# Fairness Overtime for Ambulance Allocation

Discrete Optimization for Urban Planning and Mobility
(2025SP, ORIE 5213)

Andrew Park (abp73@cornell.edu)
Leihao Fang (lf439@cornell.edu)
Nina Wang (nw364@cornell.edu)

**Abstract**

We study the ambulance allocation problem with the objective of minimizing the fairness gap, the maximum difference in coverage between any two zones over time. Using an Integer Linear Programming framework, we implement and analyze a base model, a binarized model using unary expansion, and a consistency model that limits ambulance movements between periods. Our approach introduces a unified configuration method that considers all base stations simultaneously, better aligning with real-world operations. Experimental results demonstrate that our models effectively improve fairness while accounting for operational constraints and computational efficiency.

## 1 Introduction

Ensuring equitable access to emergency services is a critical challenge in urban planning. Ambulance resources are often limited and uneven coverage can cause disparities in emergency response times. This report addresses the ambulance allocation problem by minimizing the fairness gap—the maximum difference in coverage—between zones over multiple time periods. We investigate how different optimization models and practical constraints affect the trade-off between fairness and operational efficiency, providing insights into effective and implementable resource allocation strategies.

## 2 Background and Problem Formulation

Efficient and fair allocation of ambulance resources requires balancing coverage across zones while taking into account practical constraints such as limited fleet size and operational restrictions. The *fairness gap*, defined as the maximum coverage difference between any two zones, serves as our main equity metric.

We formulate the problem as an Integer Linear Program (ILP) with the following components:

## Objective Function

$$\min\ G = \max_{i,j,t} \left|\text{Coverage}_{i,t} - \text{Coverage}_{j,t}\right| \tag{1}$$

This objective minimizes the fairness gap $G$, which represents the maximum difference in coverage between any pair of zones $i, j$ across all time periods $t$.

## Decision Variables

- $x_{i,t} \in Z_{\geq 0}$: Number of ambulances assigned to zone $i$ at time $t$.
- $q_k \in Z_{\geq 0}$: Number of times configuration $k$ is selected.
- $z_{k,t} \in \{0, 1\}$: Binary variable indicating whether configuration $k$ is selected at time $t$.

## Constraints

We define the following parameters in our model:

- $C_i$: Minimum required coverage for zone $i$
- $M$: Total number of available ambulances
- $a_{ik}$: Binary parameter indicating whether configuration $k$ covers zone $i$

1. **Coverage Constraints:**

$$\sum_k a_{ik} \cdot z_{k,t} \geq C_i, \quad \forall i, t \tag{2}$$

   Ensure each zone $i$ receives at least the required coverage $C_i$.

2. **Ambulance Availability:**

$$\sum_i x_{i,t} \leq M, \quad \forall t \tag{3}$$

   Total ambulances allocated cannot exceed the available fleet size $M$.

3. **Movement Consistency Constraints:**

$$|x_{i,t} - x_{i,t-1}| \leq \delta, \quad \forall i, t \tag{4}$$

   Limit ambulance movements between consecutive periods to at most $\delta$.

4. **Initial Deployment Constraints:**

$$x_{i,0} \leq \text{BaseCapacity}_i, \quad \forall i \tag{5}$$

   Ambulances must be initially deployed from base stations within their capacity limits.

5. **Binarization via Unary Expansion:**

$$q_k = \sum_{m=0}^{T} m \cdot z_{k,m}, \quad \sum_{m=0}^{T} z_{k,m} = 1, \quad z_{k,m} \in \{0,1\} \tag{6}$$

Integer decision variables $q_k$ are decomposed into binary variables using unary expansion to improve tractability.

This unified configuration approach jointly considers all base stations rather than analyzing them independently, better reflecting the realities of integrated emergency response systems.

## Custom Constraints that were Used with Unary Expansion

To support the joint optimization framework under unary expansion, we introduced several custom constraints in addition to standard ones. These constraints govern frequency selection, configuration linkage, movement control, and fairness.

1. **Frequency Encoding (Unary One-Hot):**

$$\sum_{k=0}^{F} z_{k,t}^{(c)} = 1, \quad \forall c, t \tag{7}$$

Each configuration $c$ must be assigned exactly one frequency value at time $t$.

2. **Configuration-Frequency Linkage:**

$$q_{c,t} = \sum_{k=0}^{F} k \cdot z_{k,t}^{(c)}, \quad \forall c, t \tag{8}$$

This links the integer-valued variable $q_{c,t}$ to the binary frequency encodings $z_{k,t}^{(c)}$.

3. **Total Configuration Count Constraint:**

$$\sum_{c} q_{c,t} = M, \quad \forall t \tag{9}$$

Ensures the total number of ambulances deployed equals the system capacity $M$ at every time step.

4. **Base-Level Ambulance Allocation:**

$$n_{i,t} = \sum_{c} q_{c,t} \cdot a_i^{(c)}, \quad \forall i, t \tag{10}$$

Computes the number of ambulances at base $i$ by summing contributions from active configurations.

5. **Flow Conservation for Ambulance Movement:**

$$n_{i,t+1} = n_{i,t} + \sum_j m_{j \to i,t} - \sum_j m_{i \to j,t}, \quad \forall i, t \tag{11}$$

Guarantees conservation of flow: ambulances arriving minus leaving equals the change in base occupancy.

6. **Movement Limit Constraint:**

$$\sum_{i \neq j} m_{i \to j,t} \leq \Delta, \quad \forall t \tag{12}$$

Limits total ambulance reassignments between time periods to promote operational stability.

7. **Zone Coverage Enforcement:**

$$y_j \leq \sum_{c,t} b_j^{(c)} \cdot q_{c,t}, \quad \forall j \tag{13}$$

Calculates the total coverage $y_j$ for zone $j$ across all time periods based on active configurations.

8. **Fairness Constraint:**

$$z \geq y_i - y_j, \quad \forall i, j \tag{14}$$

The fairness variable $z$ (our objective value) represents the maximum difference in coverage between any two zones, which we aim to minimize.

These constraints support a unified decision space over configurations, enable time-explicit control of ambulance movements, and maintain fairness and feasibility under unary-based formulation.

