Drone-Based Emergency Response Network to Opioid Overdose

Miguel Lejeune[1] and Wenbo Ma[2]

[1]Department of Decision Sciences, George Washington University, Washington D.C., USA, ✉ mlejeune@gwu.edu
[2]Department of Decision Sciences, George Washington University, Washington D.C., USA, ✉ wenboma@gwu.edu

1 INTRODUCTION

Opioid use affects about 2 million Americans and cost $78 billion in annual health care expenses. Opioid overdose is a critical condition that can cause severe neurological damage and death if not treated within minutes. The survival chance of an opiate overdose victim decreases by 10% for every minute that passes without resuscitation. Although emergency medical services (EMS) are strategically positioned to optimize access, the median response time in the US ranges from 7 to 8 minutes and can extend beyond 14 minutes in rural or high-traffic urban areas. Medical drones are emerging as a promising solution to reduce response times and enhance EMS performance. This motivates the design of a drone-based network aimed at providing timely medical treatment and at exploring its potential to improve the survival rates of overdose victims.

2 METHODOLOGY

2.1 Description

We present a new network design queuing-optimization model to assist EMS operators in responding swiftly and efficiently to opioid overdoses through the drone-based delivery of naloxone. The drone network is represented as a collection of $M/G/K$ queuing systems subjected to congestion. The model determines the optimal locations for drone bases (DB) and drones, as well as the dispatch strategies for responding to overdose incidents. Our model introduces two novel features. First, we investigate $M/G/K$ queuing systems where the number of mobile servers K (system capacity) is a decision variable. The ability to adjust the capacity of the system is a critical aspect. Integrated models that combine decisions on the location and on the number of drones can significantly improve the use of the EMS limited resources without affecting their performance. Second, unlike previous studies that assume that the average service time and rate are fixed and known, our model treats the service time, and thus the queueing delay and response time, as functions depending on the location and assignment decisions. This approach allows for capturing the decision-dependent uncertainty of service times, with expected values calculated endogenously, and can be applied to other mobile server networks (e.g., ambulances) where travel time constitutes a significant portion of the service time.

2.2 Model

The base formulation **B-IFP** of the proposed drone network design model is a nonlinear integer problem which includes fractional, polynomial, exponential, and factorial terms. Due to space limit, we present some key components of **B-IFP** as follows. The objective function (1a) minimizes the network-wide expected response time and its formulation is derived from the characteristics of $M/G/K$ queueing systems. The nonlinear constraints (1b) ensure the stability (steady-state) of the queueing system as they prevent the arrival rate at any DB from being equal or larger than the service rate.

$$\mathbf{B-IFP}: \min \frac{1}{\sum_{l \in I} \lambda_l} \Bigg[\sum_{i \in I} \sum_{j \in J_i} \frac{\lambda_i d_{ij} y_{ij}}{v} + \sum_{i \in I} \sum_{j \in J_i} \frac{\lambda_i y_{ij} \sum_{l \in I_j} (\lambda_l y_{lj} \mathbb{E}[S_{lj}^2])(\sum_{l \in I_j} \lambda_l y_{lj} \mathbb{E}[S_{lj}])^{K_j - 1}}{2(K_j-1)!(K_j - \sum_{l \in I_j} \lambda_l y_{lj} \mathbb{E}[S_{lj}])^2 \big[\sum_{n=0}^{K_j-1} \frac{\sum_{l \in I_j} \lambda_l y_{lj} \mathbb{E}[S_{lj}])^n}{n!} + \frac{(\sum_{l \in I_j} \lambda_l y_{lj} \mathbb{E}[S_{lj}])^{K_j}}{(K_j-1)!(K_j - \sum_{l \in I_j} \lambda_l y_{lj} \mathbb{E}[S_{lj}])} \big]} \Bigg] \tag{1a}$$

$$s.to \quad \sum_{i \in I_j} \lambda_i y_{ij} < K_j \frac{\sum_{i \in I_j} \lambda_i y_{ij}}{\sum_{i \in I_j} \lambda_i y_{ij} \mathbb{E}[S_{ij}]}, \quad j \in J \tag{1b}$$

2.3 Reformulation and Algorithm

Problem $\mathbf{B-IFP}$ is a very complex nonconvex integer optimization problem. We derive a lifted mixed-integer linear programming (MILP) reformulation. The reformulation method is generalizable and can be used to obtain an MILP representation of any optimization models minimizing the average network response time of a series of interdependent $M/G/K$ queueing systems with unknown and variable number of servers K.

We design a outer approximation branch-and-cut algorithm to solve the lifted MILP reformulation of **B-IFP**. The algorithm involves the dynamic incorporation of valid inequalities and optimality cuts, and uses lazy constraints to attenuate the challenges posed by the large size of the lifted constraint set. The computational study shows that the algorithm is computationally efficient and scales well. It can solve all problem instances (50) to optimality in one hour while Gurobi only solves 10% of those (see Section 3.2).

