Optimising Heterogeneous Ambulance Fleet Allocations in Jakarta

Geraint Palmer[a], Mark Tuson[a], Vincent Knight[a], Paul Harper[a], Sarie Brice[a], Leanne Smith[b], Daniel Gartner[a,c]

[a]*School of Mathematics, Cardiff University, Abacws, Senghennydd Road, Cardiff, CF24 4AG, Wales, United Kingdom*
[b]*Welsh Ambulance Services NHS Trust, Beacon House, William Brown Close, Cwmbran, NP44 3AB, Wales, United Kingdom*
[c]*Aneurin Bevan University Health Board, Lodge Road, Caerleon, NP18 3XQ, Wales, United Kingdom*

Abstract

Ambulance services have a duty of care to the clinical outcomes of the population they serve, and therefore aim to maximise the chances of survival and improve patient outcomes following a medical emergency. Despite this, although the ambulance allocation problem has been widely studied, it has predominantly focused on minimising response times or maximising coverage alone, and not explicitly for considering patient outcomes. In this paper we propose a modelling approach to consider where to best allocate different types of emergency response vehicles in order to maximise patient outcomes within a heterogeneous population. To achieve this, we develop a metaheuristic algorithm for finding better fleet allocations which is used in conjunction with a discrete-event simulation model of ambulance services with heterogeneous vehicles. A major contribution of this metaheuristic is the numerical solution of a system of equations to approximate the utilisation of vehicles. Traditionally this utilisation is problematic as it is both an input and an output of the allocation of vehicles. Our approach is informed by, and tested on, real-world data from Jakarta, Indonesia. Using our developed models, decision makers are better able to understand ambulance fleet capacity needs and allocations, and their impact on patient outcomes.

Keywords: Ambulance Planning, Emergency Medical Services, Simulation, Health Care, Optimisation, Mathematical Modelling

1. Introduction and Background

Time-critical conditions (TCCs) are a substantial cause of mortality worldwide, responsible for an estimated 54% of all deaths [18]. Emergency medical services (EMS) play a pivotal role in responding to TCCs and saving patients from life-threatening medical conditions. Most EMS research, especially planning to meet increasing ambulance demands, tends to be focussed on high-income countries including, for example, Australia [29], Germany [46], and USA [10]. In contrast, many low and middle-income countries (LMICs) lack an organised EMS system, with most ambulances used purely for patient transport and not as an emergency care vehicle [41]. However, the burden of deaths due to TCCs is much greater in LMICs than in high-income countries, the difference being around threefold [14]. For traffic accidents, strokes and heart attacks, rapid access to appropriate treatment is especially vital.

The application of our research is focused on Indonesia, supported by an award for Global Challenges Research Funds (GCRF). In Indonesia, an LMIC, there has been very little prior work on developing an EMS strategy, driven by inadequate a lack of financial investment and human capital [41, 37, 49]. Related research by the authors [13] has highlighted particular challenges in Indonesia as a lack of a single coordinated EMS, the ability or willingness to pay for an ambulance, a large geographical area, and areas of severe traffic congestion especially in Jakarta, the capital.

When our research programme commenced in October 2019, ambulance services in Jakarta were provided by many disparate, mostly private providers that charge patients for their use. We initially partnered with Ambulans 118, a non-government charitable ambulance service established in 2005 by the Indonesian Surgeons' Association, that currently operates in five cities across Indonesia: Jakarta, Palembang, Yogyakarta, Surabaya and Makasser. Unlike private providers, Ambulans 118 suggests that a donation is made for use of its EMS vehicles, but otherwise provides free emergency medical care, it is also leading on paramedic training across the country.

The overarching goal of our research was to work with the Indonesian Government, the different ambulance providers, and hospitals, to help them forecast emergency demand and make critical decisions on the best types, capacities and geographical allocations of emergency vehicles within a potentially co-ordinated EMS system, starting with Jakarta. To facilitate this, in collaboration with Ambulans 118 we organised workshops in Jakarta in February 2020, which were attended by over 170 people including doctors, nurses, paramedics, academics, and officials from the Indonesian Ministry of Health. This paper presents some aspects of the overall research programme, specifically focusing on the development of a simulation model and allocation metaheuristic for maximising patient survival.

Fundamental to the project and the research described in this paper was an initial study that we carried out in order to better understand patient needs and the barriers to use of ambulances in Indonesia. Throughout the month of December 2020, we undertook comprehensive surveys in Emergency Departments (EDs) across Jakarta and published the first known study in to EMS demand within the country [13]. Our study showed that the utilisation of ambulances by patients attending EDs is very low and of concern. The low utilisation is contributed by patients' lack of awareness of available ambulance services, patients' disinclination to use ambulances due to high costs, and long response times. All of these barriers impact on patient outcomes, especially for those with life-threatening conditions. For example, more trauma patients took a car-share ride (20%) or motorcycle (20%) to reach the ED than an ambulance (just 10%), while only 14% of critical cardiovascular patients used an ambulance compared to 67% travelling to hospital by private car or a car-share ride.