# 3 Project Experiment Setup

## Dataset Selection

We initially selected dataset `50-3004-6-7-35` arbitrarily, as it appeared first in the data folder, with small data points for fast iterations. During step 4, we observed that the model did not reallocate ambulances, but achieved an optimal solution immediately without requiring any movements. This result led our group to switch to another dataset to produce meaningful analysis of ambulance reallocation and consistency constraints.

To address this limitation, we transitioned to the more complex dataset `50-9085-6-7-35` **starting from step 4** in our experimental scenarios mentioned below, where fairness objectives could only be achieved through specific amount of ambulance reallocation. This enabled meaningful evaluation of movement patterns, fairness-efficiency trade-offs, and the effectiveness of consistency constraints.

### Experimental Scenarios

Using the dataset `50-3004-6-7-35`, we conducted the following experiments:

- **Base Model (Step 1):** Analyze fairness outcomes without consistency constraints.
- **Full-Scale Experiments (Step 2):** Evaluate fairness and computational efficiency across the full set of configurations.
- **Binarization and Unary Expansion (Step 3):** Apply unary expansion to improve model tractability and compare solution quality and runtime performance.
- **Additional Consistency Constraints for Unary Expansion Model (Step 4):** Introduce limits on ambulance movements between periods to analyze the impact of operational constraints on fairness outcomes for `50-3004-6-7-35`.

During step 4, we realized that the dataset was reaching its minimum fairness gap even with extremely limited number of configurations, without any ambulance movements. This meant that a specific number of ambulance allocation made the fairness gap minimized across time periods of 6.

As such, we switched to use dataset `50-9085-6-7-35` at this point. We conducted step 4 once more with this new dataset:

- **Additional Consistency Constraints for Unary Expansion Model (Step 4):** Introduce limits on ambulance movements between periods to analyze the impact of operational constraints on fairness outcomes for `50-9085-6-7-35`.

# 4 Methods

We implemented our models in Python using the Gurobi optimization solver for solving the ILP formulations. Configuration generation was performed to enumerate feasible ambulance allocations across bases, and experiments were structured into four progressive modeling steps:

- **Base Model:** The initial model minimizes the fairness gap without applying consistency constraints or advanced variable transformations. This serves as a baseline for evaluating fairness outcomes.
- **Unary Expansion and Binarization:** To improve solver efficiency and reduce symmetry in the ILP, we applied unary expansion to decompose integer variables into binary variables, following the formulation:

$$q_k = \sum_{m=0}^{T} m \cdot z_{k,m}, \quad \sum_{m=0}^{T} z_{k,m} = 1, \quad z_{k,m} \in \{0, 1\}$$

  This reformulation enables more efficient branch-and-bound exploration during optimization.

- **Consistency Constraints for Unary Expansion Model:** We introduced several custom consistency constraints for unary expansion model as introduced on the upper section (formula 7 to 14).
- **Visualization and Analysis:** Each step's experimental results were visualized using Matplotlib, providing graphical insights into fairness-efficiency trade-offs and the effects of model complexity on computational performance.

# 5 Results and Analysis

## Step 1: Base Model – Exploratory Analysis

Using dataset `50-3004-6-7-35`, the base model achieved an optimal solution when using around 10,000 configurations. See Figure 1 for fairness gap minimization and covering increases across the zones. Also, the team explored using different batches of configurations in and gained insights with Figure 2 and Figure 3 with coverage outcomes under different batch sizes. This result validated the correctness of the initial model.

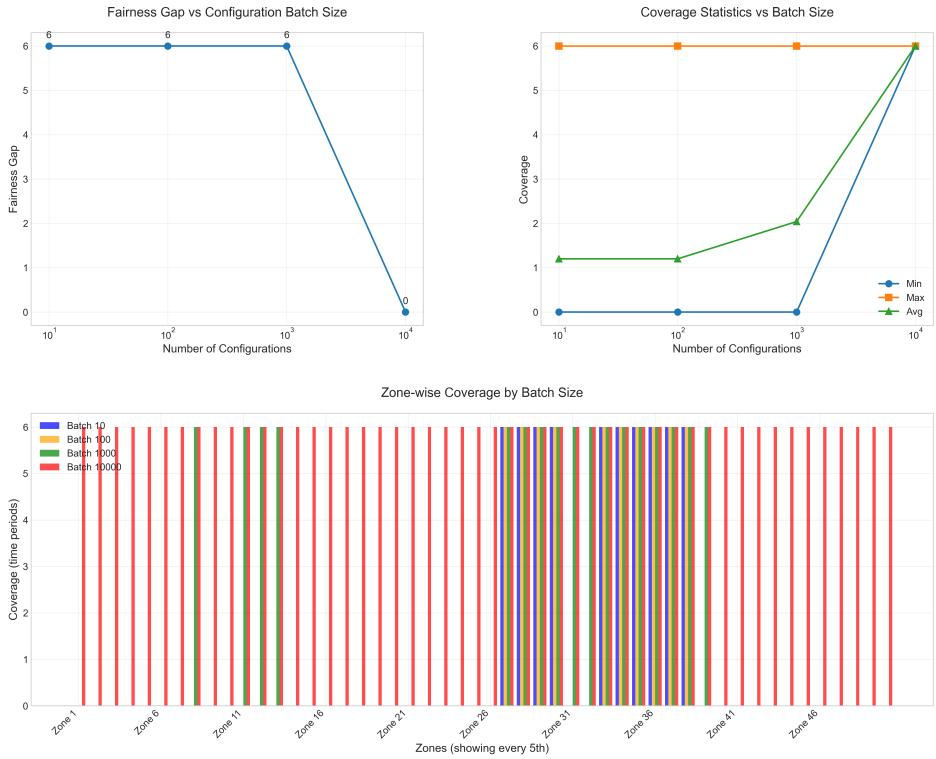


Figure 1: Fairness gap minimization and coverage distribution analysis using dataset `50-3004-6-7-35`. Left: Fairness gap decreases as configuration batch size increases. Right: Coverage statistics showing a convergence toward uniform coverage as batch size increases. Bottom: Zone-wise coverage by batch size showing the elimination of coverage disparities.

