**Fair and Efficient Ambulance Allocation in Urban Environments**

1. **Goal of the Project**

In traditional approaches to ambulance allocation, the primary focus has been on maximizing overall efficiency—to ensure that the greatest number of patients receive timely emergency care. However, as medical students major in health policy, we were particularly intrigued by the fairness-centered perspective introduced in the paper "Fairness over time in dynamic resource allocation with an application in healthcare." The authors argue that when resources are limited and decisions are made repeatedly over time, fairness must be considered alongside efficiency to ensure that no community or zone is consistently underserved.

Our project aims to explore how the number of ambulances and the length of the decision-making time horizon impact fairness across different zones in a given region. We define fairness as the minimization of disparity in average benefits across zones over a fixed time horizo, which is introduced in class, that is minimize g-h.

In this project, we will use simulated data to study a variety of scenarios with different numbers and configurations of ambulances over different time periods. Our goal is to identify patterns and thresholds, such as the minimum number of ambulances needed in a given area to achieve near-perfect equity over time.

1. **Methodology**（Using the released model in the paper）

We build our model based on the fairness-focused framework proposed in the paper Fairness over time in dynamic resource allocation. The goal is to make sure that all zones in a region receive ambulance service as equally as possible over a given number of days (called the time horizon, T).

A group of math equations

AI-generated content may be incorrect.

In this model:

* Each zone has an "average benefit", which is how often it receives coverage during the T days.
* We calculate the average benefit by selecting different ambulance deployment plans (called configurations) and deciding how many days each one should be used.
* To achieve fairness, we aim to minimize the difference between the most-covered zone and the least-covered zone. This gap represents unfairness. A smaller gap means more equal treatment across all zones.

The optimization process ensures that:

* The average benefit for each zone stays within the limits of the best and worst values.
* The total number of configuration-days used equals T.
* The number of times a configuration is used must be a non-negative whole number (e.g., you can't use a configuration 1.5 days).

1. **Dataset Description**

We utilize a publicly available dataset from the GitHub repository PhilippeOlivier/ambulances, which represents a simplified urban environment as a weighted graph. In this graph, nodes represent city zones (or demand points), and edge weights indicate travel distances or times between zones. A subset of nodes is marked as candidate ambulance bases. Based on this dataset, we define and construct the following variables and parameters for our optimization model:

* **a\_ji (Adjacency matrix):** A binary matrix indicating whether zone i can be reached from base j. It is computed by evaluating the shortest-path distance between j and i; if the distance is below a threshold (e.g., 700 meters), then a\_ji = 1, otherwise a\_ji = 0. Each base is also assumed to cover itself.
* **r\_ij (Coverage matrix):** This matrix is generated by iterating over all feasible ambulance configurations. A configuration specifies how a given number of ambulances m are distributed across a set of bases. For each configuration j, we compute a vector indicating whether zone i is covered (i.e., at least one ambulance is placed at a base that can reach zone i). The complete matrix r\_ij is then used as input to the optimization model, serving as the benefit tau\_ij.
* **T (Time horizon):** A user-defined parameter representing the number of days or time periods to be scheduled. It defines the total number of ambulance deployment decisions to be made.
* **q\_j (Configuration usage):** A decision variable representing the number of days configuration j is used over the time horizon T. It is constrained such that the sum of q\_j over all configurations equals T.
* **y\_i (Average benefit):** A continuous variable representing the average coverage received by zone i across all T days. It is computed as a weighted average based on the q\_j values and the r\_ij coverage matrix.
* **g, h (Maximum and minimum benefit):** Auxiliary variables used to track the most and least served zones. The optimization objective is to minimize the gap g - h, representing the fairness gap in service levels.
* **n, m (Zones and ambulances):** The total number of zones (n) is derived from the number of nodes in the graph. The number of ambulances (m) is a controllable parameter explored experimentally to evaluate its impact on fairness.
* **k (Number of configurations):** Determined by enumerating all possible allocations of m ambulances over the candidate bases. This number can grow combinatorially but remains tractable for small m and moderate base counts.

Together, these variables and parameters define a mixed-integer linear program that schedules ambulance configurations to maximize equity in emergency service coverage across time and space.