The first contribution of this paper is an extension the work of [24] to consider both heterogeneous patients and heterogeneous fleets. Here a maximal survival objective is formulated, from vehicle allocations and utilisations, which are themselves found by numerically solving a relationship on the vehicles' share of the

demand. This is then used in a metaheuristic to find improved allocations that aim to maximise patient survival. Secondly, we develop a comprehensive sequential discrete event EMS simulation that models and evaluates heterogeneous fleet allocations of emergency vehicles by utilising a novel approach in which transit jobs, rather than patients, are framed as the queueing 'customers'. Our methodology utilises two discrete event simulations of the same system that are run sequentially and together combine to form the logic of a single simulation of a heterogeneous fleet; and key performance indicators (KPIs) are calculated by making use of survival functions. Thirdly, we demonstrate the use and impact of our modelling framework applied to current and proposed ambulance allocations in the city of Jakarta, thus supporting Government-level decision making with an overall goal to improve the lives of those living in LMICs. This enables decision support and managerial insights in Indonesia, although the developed modelling framework could be readily applied to other locations.

The paper is structured as follows: Section 2 gives an overview of the literature on modelling ambulance allocations. Section 3 describes the current ambulance service behaviour and sets out the mathematical notation used throughout the paper. Section 4 describes an optimisation model, including the survival objective, utilisation considerations, and a metaheuristic algorithm to find better fleet allocations. Section 5 outlines the logic of the sequential discrete event simulation models for heterogeneous fleets. Section 6 gives a case study, describing the current emergency service situation in Jakarta and applying the optimisation and simulation results. Section 7 discusses the findings and contributions of the paper.

2. Literature Review

There is a rich literature on operational research applied to EMS location and related allocation problems. Literature reviews that include these topics are Aringhieri et al. [3], Bélanger et al. [9], Farahani et al. [17], Li et al. [27], Liu et al. [28], Reuter-Oppermann and Vile [38], Mukhopadhyay et al. [33] and Wang et al. [48]. Ambulance allocation problems such as that concerned in this paper can be categorised as a *preparedness* problem [33].

Emergency medical services can be evaluated using different performance measures with coverage and response time being the predominant metrics [31]. However, missing a response time target by just one second, for example, would be considered as a 'failure' in many models with no appreciation at all of the impact on patient survival or outcome [31]. This seems somewhat restrictive and short-sighted, which is why we focus on survival as a metric and compare and contrast papers using that metric with our approach. Here we consider some closely related papers, providing a summary of each and comparing and contrasting them with our modelling and solution approach.

Amorim et al. [2] propose an integrated strategic and tactical planning approach which features an optimisation model and a local search heuristic based on Gaussian Processes. Their methodology is applied to the city of Porto while reporting on performance metrics such as survival. Our approach is different because we use a discrete-event simulation framework to evaluate results from a population based heuristic to determine the number of ambulances required in each station.

Boutilier and Chan [11] develop an integrated location-queueing model that incorporates existing EMS response times in a drone network. They use a p-median approach and an Erlang loss model. Although survival is not their main optimisation criterion, they report how many patients would survive using their approach.

Erkut et al. [16] can be considered the first modelling approach that considers decaying survival probabilities during response time. The rationale is, because survival can be thought of as a more robust and generic objective for EMS performance measurement than coverage or average response time. In an application of a recently developed Maximal Survival Location Problem model (MSLP), the authors use data from Edmonton,

Canada and show that maximising survival is superior to other objectives in clinical outcome for cardiac arrest patients.

McCormack and Coates (2015) [30] focus on the optimisation of EMS vehicle fleet allocation and base station location through the use of a genetic algorithm (GA) with an integrated EMS simulation model. Their objective is maximization of the overall expected survival probability across patient classes. Applications of the model were undertaken using real call data from the London Ambulance Service. The difference between their modelling approach and our is that we have a different survival function, focus on a different heuristic optimisation and evaluate our approach with data from a developing country, and have higher traffic volume and congestion in our data.

Bélanger et al. [8] combine optimisation and simulation to find both ambulance locations and dispatch policies. Here, binary integer programming is used to minimise total response time, where ambulance availability, or utilisation is unknown. In this work this is overcome by iteratively feeding the solution of the optimisation programme into a simulation to find the availability parameters, until convergence. Our work differs as we solve for ambulance utilisation explicitly.

Toro-Díaz et al. (2013, 2015) [43, 44] build optimisation models for both the allocation and the dispatch, or preference, rules for EMS services. They utilise the hypercube queueing model for ambulance availability with a integer programming and metaheuristic algorithms. In particular, they consider a number of different objectives including mean response times, expected coverage [15], and the Gini index on individual response times as measures of fairness in the system.

Knight, Harper and Smith [24] provide a supporting extension of the MSLP that allows it to be applied more generally to real-world EMS systems, acknowledging that different patients have varying levels of expected survival probabilities. The Maximum Expected Survival Location Model for Heterogeneous Patients, MESLMHP, allows multiple classes of patient groups to be defined, where previously only cardiac arrest patients formed part of the objectives for primary response. In reality any emergency patient has a necessity for swift attendance, and a timely response to any incident type may impact on clinical outcome in some way. It is for this reason MESLMHP is designed to be generic enough to accommodate any number of patient groups, with each class weighted dependent upon the relative urgency of the incident. More than just a contribution to the model's demand input, these patients are included in the optimisation when maximising total population survival probability. Our work directly extends this work, to include heterogeneous vehicle fleets, while developing a different fixed-point numerical solution to the problem of approximating vehicle utilisations.

