly, the maximum wait time for both models is identical at 7,968 seconds. At first glance, this may seem counterintuitive, as one might expect the Min-Max model to produce a smaller maximum wait time. However, this outcome is due to the incident in zipcode 10013, which consistently experiences significantly high wait times regardless of which station dispatches ambulances. Among all available options for zipcode 10013, station M8 provides the shortest wait time, which is 7,968 seconds. This explains why both models ultimately choose station M8 to dispatch an ambulance to the incident in zipcode 10013. This also suggests that for zipcode 10013, it may be necessary to establish new ambulance stations nearby to significantly reduce the wait time for this area.

	Min-Sum Model	**Min-Max Model**
Total Wait Time(s)	80059	169691
Max Wait Time(s)	7968	7968
Min Wait Time(s)	362	1230

Table 2: Comparison of Wait Times for Min-Sum and Min-Max Models

Figure 1 shows the distribution of wait time for the two models. The Min-Sum Model (orange bars) has a higher concentration of incidents with shorter wait times (less than 2,000 seconds). This highlights the model's goal of minimizing the overall (or total) wait time by focusing on reducing shorter wait times across all incidents. The Min-Max Model (blue bars) spreads the wait times more evenly, with higher frequencies in the middle to longer ranges (e.g., 3,000–5,000 seconds). This is consistent with the model's focus on ensuring no incident has an excessively high wait time.

In this specific case, due to the unique situation of zipcode 10013, the Min-Sum model clearly outperforms the Min-Max model. Both models result in the same maximum wait time, meaning the Min-Max model fails to improve the longest wait time. However, the Min-Sum model achieves a significantly smaller total wait time, making it the better choice in this scenario.



(a) Wait time Distribution

(b) KDE of wait time

Figure 1: Basic Model: Distribution of wait time

4.2 Comparison of Basic & Robust Models

The basic Min-Sum and Min-Max models were enhanced to address uncertainty in demand and capacity using robust optimization techniques. A detailed analysis of ambulance capacity and wait time distribution under these models was conducted, with the results presented in the Appendix. This section focuses on highlighting the key differences between these optimization approaches.

Distinctive Model Behaviors

- The basic Min-Sum model exhibits aggressive optimization with a sharp concentration of incidents under 2000 seconds, while its robust counterpart shows a more spread-out distribution with a gentler peak. This difference reveals how robustness sacrifices immediate efficiency for reliability, increasing the total wait time from 80,059 to 155,888 seconds.
- A fascinating pattern emerges in station utilization: the basic models show extreme allocation tendencies (either full capacity or zero ambulances), while robust models maintain a minimum threshold across all stations. This is particularly evident in Station M7, which goes from 0-2 ambulances in basic models to 14-16 in robust models, demonstrating how uncertainty consideration fundamentally alters resource distribution strategy.

Performance Trade-offs

- The robust models reveal an intriguing bimodal distribution pattern, particularly in the Min-Max version, with a secondary peak around 5000-6000 seconds. This suggests a deliberate load-balancing mechanism that accepts some medium-duration waits to prevent system-wide failures, a feature absent in the basic models' more optimistic allocations.
- Both model types maintain the same maximum wait time of 7,968 seconds, but their approaches to handling this constraint differ dramatically. The basic models achieve this through concentrated resource allocation, while robust models distribute resources more evenly, suggesting different strategies for managing worst-case scenarios.

System Resilience Characteristics

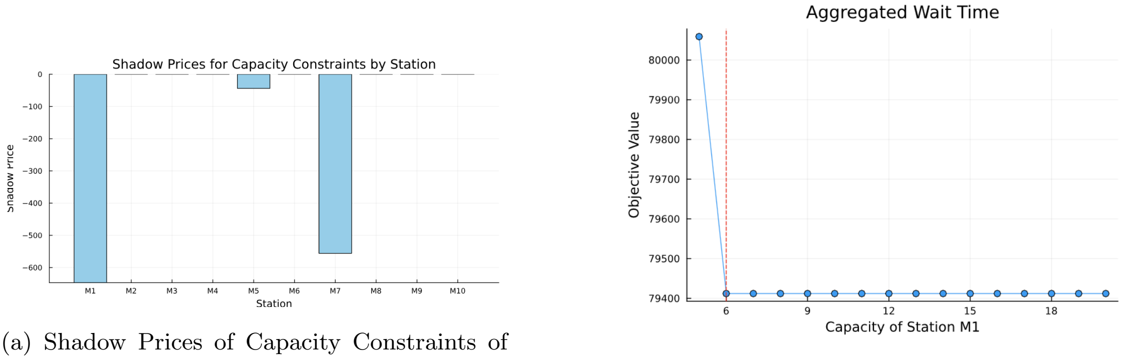
- The robust models demonstrate sophisticated risk mitigation through their station allocation patterns. While basic Min-Sum heavily favors stations M1 and M4, robust models ensure no station operates below a critical threshold, creating a more resilient network capable of handling demand fluctuations.
- An unexpected finding is that despite higher total wait times, robust models show more consistent performance across different demand scenarios. This is evidenced by their broader KDE distributions and more balanced station utilization, suggesting that operational stability comes at the cost of immediate efficiency but provides better long-term service reliability.