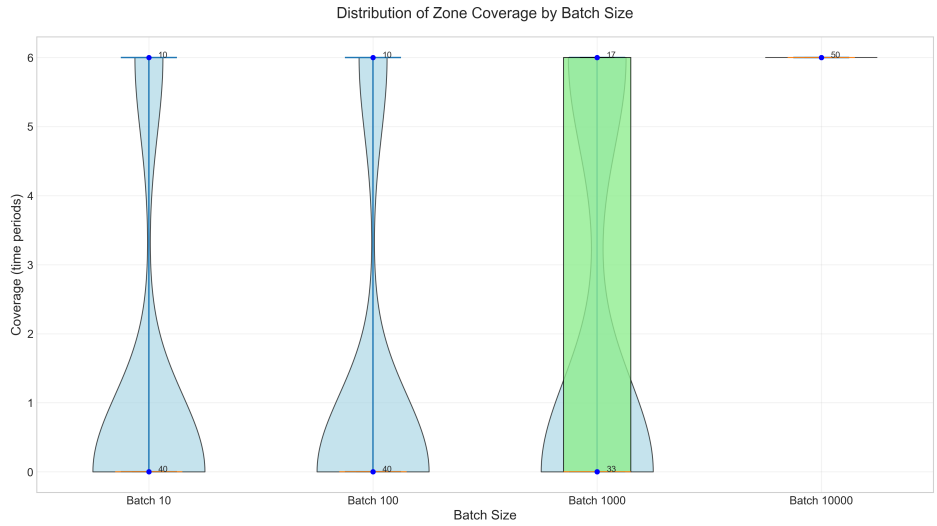


Figure 2: Zone-level coverage distribution across different batch sizes using dataset `50-3004-6-7-35`. The violin plots show how coverage distribution narrows and becomes more uniform as the batch size increases from 10 to 10,000 configurations.

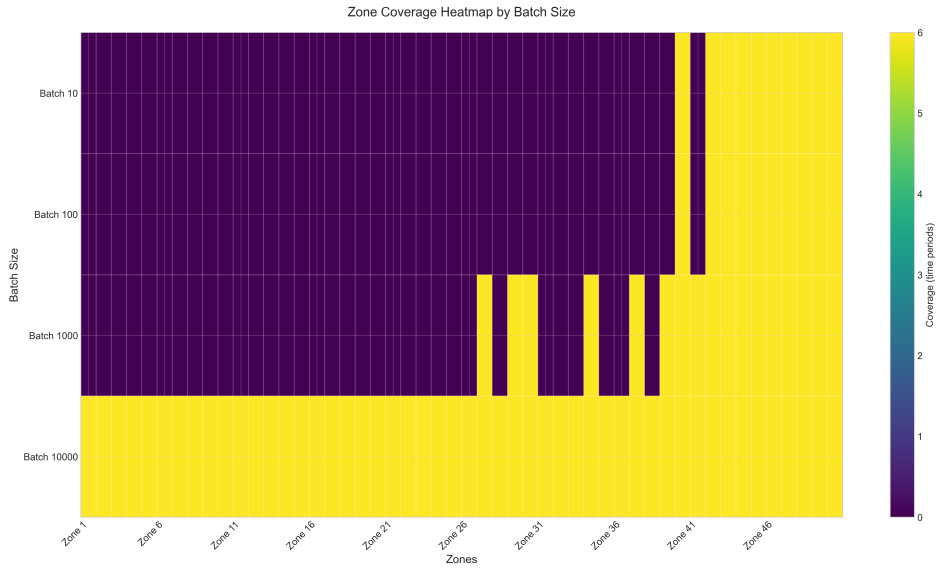


Figure 3: Zone-level coverage heatmap comparing coverage patterns across different batch sizes. The transition from dark (low coverage) to bright yellow (high coverage) illustrates how larger batch sizes provide more uniform and equitable coverage across all zones.

## Step 2: Base Model - Using All Configurations

With `50-3004-6-7-35` dataset, we evaluated fairness outcomes and computational efficiency across the full-sized column generated configurations. Figure 4 shows that the model reaches its optimal result around 10,000 configurations, which means we do not have to use all the data points with this dataset.

Figure 5 illustrates the trade-off between solution quality and runtime. As the number

of configurations increases, solution quality improves, but at the cost of higher computation time. Fairness outcomes, as shown in Figure 6, also improve significantly when larger time periods are considered.