3. Problem Statement & Notation

We first present some notation. We have the following sets:

- $\mathscr{P}$ is the set of pick-up locations, indexed by p;
- $\mathscr{A}$ is the set of ambulance locations, indexed by a;
- $\mathscr{Y}$ is the set of hospitals, indexed by y;
- $\mathscr{K}$ is the set of medical specialities, indexed by k;

to incorporate heterogeneous fleets the set of specialities $\mathscr{K}$ is partitioned into two sets, $\mathscr{K}_A$, those patients that can be seen by secondary vehicles, and $\mathscr{K}_B$, those patients who are seen by primary vehicles only.

An allocation is given by the tuple $\left(Z_a, \tilde{Z}_a\right)$, where Z_a is the number of primary vehicles allocated to ambulance location a; and $\tilde{Z}_a$ is the number of secondary vehicles allocated to ambulance location a. Also let:

- λ_{pk} be the rate at which patients of speciality k make calls from pick-up location p;
- $s_k(t)$ be the survival function function associated with patients of speciality k;
- π_a be the average utilisation of primary vehicles stationed at location a;
- b_{pa} be the expected travel time from ambulance location a to pick-up location p; and let B_{pa} be a random variable representing the time it takes for an ambulance to drive this distance;
- C_{py} be a random variable representing the travel time from pick-up location p to hospital y;
- D_{ya} be a random variable representing the travel time from hospital y to ambulance location a;
- F_{pa} be a random variable representing the travel time from pick-up location p to ambulance location a;
- G_k be a random variable representing the time the ambulance spends with patients of speciality k at the pick-up location;
- J_k be a random variable representing the time the ambulance spends with patients of speciality k at the hospital;
- Θ be a random variable representing the time the ambulance spends re-fuelling, re-stocking, and resting between transit jobs;
- q_{pky} be the probability that a patient of speciality k from pick-up location p is taken to hospital y. Note that $\sum_{y\in\mathscr{Y}} q_{pky} < 1$ is possible, that is a patient may not go to any hospital, in which case the ambulance returns to their ambulance location.

To incorporate heterogeneous vehicles, let also $\tilde{\pi}_a$ denote the average utilisation of secondary vehicles stationed at location a; and $\tilde{b}_{pa}$, $\tilde{B}_{pa}$, $\tilde{C}_{py}$, $\tilde{D}_{ya}$, and $\tilde{F}_{pa}$ denote the corresponding travel times for secondary vehicles.

Patients will call an ambulance from one of the pick-up locations $p \in \mathscr{P}$. If *all* ambulances are busy, then that call will be abandoned and it is assumed that the patient will find their own way to the hospital through private or public transportation, which is an appropriate and justified assumption for the case of Jakarta [13]. Otherwise, a central control centre will dispatch the closest available primary vehicle. That vehicle will travel from its location to the patient (B_{pa}), spend some time on the scene treating the patient (G_k), travel to the hospital (C_{py}), spend some time handing over the patient at the hospital (J_k), travel back to it's original location (D_{ya}), then spend time refilling and refuelling (Θ). There is also a probability that the patient does not need a hospital, and so the vehicle will travel from the pick-up location back to it's original vehicle location (F_{pa}). These routes are shown visually in Figure 1. Note that these routes cannot be interrupted, that is, an ambulance must return to it's station for refilling before it is available to respond to another call, to ensure it is cleaned, and fully stocked for it's next call. Note also that B_{pa} and F_{pa} need not necessarily be equal, as travel times need not necessarily be symmetrical, this could be due to a number of reasons, such as non-symmetric road networks, and differences in driver urgency on each leg of the journey.

Therefore the total time an ambulance is busy will be T_{paky} given by Equation 1 if a hospital is required, and service time T_{pak} given by Equation 2 if no hospital is required.

$$T_{paky} = B_{pa} + G_k + C_{py} + J_{yk} + D_{ya} + \Theta \tag{1}$$

$$T_{pak} = B_{pa} + G_k + F_{pa} + \Theta \tag{2}$$

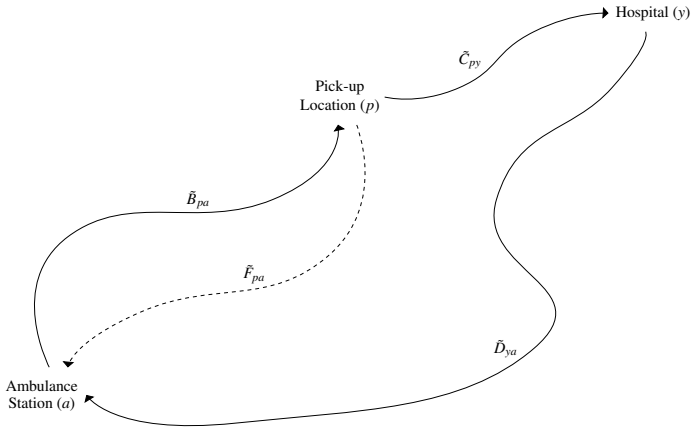


Figure 1: Ambulance routes for a given transit job.