This comparative analysis reveals that while basic models optimize for ideal conditions, robust models create a more resilient system capable of handling real-world uncertainties, albeit at the cost of increased average wait times. The trade-off between efficiency and reliability becomes particularly evident in the distinct patterns of resource allocation and wait time distributions.

4.3 Sensitivity Key Findings

In the sensitivity analysis for the mean sum model, we find that the shadow prices for most stations are zero, except for Station M1, Station M4 and Station M6 (Figure 2.a), where the shadow prices are -647, -44 and -551, respectively. This indicates that these three stations are critical bottlenecks, as increasing their capacity would significantly reduce the overall waiting time. In the sensitivity analysis graphs for these stations (Figure 2 and 3), we plotted the aggregated waiting time against their capacity. We observed distinct knee points: Station M4 has a knee point at a capacity of 6, Station M4 has a knee point at a capacity of 11, while Station M6's knee point is 15. These findings align with the capacity vector we have, where the capacities for Station M1, M4, and M6 are only 5, 9, and 8, respectively, which are much smaller than the capacities of other stations.

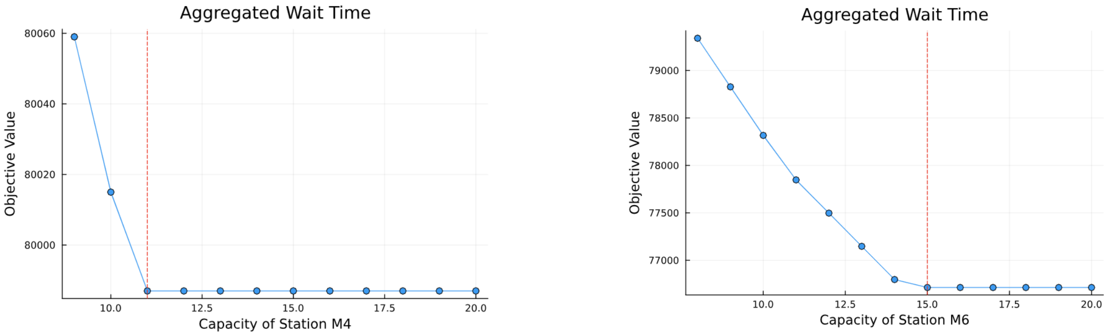
This analysis suggests that increasing the ambulance capacity for Station M1 from 5 to 6, for Station M4 from 9 to 11, and for Station M6 from 8 to 15 would effectively alleviate bottlenecks and minimize waiting times, leading to an overall improvement in the system's efficiency.



(a) Shadow Prices of Capacity Constraints of Stations

(b) Station M1 Capacity Constraints

Figure 2: Sensitivity Analysis Overall & M1



(a) Station M4 Capacity Constraints

(b) Station M6 Capacity Constraints

Figure 3: Sensitivity Analysis M4 & M6

However, in the sensitivity analysis for the min-max model, we observed that the shadow prices for all stations were zero, indicating that increasing the capacity of any single station does not reduce the maximum waiting time. This outcome makes sense because when minimizing the maximum waiting time, the optimization focuses on equalizing the waiting time across all stations, rather than reducing the total waiting time. In this model, increasing the capacity at one station simply shifts the bottleneck to another station with a lower capacity. Since the max-wait objective is constrained by the station with the worst-case waiting time, adding capacity to a single station does not improve the overall maximum waiting time unless a systemic redistribution of capacity across multiple stations is considered.

5 Impact and Conclusion

The optimization models of ambulance dispatchis designed to minimize emergency waiting times, which is critical for saving lives as every minute of delay can significantly reduce survival chances in life-threatening situations. While many models aim to minimize average wait times (min-sum), we incorporate a min-max approach to address social equity concerns. As the min-sum model generally achieves lower average waiting times, but it could exhibit significant outliers with excessively long wait times, which can exacerbate social inequalities by disproportionately affecting underserved areas. In contrast, we expect the min-max model to prioritize reducing the longest wait times, potentially ensuring a more equitable distribution of emergency response services. The application of these models to large-scale emergency data could balance efficiency and equity in public service.

The wait times for the robust Min-Sum and Min-Max models indicate that the total wait time is significantly lower for the Min-Sum model (80059 seconds) compared to the Min-Max model (169691 seconds), as expected. However, surprisingly, the maximum wait time is the same for both models at 7,968 seconds (approximately 132 minutes). This contradicts our expectation that the Min-Max model would outperform the Min-Sum model in minimizing the longest wait times. Upon further investigation, we identified ZIP code 10013 as a critical factor contributing to this outcome. This area, covering parts of Chinatown, SoHo, and Little Italy in Manhattan, exhibits extremely high waiting times due to its dense business activity, high population, and potentially unsafe conditions, making it a high-risk neighborhood requiring additional resources. This analysis reveals that the model not only serves as a template for emergency response optimization but also provides actionable insights into specific areas, such as the need for targeted interventions in ZIP code 10013. By reallocating resources to this area, we could significantly improve equity and emergency response outcomes.