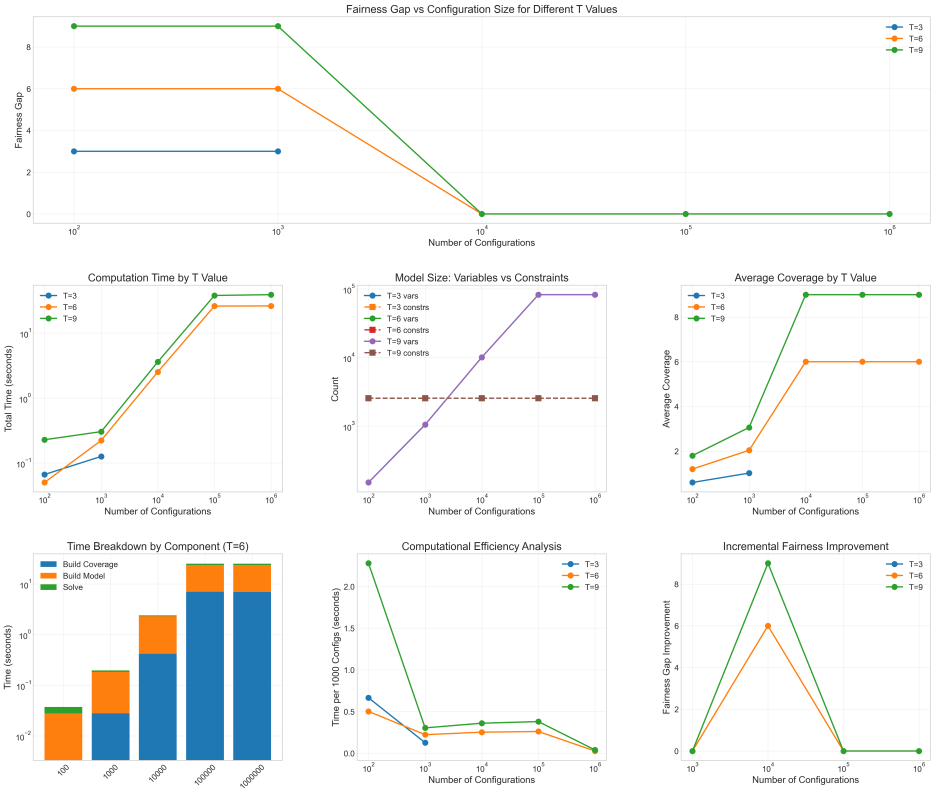


Figure 4: Comprehensive analysis of model performance with increasing configuration sizes for dataset `50-3004-6-7-35`. The upper-left plot shows fairness gap reaching zero at approximately 10,000 configurations. The bottom-right plot (Incremental Fairness Improvement) demonstrates that the most significant improvements occur when transitioning from 1,000 to 10,000 configurations, with diminishing returns beyond that point.

## Step 3: Binarization and Unary Expansion

Applying unary expansion reduced model symmetry and improved solver performance with dataset `50-3004-6-7-35`. As shown in Figure 7, the binarized model achieved better solution quality under comparable or reduced computation times. Surprisingly, the heatmap in Figure 8 indicated that the dataset itself was reaching an optimal result with extremely limited resources, so we started to doubt about the dataset, eventually decided to use another dataset `50-9085-6-7-35` for step 4. The radar chart in Figure 9 summarizes efficiency improvements across key performance metrics. Figure 10 provides a detailed comparison across modeling strategies.

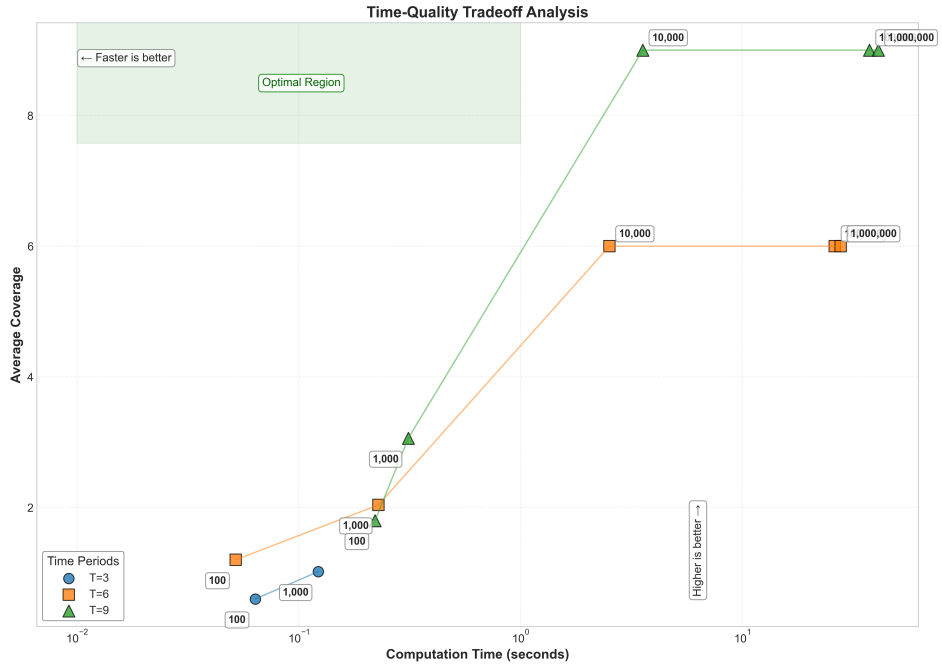


Figure 5: Trade-off analysis between computation time and solution quality across different configuration sizes and time horizons (T=3, T=6, T=9). Larger time horizons (T=9) consistently yield better average coverage but at increased computational cost.

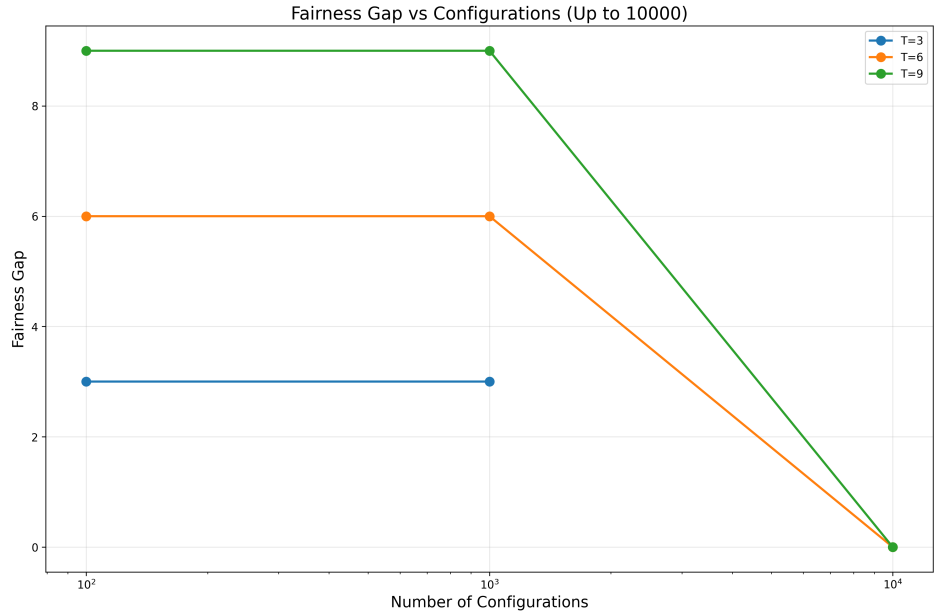


Figure 6: Fairness gap (difference in zone coverage) as a function of the number of configurations, under different time intervals ($T = 3$, $T = 6$, $T = 9$). Larger configuration sets significantly reduce the fairness gap.