There are two types of vehicle: primary vehicles, typically ambulances, that must be dispatched to all patients; and secondary vehicles, typically rapid response motorcycles, that can travel faster than primary vehicles, and are dispatched to some patients to respond to an emergency faster than the primary vehicle. Secondary vehicles cannot transport patients, but are used in conjunction with primary vehicles to reduce response times. From a resource cost perspective, one primary vehicle costs the same to purchase and to run as three secondary vehicles.

4. Optimising Allocations for Maximal Survival

Here we outline a method of finding vehicle allocations that maximise the expected survival of patients. The problem is, for a given set of vehicle locations (that is ambulance stations), a given total number of primary vehicles, and a given total number of secondary vehicles, how many vehicles of each type should be stationed at each location to maximise expected survival across all patients.

4.1. Survival Functions

A key concept here is the survival curve of a patient. In reality, some emergency incidents do not result in a substantive deterioration of a patient's status over time; however, in all situations, there is a reasonable cut-off beyond which a patient should not expect to have to wait for care, and mixture of theoretical survival functions and step functions can be utilised. Survival probabilities for critical incidents, calculated from a theoretical monotonically decaying survival function reported in the literature [45] are used to demonstrate an attainable level of success from a response. One particular survival curve $s(t)$ of Equation 3, represents survival until hospital discharge following cardiac arrest; its origins are explained in detail by [24], and gives the probability of survival if seen within a time t in the form of a logistic function. Figure 2 shows the difference between using this survival curve and a hard cut-off of 8 minutes. However, hard cut-off curves like Equation 4, with a cutoff of L can still be used to represent meeting artificially selected targets, for example for transportation jobs.

$$s(t) = \left(1 + e^{0.26+0.139t}\right)^{-1} \tag{3}$$

$$s_L(t) = \begin{cases} 1 & \text{if } 0 \leq t \leq L \\ 0 & \text{if } t > L \end{cases} \tag{4}$$

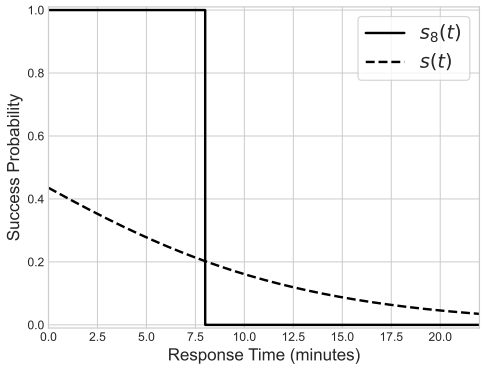


Figure 2: Survival function $s(t)$, estimated by [45] compared with $s_L(t)$ with a hard cutoff of $L = 8$, represented as a step function.

4.2. Maximal Expected Survival Location Model with Heterogeneous Patients and Heterogeneous Fleet

Here we propose a Maximal Expected Survival Location Model with Heterogeneous Patients and Heterogeneous Fleet (MESLMHPHF). It is an extension of the MESLMHP model given in [24], which did not consider heterogeneous fleets. This is a model of survival that can be used as an objective function for optimisation algorithms. It is a weighted expected survival function as the objective function which considers heterogeneous patients (different specialities) and heterogeneous fleets (both primary vehicles, EAs, and secondary vehicles, RRVs), and is constructed by appropriately summing these survival curves, described in Section 4.1, across the population, multiplying by ambulance availability where appropriate. The MESLMHPHF is given by Equation 5.

$$g\left(Z_a, \tilde{Z}_a\right) = \sum_{p \in \mathscr{P}} \sum_{a \in \mathscr{A}} \left(\sum_{k \in \mathscr{K}_A} w_k \lambda_{pk} \hat{\Psi}_{kpa} + \sum_{k \in \mathscr{K}_B} w_k \lambda_{pk} \Psi_{kpa} \right) \tag{5}$$

where w_k is a weight associated with patient speciality type k. For this study we assume $w_k = 1$ for all k, which allows the interpretation of $g\left(Z_a, \tilde{Z}_a\right)$ as the expected number of patients surviving per time unit. Now Ψ_{kpa} can be interpreted as the probability of a patient $k \in \mathscr{K}_B$ at pick-up location p being seen by a primary vehicle from location a and surviving; while $\hat{\Psi}_{kpa}$ would be the probability of patients of speciality $k \in \mathscr{K}_A$ at pick-up location p being seen by any vehicle from location a and surviving. Equation 5 is the weighted sum over the expected survival probabilities of all patient specialities, all patient pick-up locations, and all ambulance stations. These survival probabilities are given by Equations 6 and 7 respectively.

$$\Psi_{kpa} = s_k\left(b_{pa}\right)\left(1 - \pi_a^{Z_a}\right) \prod_{\alpha \in \mathscr{A}} \pi_\alpha^{\left(Z_\alpha \beta_{p\alpha a}\right)} \tag{6}$$

$$\begin{aligned} \hat{\Psi}_{kpa} = {} & s_k\left(\tilde{b}_{pa}\right)\left(1 - \tilde{\pi}_a^{\tilde{Z}_a}\right) \prod_{\alpha \in \mathscr{A}} \tilde{\pi}_\alpha^{\left(\tilde{Z}_\alpha \beta_{p\alpha a}\right)} \pi_\alpha^{\left(Z_\alpha R_{p\alpha a}\right)} \\ & + s_k\left(b_{pa}\right)\left(1 - \pi_a^{Z_a}\right) \prod_{\alpha \in \mathscr{A}} \pi_\alpha^{\left(Z_\alpha \beta_{p\alpha a}\right)} \tilde{\pi}_\alpha^{\left(\tilde{Z}_\alpha \left(1 - R_{p\alpha a}\right)\right)} \end{aligned} \tag{7}$$