1. **Coding**

We used Python and the Gurobi optimization solver to build and test our model. Our code is modular and allows changes in key parameters like the number of ambulances or the time horizon.

The workflow includes:

* 1. **Data Loading**: We load a city road network (graph) and base locations from a dataset.
  2. **Service Matrix Building**: We determine which zones can be reached from each base within a given distance threshold. This creates a matrix that shows which base can serve which zone.
  3. **Configuration Generation**: Based on how many ambulances we have, we generate all possible ways to assign them across the bases.
  4. **Coverage Matrix Creation**: For each configuration, we determine which zones are covered and build a matrix that captures this.
  5. **Optimization Setup**: Using Gurobi, we define variables for how many days to use each configuration, the average coverage of each zone, and the highest and lowest coverage values.
  6. **Running the Solver**: We run the model for different values of T (e.g., 1, 5, 15, 30, 60) and different ambulance counts (e.g., 2, 3, 5) to see how fairness improves over time and with more resources.

Each scenario outputs how evenly coverage is distributed, helping us understand how to balance resources and time to achieve fairness.

1. **Result Analysis**

In a baseline scenario of 100 city regions (6 candidate EMS sites), we evaluate the model's equity performance by adjusting the total dispatch time T and the number of ambulances m, using the difference in the average number of days of access to the service (i.e., the equity gap g-h) in each region as an indicator. The resultant data are presented in visual graphs in the figure and are combined with theoretical analyses and real-world insights to discuss major trends, tipping points, and the impact of city size on fairness.

* 1. **Impact of Scheduling Length T on Fairness**

A graph with a line graph and numbers

AI-generated content may be incorrect.

**Figure 1**

The trend of the fairness gap (g-h) with increasing total time T for different numbers of ambulances can be seen in Figure 1 (100 regional scenarios). It can be seen that the fairness gap decreases significantly with longer dispatch cycles, but with decreasing marginal improvement:

* At very small T (e.g., T = 1 day), some regions may not be covered at all due to the fact that only a single day of fixed deployment can be used, resulting in a fairness gap close to 1 (the optimal region receives 100% coverage and the worst region 0).
* As T increases from 1 to 5, service rotation between different regions becomes possible and the equity gap decreases dramatically (the curve in Figure 1 shows a steep drop around T=5). This is because the model allows for gradual coverage of otherwise unserved areas by changing the daily ambulance deployment points over a longer period of time.
* However, as T continues to increase, the increase in fairness levels off: for example, increasing T from 15 to 60 days still narrows the fairness gap but at a slower rate. Once the scheduling period is long enough for all regions to be covered at least once (the curve in Fig. 1 levels off around T ≈ 30 days), continuing to extend the time has limited effect on closing the gap. At this point, most of the regions have received more equal service coverage, the extra days are mainly used to fine-tune the frequency of coverage for each region, and the fairness improvement enters a phase of diminishing marginal benefits.

From a theoretical analysis, time extension provides the opportunity to compensate for resource shortfalls through rotational scheduling: even if the number of ambulances is limited, ample cycles allow services to be rotated repeatedly between regions to maximize coverage in the most disadvantaged regions. In reality, however, extending dispatch cycles indefinitely is not an efficient strategy because beyond a certain length of time, the cumulative coverage in almost all regions is close to equilibrium, and further rotations can only result in small improvements in equity. Therefore, urban planners should focus on finding cycles that are “long enough” - e.g., about one month (30 days) in this case to bring about a significant reduction in inequity - rather than pursuing ultra-long cycles to avoid unnecessary inequities. rather than pursuing extremely long cycles to avoid unnecessary scheduling complexity.

* 1. **Number of ambulances m Impact on fairness**

A chart with numbers and a number of times

AI-generated content may be incorrect.