## Step 4: Custom Consistency Constraints

With some doubts about the `50-3004-6-7-35` dataset, we added more custom constraints to it, and were convinced to switch to another dataset.

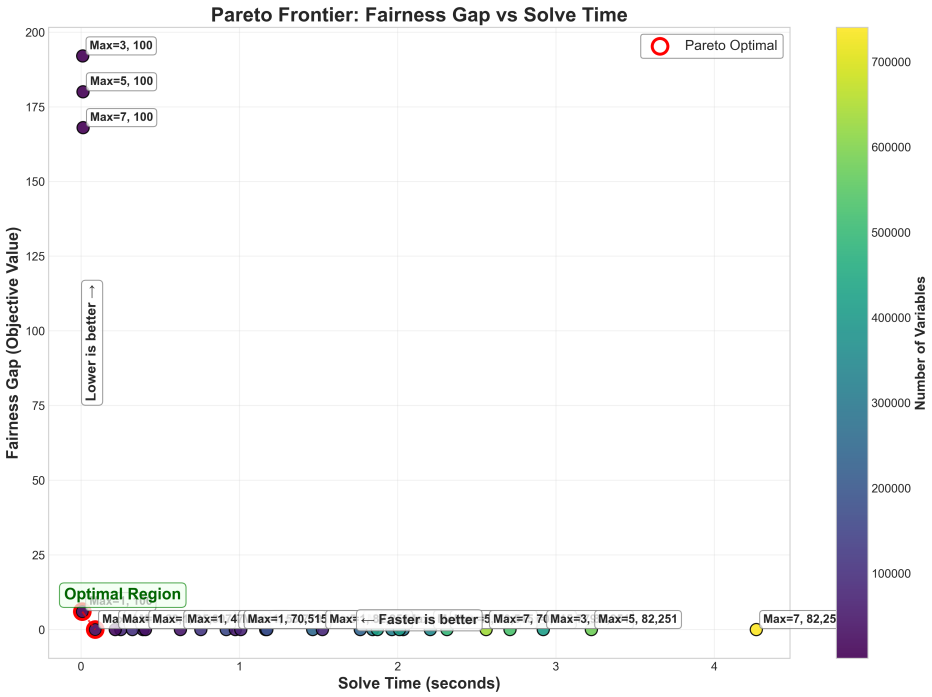


Figure 7: Pareto frontier analysis showing the trade-offs between fairness gap and solve time for different model configurations. Points on the frontier (circled in red) represent optimal solutions that cannot be improved in one dimension without sacrificing the other. Binary selection models tend to appear more frequently in the optimal region, offering better balance between solution quality and computational efficiency.

As you can see here with the result with `50-3004-6-7-35` dataset, it reached its optimal result without any re-allocations, as shown in Figure 11, Figure 12, and Figure 13.

This meant that the dataset itself had a solution with our settings (T = 6, NUM AMBULANCES = 35, MAX CONFIG FREQUENCY = 3, MAX MOVEMENT = 10), the model found an optimal solution using only 1,000 configurations without requiring any ambulance reallocation.

So that we switched to dataset `50-9085-6-7-35` and the following is the result. After establishing that the 50-3004-6-7-35 dataset resulted in optimal solutions without requiring ambulance reallocation, we transitioned to the more complex 50-9085-6-7-35 dataset to better evaluate the impact of our consistency constraints. Introducing consistency constraints stabilized ambulance allocations over time by limiting excessive movement. Figure 16 shows how fairness and operational feasibility were balanced under these constraints. While strict movement limits slightly increased the fairness gap, they resulted in more practical and implementable allocation strategies. Figure 14 illustrates the resulting allocation patterns across time, and Figure 15 shows the movement happening between bases.

To produce a meaningful result, we switched to dataset `50-9085-6-7-35` and the following is the result. Introducing consistency constraints stabilized ambulance allocations over time by limiting excessive movement. Figure 16 shows how fairness and operational feasibility were balanced under these constraints. While strict movement limits slightly increased

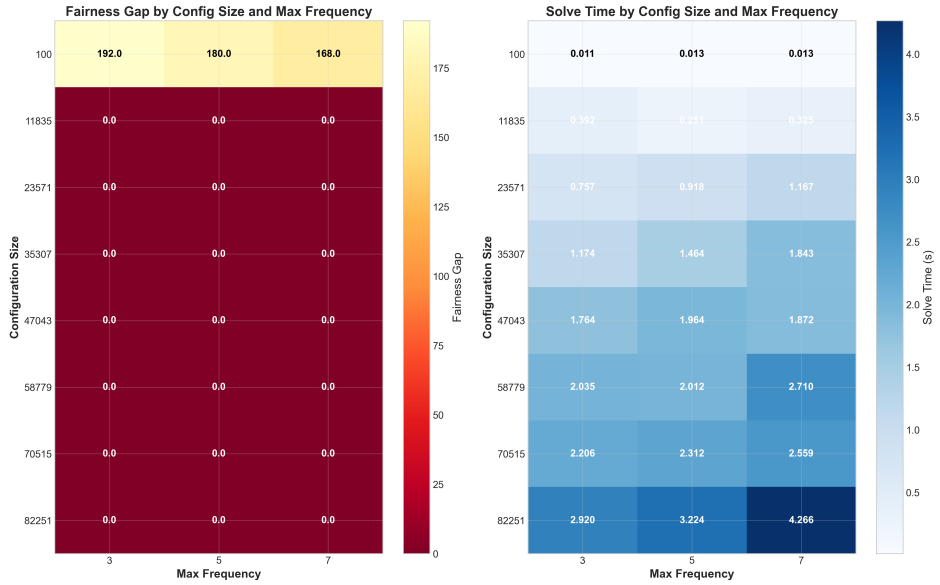


Figure 8: Performance heatmaps showing fairness gap (left) and solve time (right) as functions of configuration size and maximum frequency parameter. Note that even with only 100 configurations, T=7 achieves near-optimal fairness, indicating that the `50-3004-6-7-35` dataset has inherent structure allowing for efficient optimization with minimal resources.