Here $R_{pa_1a_2}$ and $\beta_{pa_1a_2}$ are binary variables indicating the preference of sending a vehicle from one station to another, that is the dispatch rule. Here we use the closest or fastest vehicle, so $\beta_{pa_1a_2}$ indicates if a vehicle of the same type can reach p quicker from a_1 than a_2, defined in Equation 8, while $R_{pa_1a_2}$ indicates if a

primary vehicle at a_1 can reach p quicker than a secondary vehicle at a_2, defined in Equation 9. Note that here these parameters are defined using travel times, but can be generalised to account for any type of preference, including for example proximity, importance, or efficiency.

$$\beta_{pa_1a_2} = \begin{cases} 0 & \text{if } a_1 = a_2 \\ 1 & \text{if } b_{pa_1} \leq b_{pa_2} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

$$R_{pa_1a_2} = \begin{cases} 1 & \text{if } b_{pa_1} \leq \tilde{b}_{pa_2} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Equation 6 is the probability of a patient surviving (their survival function), multiplied by the availability of primary vehicles at that ambulance station, multiplied by the unavailability of vehicles at closer stations. Equation 7 extends this to two vehicles types, the first part of the sum repeating the logic for the faster secondary vehicles, and the second part of the sum adapting that logic for primary vehicles, who will only be contribute the patient's survival if all faster secondary vehicles are also busy. These interpretations are given as annotations to the equations in Appendix A.

A key consideration is the vehicle utilisations π_a and $\tilde{\pi}_a$; discussed in the next subsection.

4.3. Utilisation Considerations

The models above assume that the utilisations of each vehicle type at each vehicle location is known, which is a potentially restrictive assumption. These utilisations π_a and $\tilde{\pi}_a$ depend on the demand to station a, which itself depends on the vehicle allocations. One method of overcoming this, used in [24], is to consider utilisation as the ratio of demand and service rates, shown in Equations 10 and 11. Here we let $\frac{1}{\mu}$ and $\frac{1}{\tilde{\mu}}$ be the average job times of primary and secondary vehicles, then λ_a and $\tilde{\lambda}_a$ represent the share of the demand seen by primary and secondary vehicles from vehicle location a, respectively.

$$\pi_a = \frac{\lambda_a}{\mu} \tag{10}$$

$$\tilde{\pi}_a = \frac{\tilde{\lambda}_a}{\tilde{\mu}} \tag{11}$$

Note that here we assume that average job times are not dependent on the vehicle location a, although this may be unrealistic, given that the travel between locations are a key part of a vehicle's job time (B_{pa}, D_{ya}, F_{pa}). However, as discussed in Section 3, there are a number of other components to an ambulance job including time on site (G_k), hospital handover time (J_{yk}), and vehicle refill time (Θ). For the case of Jakarta, we justify the assumption of job times not being location dependent by considering the proportion of an ambulance job's time that is location dependent, ρ, defined in Equation 12,

$$\rho = \begin{cases} \frac{B_{pa}+D_{ya}}{T_{paky}} & \text{if hospital required,} \\ \frac{B_{pa}+F_{pa}}{T_{pak}} & \text{if no hospital required,} \end{cases} \tag{12}$$

then, using the simulation described in Section 5, we consider the distribution of ρ across all simulated jobs under the current situation and allocation in Jakarta, shown in Figure 3. We see for primary vehicle jobs on average 24% of a job time is location dependent, while it is 36.5% for secondary vehicles. Considering

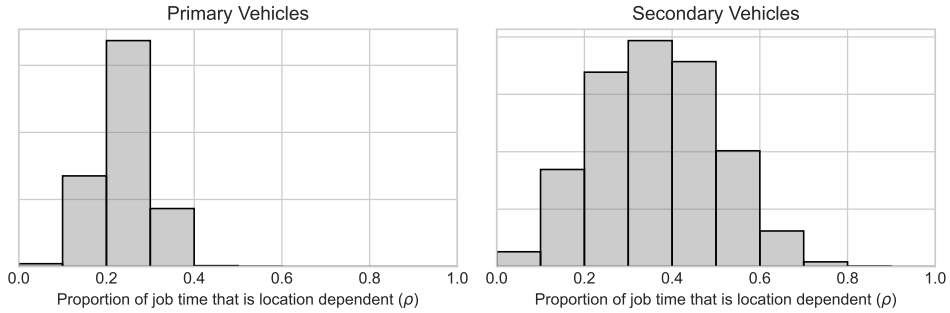


Figure 3: Distribution of the proportion of an ambulance job consists of location dependent components, for primary and secondary vehicles.

location-dependent service rates would be an interesting further research question, potentially improving the model's performance by incorporating travel times into the MESLMHPHF. However, given that in the case of Jakarta the preferences $\beta_{pa_1a_2}$ and $R_{pa_1a_2}$ are themselves defined by travel times (Equations 8 and 9), this may not have much affect for this study.