Furthermore, to address uncertainties in demand and capacity, we added robustness to both the Min-Sum and Min-Max models, resulting in significant differences in resource allocation and performance. While the basic Min-Sum model achieves aggressive optimization with sharply concentrated wait times under 2000 seconds, its robust counterpart sacrifices immediate efficiency for reliability, spreading out the wait time distribution and increasing the total wait time from 80,059 to 155,888 seconds. Another key observation is the distinct station utilization patterns: the basic models exhibit extreme allocation tendencies (e.g., zero or full capacity at stations), whereas robust models ensure no station operates below a critical threshold, fostering a more resilient network. Despite both models maintaining the same maximum wait time of 7,968 seconds, the robust models results in better long-term stability, as seen in the robust models' broader station utilization and more balanced performance across demand scenarios. While robustness increases average wait times, it significantly enhances the system's resilience, providing valuable insights into the trade-offs between efficiency and reliability in real-world emergency response optimization.

Additionally, based on our min-avg and min-max models, we conducted a sensitivity analysis to evaluate how capacity constraints at each station impact system performance, specifically analyzing the shadow prices associated with these constraints. Shadow prices reveal the marginal benefit of increasing capacity at a particular station in reducing overall or maximum waiting times. This analysis allowed us to identify critical bottleneck stations where additional staffing or resources would have the greatest impact on improving response times. For instance, Station M1 has a shadow price of -647, suggesting that increasing its capacity from 5 to 6 would significantly reduce waiting times. Similarly, Station M4, with a shadow price of -44, shows a knee point at a capacity of 11, indicating that increasing its capacity from 9 to 11 would alleviate congestion. Station M6, with a shadow price of -551, exhibits a knee point at a capacity of 15, suggesting that increasing its capacity from 8 to 15 would dramatically enhance system performance. Using these granular insights, we provided targeted recommendations for reallocating ambulance capacity and staffing to optimize station performance and enhance system-wide efficiency and equity.

Future Work

This study underscores the critical role of optimization in enhancing Manhattan's EMS response system. By comparing min-sum, min-max, and robust models, we demonstrate the trade-offs between efficiency, equity, and resilience. While the min-sum model optimizes total response times, the min-max and robust approaches ensure equitable and reliable service delivery under uncertainty. Strategic staffing adjustments informed by sensitivity analysis provide actionable recommendations to alleviate bottlenecks, enhancing system performance. Looking forward, integrating real-time data on traffic conditions, incident severity, and special events will refine the model's applicability and enhance its dynamic, data-driven decision-making capabilities. These advancements will ensure Manhattan's EMS system remains adaptive to evolving urban challenges. Additionally, implementing multi-stage decision-making, which considers both initial and updated information for certain call types (e.g., initial and end-time data), would optimize dynamic resource allocation. Including contextual indicators like special events or traffic conditions would further enhance the model's adaptability to real-world scenarios, such as high-demand periods or geographically localized incidents. By embedding equity considerations into operational strategies, these extensions exemplify a forward-thinking approach to urban healthcare management, setting a benchmark for global EMS optimization.

7 Appendix

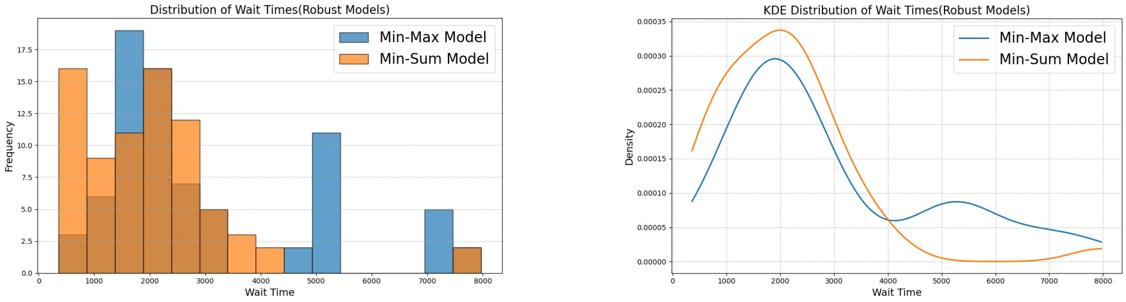
7.1 Robust Model Key Findings

Station Code	Station Capacity	# of Ambulance Dispatched	
		Min-Sum Model	Min-Max Model
M1	5	4	4
M2	26	3	9
M3	28	8	8
M4	9	7	7
M5	16	12	10
M6	8	6	6
M7	20	16	14
M8	11	8	8
M9	16	12	10

Table 3: # of dispatches by station for Robust Min-Sum and Robust Min-Max Models

	Min-Sum Model	Min-Max Model
Total Wait Time(s)	155888	225587
Max Wait Time(s)	7968	7968
Min Wait Time(s)	478	362

Table 4: Comparison of Wait Times for Robust Min-Sum and Robust Min-Max Models



(a) Wait time Distribution

(b) KDE of wait time

Figure 4: Robust Model: Distribution of wait time