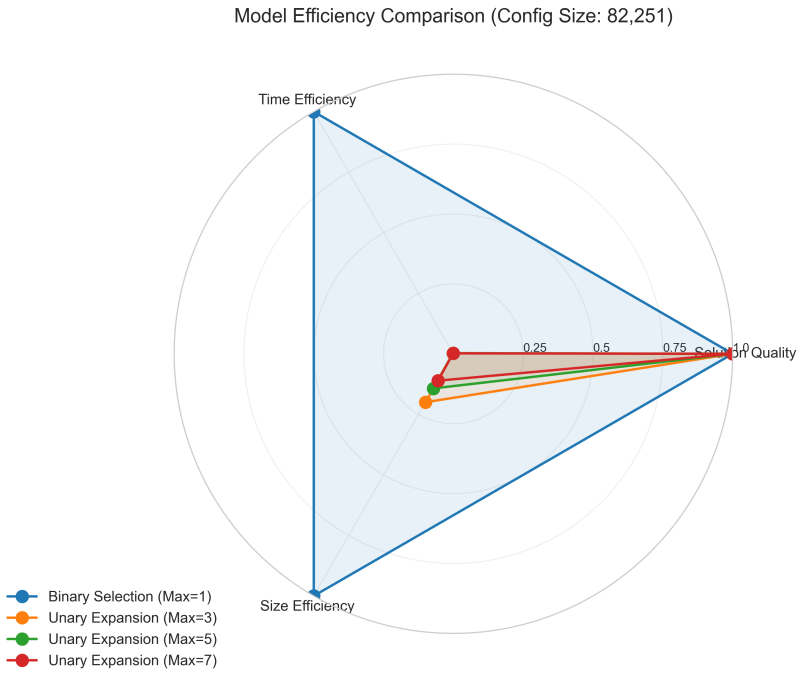


Figure 9: Radar chart summarizing model efficiency for a configuration size of 82,251. Binary selection (Max=1) shows superior performance across size, time, and solution quality dimensions.

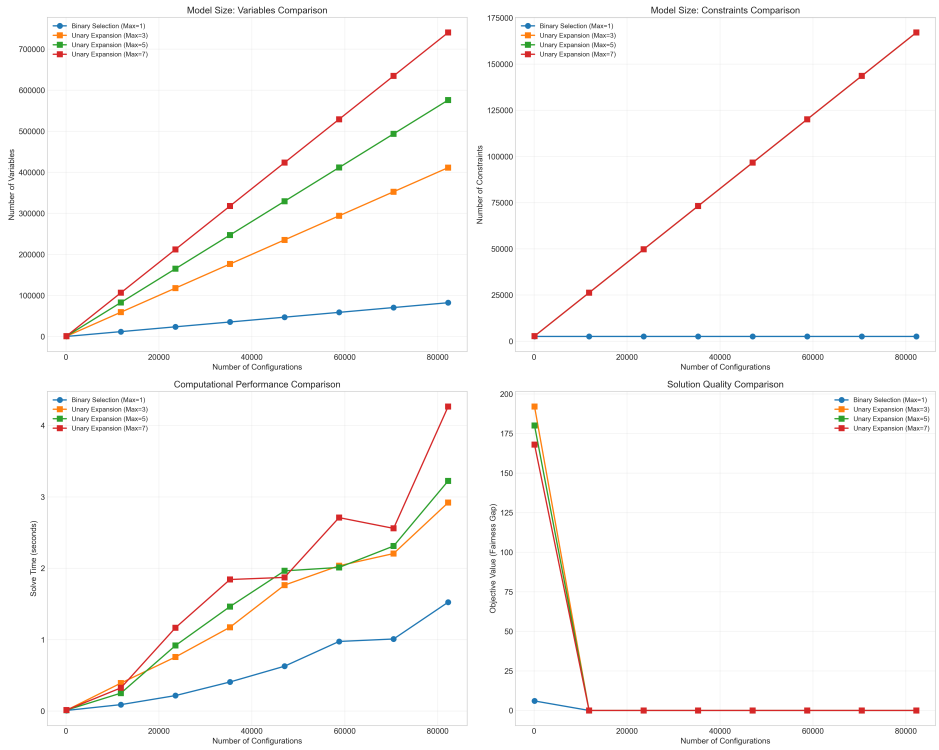


Figure 10: Comparison between binary selection and unary expansion methods across four dimensions: number of variables, number of constraints, computation time, and solution quality. Unary expansion leads to larger models but generally faster and better-quality solutions.

the fairness gap, they resulted in more practical and implementable allocation strategies. Figure 14 illustrates the resulting allocation patterns across time, and Figure 15 shows the movement happening between bases.

## Key Takeaways

- The base model alone is insufficient for real-world applications due to its inability to account for operational constraints.
- Unary expansion improves computational efficiency and solution quality by simplifying variable structures.
- Consistency constraints introduce realistic movement limitations with only a modest impact on fairness outcomes.
- The final model achieves a balance between fairness, computational efficiency, and practical ambulance deployment strategies.
- On limited occasions, the optimization can reach its best result without noticeable re-allocations, due to the dataset itself having an optimal answer, which is easy to infer with Gurobi.