The demand experienced by vehicles at each location $a \in \mathscr{A}$ would indeed be location specific, and would depend on the number and utilisation of the vehicles at every other location. This is addressed in the next subsection.

4.4. Numerical approximation of utilisation

For primary vehicles, which operate independently of secondary vehicles, the relationship between demand experienced at each location, λ_a, and the allocations and utilisations of all other locations, is given by Equation 13. For secondary vehicles, the relationship between demand experienced at each location, $\tilde{\lambda}_a$, and the allocations and utilisations of all other locations, is given by Equation 14. Note that this also depends on the utilisations of the primary vehicles, π_a.

$$\lambda_a = \sum_{p\in\mathscr{P}} \sum_{k\in\mathscr{K}} \lambda_{pk} \left(1 - \left(\frac{\lambda_a}{\mu}\right)^{Z_a}\right) \prod_{\alpha\in\mathscr{A}} \left(\frac{\lambda_\alpha}{\mu}\right)^{(Z_\alpha \beta_{p\alpha a})} \tag{13}$$

$$\tilde{\lambda}_a = \sum_{p\in\mathscr{P}} \sum_{k\in\mathscr{K}_A} \lambda_{pk} \left(1 - \left(\frac{\tilde{\lambda}_a}{\tilde{\mu}}\right)^{\tilde{Z}_a}\right) \prod_{\alpha\in\mathscr{A}} \pi_\alpha^{(Z_\alpha R_{p\alpha a})} \left(\frac{\tilde{\lambda}_\alpha}{\tilde{\mu}}\right)^{(\tilde{Z}_\alpha \beta_{p\alpha a})} \tag{14}$$

We propose finding the true vehicle utilisations by first solving Equation 13 for the λ_a; determining the primary vehicle utilisations with Equation 10; then using these to solve Equation 14 for the $\tilde{\lambda}_a$; and determining the secondary vehicle utilisations with Equation 11. These can be solved numerically. In our implementation this is solved using the MINPACK hybrd and hybrj algorithms by using Scipy's `fsolve` function [47]. It should be noted that in this section we assume that all λ_{pk} are static and not time-dependant, which is an assumption that will be used throughout the optimisation methodology, however in the simulation we can account for time-dependent demand.

Now note that, in Equation 13, we have that

$$\sum_{p\in\mathscr{P}} \sum_{k\in\mathscr{K}} \left(1 - \left(\frac{\lambda_a}{\mu}\right)^{Z_a}\right) \prod_{\alpha\in\mathscr{A}} \left(\frac{\lambda_\alpha}{\mu}\right)^{(Z_\alpha \beta_{p\alpha a})} < 1,$$

and so $\sum_{a\in\mathscr{A}}\lambda_a < \sum_{p\in\mathscr{P}}\sum_{k\in\mathscr{K}}\lambda_{pk}$, meaning that there is lost demand. This corresponds to the patients who are abandoned, or take private or public transport to a hospital instead of waiting for an ambulance. Similarly, in Equation 14, we have that

$$\sum_{p\in\mathscr{P}}\sum_{k\in\mathscr{K}_A}\left(1-\left(\frac{\tilde{\lambda}_a}{\tilde{\mu}}\right)^{\tilde{Z}_a}\right)\prod_{\alpha\in\mathscr{A}}\pi_\alpha^{(Z_\alpha R_{p\alpha a})}\left(\frac{\tilde{\lambda}_\alpha}{\tilde{\mu}}\right)^{(\tilde{Z}_\alpha\beta_{p\alpha a})} < 1,$$

and so $\sum_{a\in\mathscr{A}}\tilde{\lambda}_a < \sum_{p\in\mathscr{P}}\sum_{k\in\mathscr{K}_A}\lambda_{pk}$. This lost demand represents both abandoned calls, but also calls where secondary vehicles are not deployed as a primary vehicle could reach them first.

4.5. Metaheuristic Optimisation

In order to maximise the MESLMHPHF objective function presented in Section 4.2, we use an evolutionary algorithm. A population of possible solutions is created and ranked according to the objective, then for each generation of the metaheuristic a proportion of the best performing solutions is kept, and are mutated to complete the population for the next generation. To encourage exploration, the number of times each solution is mutated begins high and gradually decreases throughout the run of the simulation. The metaheuristic takes five hyper-parameters: N the population size, κ the number of solutions to retain to the next generation, m_0 the initial mutation rate; c the cooling rate; and H the number of generations. The metaheuristic is described in Algorithm 1.

Algorithm 1: Evolutionary Algorithm used to find better allocations.

```
Create a population of N random allocations (Z_a, Z̃_a);
for i ← 0 to H do
    Rank the population according to Equation 5;
    Keep the top κ solutions;
    m ← ⌈m_0 − max(m_0, i × c)⌉;
    for j ← 0 to N − κ do
        Randomly choose and copy a solution from the population;
        Mutate that copy of the solution m times;
        Place mutated copy in a separate population;
    end
    Combine populations;
end
Rank the population according to Equation 5; Output the top ranking solution.
```

Mutations consist of one of four possible changes: randomly change the location of one primary vehicle; randomly change the location of one secondary vehicle; randomly remove one primary vehicle and add secondary vehicles to three randomly chosen locations (as one primary vehicle costs the equivalent to three secondary vehicles); and randomly remove three secondary vehicles and add a primary vehicle to a randomly chosen location. If it is required that exact vehicle numbers need to be retained, then only for first two mutations need be chosen.