**Figure 2**

The trend of the fairness gap (g-h) with increasing total time T and increasing number of ambulances m for different numbers of ambulances can be seen in Figure 2 (100 regional contexts):

* When the number of ambulances reaches 6 or 7, the fairness gap is 0 in all T cases, i.e., the system achieves “full fair coverage”.
* Instead, when m = 1~4, the gap value remains consistent (e.g., 0.267 or 0.25), and an increase in T does not significantly improve the gap. This is different from our intuitive understanding of “the longer the time, the fairer the fairer the fairer the fairer the fairer the fairer the fairness of the system. If the number of ambulances is insufficient, even if the scheduling time is increased, true coverage rotation cannot be achieved and the system fairness is stuck at a certain value.
* Similarly, in m = 1~4, the gap values are almost the same, indicating that the system is in the same “resource-constrained plateau” in this resource interval until it reaches 6 vehicles before jumping into the 0-gap state.

These phenomena suggest that the fairness is mainly limited by the total amount of resources rather than the scheduling period itself. For m < 5, insufficient resources prevent the model from adequately covering the rotation, and the gap is “stuck” in a plateau region; once the “critical resource threshold” is crossed (m ≥ 6), the system jumps to a fully fair state.

* 1. **Critical points, equipotential lines and marginal effects**

A graph of a graph

AI-generated content may be incorrect.

**Figure 3**

Figure 3 shows the contour plots of the fairness gap pairs T and m (100 regional contexts). The contour lines connect points with the same fairness gap under different combinations, and the colors from dark to light equally indicate high to low gap regions. A critical turn region can be clearly seen in the figure:

* In the lower left of the figure (low m, short T) are clustered dark-colored, equipotential zones with high equity gaps; while toward the upper right (high m, long T) are light-colored, almost equitable zones.
* When the contours change from dense to sparse, it indicates a change in the rate of improvement - which often corresponds to a tipping point. For example, as can be observed in the figure, when the number of ambulances increases to approximately 4 to 5, or the dispatch length exceeds 20 to 30 days, the contours begin to become noticeably sparse and parallel, implying that the marginal benefit of the equity improvement declines with any further increase in resources or time.

This phenomenon corresponds to the two key tipping points we mentioned earlier: the coverage tipping point and the frequency fine-tuning point. The coverage tipping point refers to the state reached when all areas have been served at least once, before which every additional day or vehicle significantly reduces the proportion of areas not yet covered (the narrower spacing of contours in the figure indicates that the gap is decreasing rapidly); once this point is crossed, the system enters the frequency fine-tuning phase, when inequity mainly stems from the slight difference in the frequency of service in each area, and continuing to invest in resources only slowly reduces this difference (the thinner contours indicate that the gap is decreasing slowly).

Theoretically, contour plots reveal the substitutability of time and resources and the effect of critical saturation: to a certain extent, an increase in the number of dispatch days can compensate for ambulance shortfalls, and conversely an increase in the number of ambulances can achieve a similar level of equity in a shorter period of time. However, there is a limit to this substitutability, and when a certain combinatorial threshold is exceeded, the system is close to full equity and it is difficult to significantly change the outcome by continuing to invest in either element.

For city managers, this provides an intuitive basis for strategy development: on the one hand, it is important to ensure that at least a critical level of resources is reached (at least a number of vehicles close enough to cover all EMS stations, as shown in the figure); on the other hand, the length of the rotation plan should be developed in such a way that it ends near the critical point, with the payoff for increasing the number of dispatch cycles after that point diminishing. For example, the analysis in this model implies that dispatching for about a month with a near-full-coverage number of ambulances already achieves near-optimal fairness, and that investing resources beyond that combination further smoothes out the nuances but is less cost-effective.

1. **Conclusion**

In this project, an equity-oriented ambulance allocation model is constructed to explore the effects of the number of ambulances (m) and dispatch time (T) on equity with the objective of minimizing the regional average coverage gap. The results show that extending the dispatching period reduces the fairness gap with diminishing marginal benefits; when there are not enough ambulances, the fairness improvement is bottlenecked and cannot be significantly improved by extending T; when m reaches a critical value (about 6 ambulances), the coverage is equalized, and the fairness gap is reduced to 0. Thus, it reveals that the dispatching time and resource investment are complementary and have limits in improving the fairness. It is recommended to ensure that the number of ambulances reaches the threshold required for coverage and that the dispatch cycle does not exceed the point of diminishing returns to fairness in order to balance fairness and efficiency.