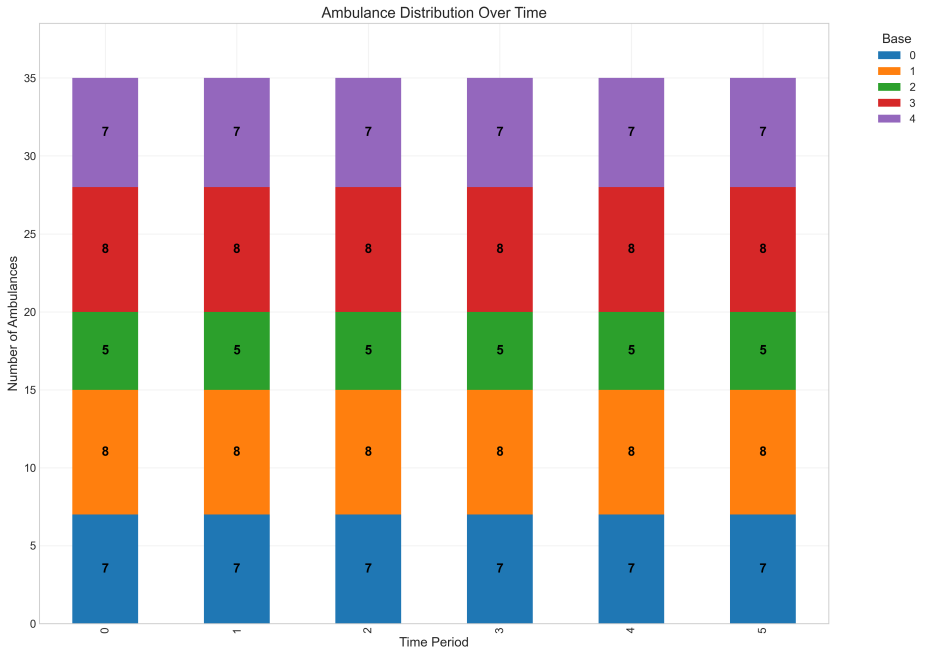


Figure 11: Ambulance distribution across bases for dataset `50-3004-6-7-35` showing identical distribution across all time periods. The consistent height of colored segments demonstrates that the optimal solution requires no ambulance reallocation over time, confirming our hypothesis about this dataset's special structure.

# 6 Discussion

Initial experimental results were unexpected, leading us to thoroughly review our modeling setup. Despite initial doubts, repeated verification confirmed the correctness of our implementation, and we recognized that Gurobi efficiently solved the scenarios provided, both for `50-3004-6-7-35` and `50-9085-6-7-35`.

While our models perform well under the tested datasets, applying them to larger instances (with larger number of bases) remains a challenge due to computational resource requirements. Our findings provide a foundation for future studies exploring fairness in resource allocation at larger scales.

# 7 Conclusion

We explored fairness-oriented ambulance allocation through progressively advanced optimization models, incorporating real-world constraints and solver efficiency techniques. Our final model balances fairness and operational practicality, providing actionable insights for emergency response planning.

Future work should extend this framework to larger datasets and explore additional operational scenarios, using this study as a reference for scalable fairness optimization.

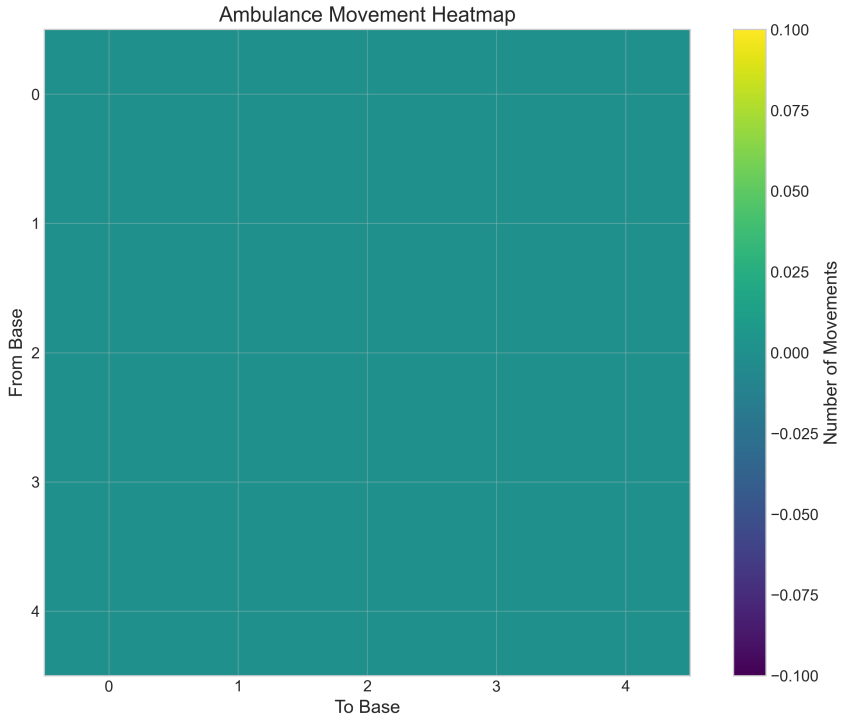


Figure 12: Ambulance movement heatmap for dataset `50-3004-6-7-35` showing uniform green color indicating zero movement between any bases. This confirms that the optimal allocation remains static throughout all time periods for this dataset.

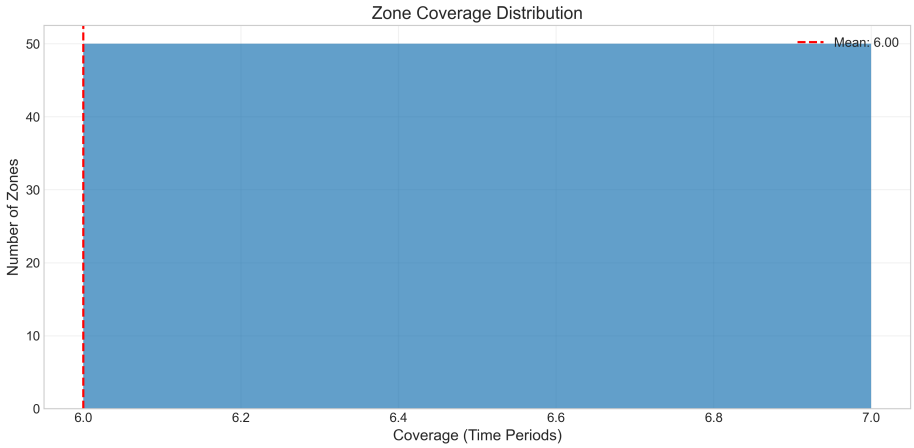


Figure 13: Summary of results under consistency constraints: fairness gap remains near zero, model size scales linearly, and solve time increases modestly. Most importantly, ambulance movements are significantly reduced, improving implementability.