The source code for the optimisation metaheuristic, including the evaluation of the MESLMHPHF objective function and the solution method for the utilisation considerations, is available at [25] and development took place at `https://github.com/drvinceknight/HeterogeneousAmbulanceFleetAllocations`. Numerical results are discussed in Section 6.

5. Simulating Allocations

Discrete event simulation (DES) is a common methodology that allows us to investigate given scenarios under uncertainty. It is a common methodology for modelling emergency medical services, with a review given in [1]. Here it will be used to quantify the effectiveness of given ambulance allocations by measuring a range of key performance indicators (KPIs), such as the average response time, the percentage of abandoned calls, vehicle utilisations, and expected survival based on response times. The simulation is built using the Ciw library [34], and the model logic is described below. Its central ideas include modelling transit jobs as customers, rather than the patients themselves, and simulating primary and secondary vehicles as two simulations sequentially, with the output of the former being the input of the latter. Separating out primary and secondary vehicle logic simplifies the model logic, allows for modular simulations that can be run in isolation, and so easier to adapt and to maintain.

5.1. Simulating Primary Vehicles

For primary vehicles the logic to simulate is as follows: Patients make a call from one of a number of pick-up locations and await an ambulance to pick them up and take them to an appropriate hospital. Ambulances are stationed at a number of ambulance locations; when a patient makes a call, all free ambulances from any location calculate their expected time to reach that patient, and the ambulance with the smallest expected time to the patient is called out. The ambulance drives from its current ambulance location to the patient's pick-up location, then if a hospital is required, from the patient's pick-up location to the hospital, and then from the nearest hospital back to their original ambulance location. If a hospital is not required, the ambulance returns to their original ambulance location.

Rather than considering patients as customers in a queue, here the situation is re-framed to model transit jobs as customers, with the ambulances as servers. This is similar to the approach of [23], whereas ambulances are now simple servers, their stations are nodes in a queueing network, with a routing decision representing the ambulance preference.

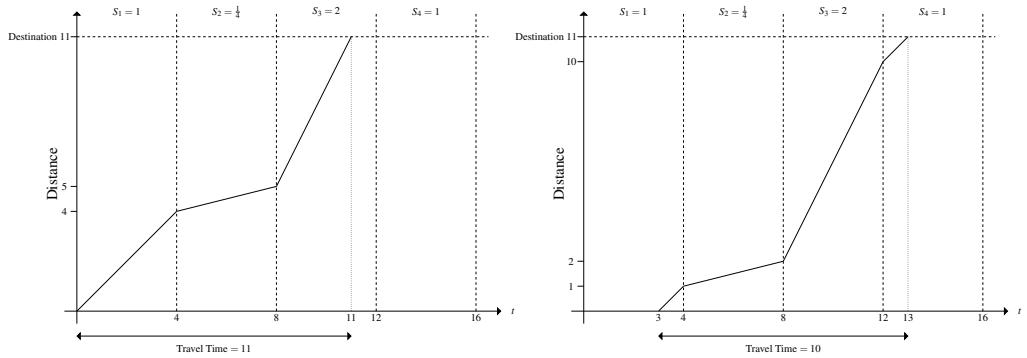
Transit jobs can be categorised into classes corresponding to the pick-up locations $\mathscr{P}$ and speciality $\mathscr{K}$. Jobs of class $(p,k) \in \mathscr{P} \times \mathscr{K}$ arrive with rate λ_{pk}. Servers can be categorised into classes corresponding to the ambulance locations $\mathscr{A}$. Service times are server-dependent, that is the service time of a transit job of type (p,k), being served by a server of class a, will have service time T_{paky} if a hospital is required and T_{pak} if no hospital is required. These were given by Equations 1 and 2 respectively, in Section 3. That is T_{paky} and T_{pak} is the overall time the ambulance spends dealing with that transit job and is unavailable to receive any more transit jobs.

From a patient's point of view their service only lasts $G_k + C_{py} + J_{yk}$, (or G_k if no hospital is required), and their waiting time is B_{pa} plus the time waiting for an ambulance to be dispatched. However in our case, if there are no free ambulances at the time of call, it is assumed that patients find their own care or transport, and so the call is abandoned. Therefore, the time waiting for dispatch is always zero.

Furthermore, all service times are time dependent, and calculated from travel distances and approximate hourly traffic levels, described in Section 5.2.

5.2. Travel Time Calculations

Travel times within a city such as Jakarta are not constant throughout the day due to the variability in traffic, and it is necessary to capture these in models of emergency medical services [40]. Here, each day is split into a set of periods, $\mathscr{H}$ indexed by h. Within each period traffic levels are modelled as piecewise linear functions similar to [22], and are here considered constant but differing from period to period. Traffic levels



(a) Beginning at $t = 0$, a travel time of 11.

(b) Beginning at $t = 3$, a travel time of 10.