3 Data-Driven Tests and Insights

To demonstrate the applicability of the proposed approach, we conduct extensive numerical tests using publicly available real-life data from the city of Virginia Beach.

3.1 Response Time Reduction and Network Robustness

We analyze how the response time, a critical metric for EMSs, can be improved by using drones. We first carry out an in-sample analysis before cross-validating the results and assessing, using out-of-sample data, the stability of the results and the robustness of the designed network.

In-Sample analysis: On the training set with which the drone network is built, the (simulated) average response time enabled by the drone network is 1 minute and 32 seconds, as compared to 8 minutes and 56 seconds of the current ambulance network. In other words, the drone network reduces the response time by 82.9% on average

Out-of-sample analysis: We carry out a cross-validation analysis to assess how the training data-based (in-sample) networks perform on testing (out-of-sample) data not used to design the network.

The network built with the training data reduce in a striking manner the out-of-sample response time. The average out-of-sample response time is of 1 minutes and 38 seconds while it amounts to 9 minutes and 19 seconds for the ambulance network in Virginia Beach. This corresponds to a 82.4% reduction in the average response time.

3.2 Computational Efficiency

We evaluate the computational efficiency and scalability of the proposed reformulations and algorithms with respect to the size of the problem, namely the number $|I|$ of overdose incidents (OI) to which the network must respond. Performance profile plots highlight the increased benefits gained with our approaches as the size of the problem and the volume of OIs increases.

We have created ten problem types that differ in the number of OIs ranging from 50 to 500, by increment of 50: $|I| \in \{50, 100, 150, 200, 250, 300, 350, 400, 450, 500\}$. Each instance type is identified by the tuple $(q, p, |I|)$ and we have generated five problem instances for each instance type. This gives a total of 50 instances which we solve with the following approaches:

- `REFO`: Direct solution of MILP problem with Gurobi's default settings.
- `R-OA-B&C`: Outer Approximation branch-and-cut algorithm.

The two approaches are compared in terms of solution times and the number of instances for which optimality can be proven in one hour. The horizontal axis in Figure (1) indicates the running time in seconds while the vertical axis represents the number of instances solved to optimality within the corresponding running time.

The performance profile shows clearly that the proposed outer approximation branch-and-cut method `R-OA-B&C` is much faster, scales better, and allows the solution of many more instances than the direct solution method `REFO`. Figure (2) shows the average (across five instances) solution times for each instance type and associated size $|I|$. When an instance cannot be solved to optimality within one hour, the solution time is estimated to be 3600 seconds, which explains that for all instances of size $|I| \in [100, 500]$ the solution time for `REFO` is 3600 seconds. Figure (2) highlights the added benefits of `R-OA-B&C` for larger and more complex instances.

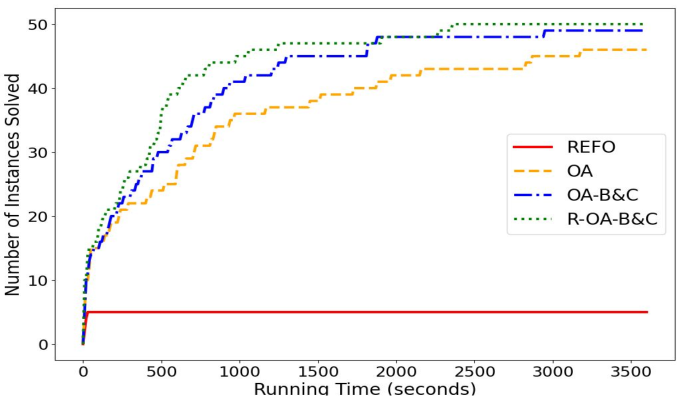

Figure 1: Performance Profile

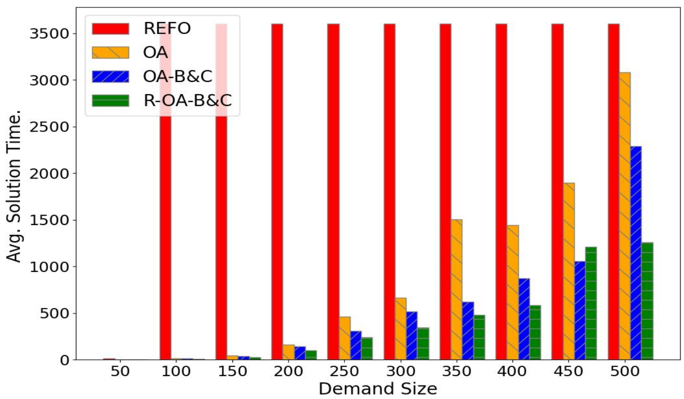

Figure 2: Average Solution Time by Demand Size $|I|$