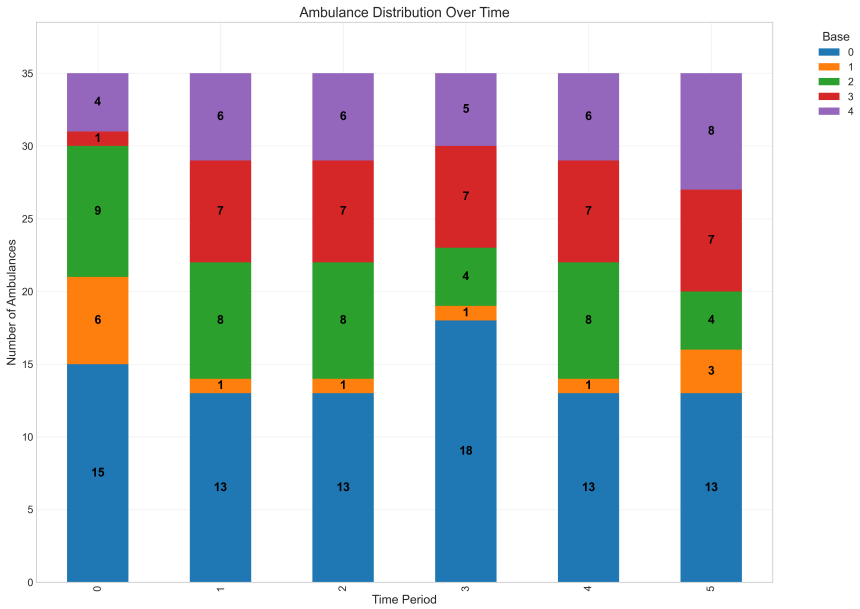


Figure 14: Ambulance distribution over time for dataset `50-9085-6-7-35` showing dynamic reallocation patterns. Unlike the previous dataset, this configuration requires strategic ambulance movements between bases over time to maintain optimal fairness. Note the changing proportions of colored segments indicating shifts in ambulance counts at each base.

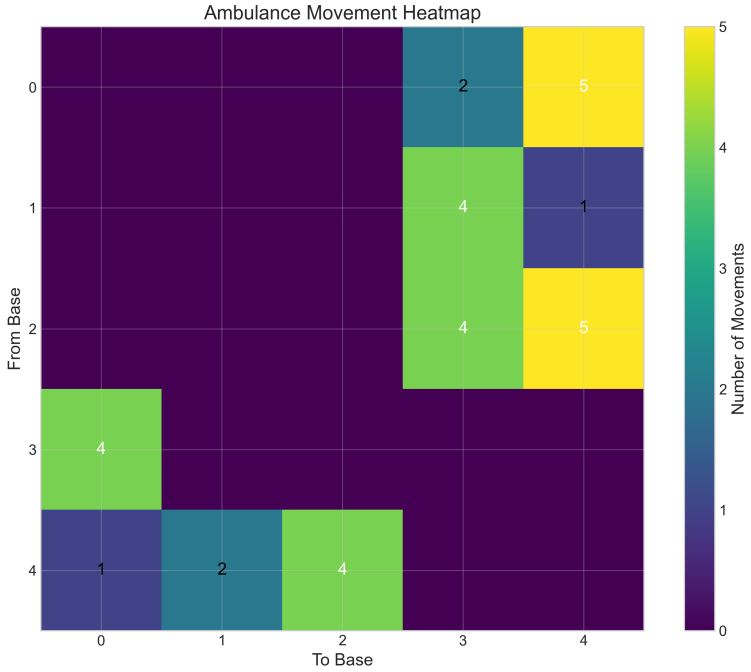


Figure 15: Ambulance movement heatmap for dataset `50-9085-6-7-35` showing significant ambulance relocations between specific base pairs. The color intensity indicates movement volume, with the highest movements occurring between bases 0-4, 2-4, and 3-2. These strategic movements are essential for maintaining fairness while respecting movement constraints.

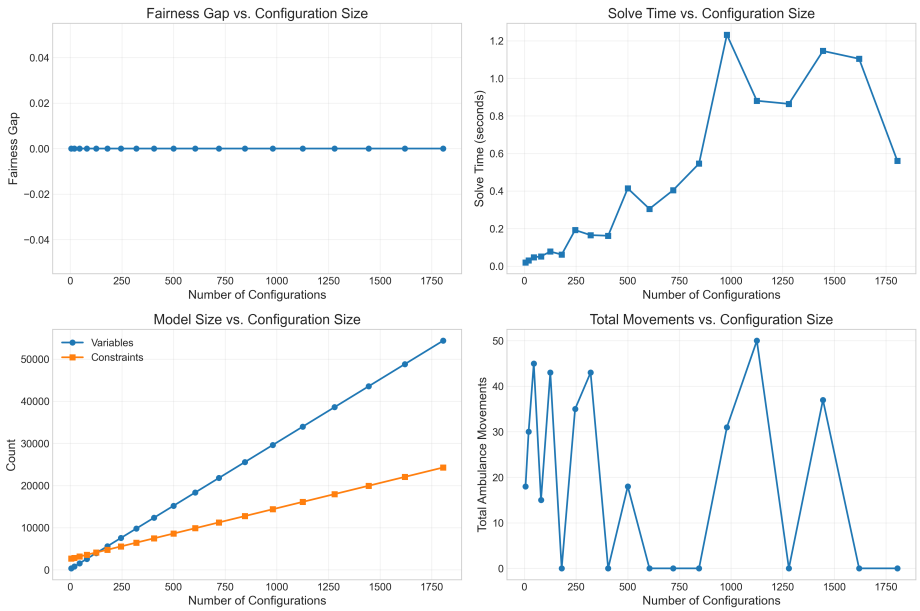


Figure 16: Performance analysis for dataset `50-9085-6-7-35` with consistency constraints. Top-left: Fairness gap remains near-zero across all configuration sizes. Top-right: Solve time scales with configuration size, showing our model's robustness. Bottom-left: Linear scaling of model size with configuration count. Bottom-right: Total ambulance movements fluctuate with configuration size, showing the optimizer's ability to find diverse movement patterns while maintaining fairness.