Figure 4: Example of calculating travel times for two different starting times.

influence the speed of the ambulance through a delay factor d_h associated with each period. Therefore, the expected speed S_h at which an ambulance travels during period h is given by Equation 15,

$$S_h = sd_h \tag{15}$$

where s is some given baseline speed. The relationship between time travelled and distance covered is piece-wise linear with slope S_h in period h. Therefore, travel times are calculated using this relationship.

As an example, consider the scenario where for the first period $t \in (0,4)$ we have $S_1 = 1$; for $t \in (4,8)$ we have $S_2 = \frac{1}{4}$, for $t \in (8,12)$ we have $S_3 = 2$, and for $t \in (12,16)$ we have $S_4 = 1$. Consider that the vehicle must travel 11 units. If the vehicle begins its journey at $t = 0$, then the expected travel time would be 11 time units, shown in Figure 4a. However if the vehicle begins its journey at time $t = 3$ then the expected travel time is 10 time units, shown in Figure 4b.

This is the method used to calculate the expected values of B_{pa}, C_{py}, D_{ya} and F_{pa}. Furthermore it is assumed that travel times for each segment of the route (from vehicle location to patient, from patient to hospital, and from hospital back the the vehicle's location) follow a Triangular distribution around the calculated expected travel time, between 75% and 125% of the value.

5.3. Time-dependent Demand

Additionally, it is assumed that calls arrive according to a Poisson distribution, with rates λ_{pk}, and that these rates are time dependent. Each day is split into four 6-hour periods, the morning, afternoon, evening and night, and each speciality k and pick-up location p will have a different call arrival rate for each of these periods.

For some locations and specialities the number of observed calls might be very low, and so the λ_{pk} rate would be very low. This can cause synchronicity issues when sampling arrivals [36], where artificially long inter-arrival times can be introduced at the beginning of one time period due to the low arrival rate in the previous time period. In order to overcome this here, rather than sample inter-arrival times iteratively from an Exponential distribution, an entire schedule of arrival dates are sampled at the beginning of the simulation run, first by sampling the number of arrivals required in each time period from a Poisson distribution, and then by sampling specific dates within that time period using a Uniform distribution. This mechanism, and alternative to thinning [26], was implemented in the Ciw software in version v2.3.3 [42], and described in the documentation here: `https://ciw.readthedocs.io/en/latest/Guides/time_dependent.html`.

5.4. Simulating Secondary Vehicles

Secondary non-transit vehicles, can in general travel faster than primary transport vehicles, and so can be dispatched at the same time as a primary vehicle but reach the patient earlier, reducing the response time for that patient and so increasing their survival probability. They are only dispatched for patients of speciality $k \in \mathcal{K}_A$, and if the closest secondary vehicle can reach them before the closest primary vehicle. Here secondary vehicles are simulated in a second discrete event simulation, run sequentially, but simulating the exact same time period and events as the first simulation. This is similar to sequential hybrid simulation methodology [12, 32], however in this case the two combined components are both DES, and are combined to simplify the logic of each component, maintain modularity and so ease model reusability and future adaptations.

This is possible as there are no synchronicity issues between the two components. Primary vehicles operate independently of secondary vehicles, that is the way primary vehicles respond to a patient is unaffected by the presence or lack of a secondary vehicle. Therefore, primary vehicle logic is not compromised by simulating primary vehicles in isolation. Secondary vehicles are impacted by the behaviour of primary vehicles, they must remain with the patient until the primary vehicle arrives, and so must be simulated after the simulation of primary vehicles, taking as inputs the exact list of events that occurred. That is the logic of the secondary vehicles is determined by observing the actions of primary vehicles and reacting to them. Simulation results are combined for each individual patient, to determine their response time caused by either primary or secondary vehicles.

Secondary vehicles are chosen in the same way as primary vehicles, by choosing out of the currently free vehicles the one with the smallest expected time to patient. Their service times are reactionary to what occurred with the primary vehicle. The decision to send a primary vehicle or not depends on the previously estimated travel time for the primary vehicle, not the actual travel time experienced by that primary vehicle: if a secondary vehicle is estimated to reach there first, they will be dispatched. Let $\tilde{B}_{pa_2}$ be the time it takes for the secondary vehicle to travel from its location a_2 to the patient pick-up location p; B_{pa_1} is the time is took for the primary vehicle to get from its location a_1 to the patient pick-up location p; G_k is the time the primary vehicle spent with the patient; and $\tilde{F}_{pa_2}$ is the time it takes to return to the secondary vehicle's location a_2 from the patient pick-up location p. Note that B_{pa_1} and G_k are exact values obtained from the initial simulation of the primary vehicles, while $\tilde{B}_{pa_2}$ and $\tilde{F}_{pa_2}$ are random variables to sample in the subsequent simulation. Travel times are calculated as described in Section 5.2, replacing the primary vehicle delay factor, d_h, with a delay factor for secondary vehicles, $\tilde{d}_h$. This accounts for secondary vehicles travelling faster and reacting to traffic differently to primary vehicles, similar to the method used in [19] to differentiate the travel times of ambulances with and without flashing lights.

Exact synchronicity considerations are shown in Figure 5:

- whenever the secondary vehicle reaches the patient before the primary has left, vehicle, they remain with the patient until the primary vehicle leaves;
- if secondary vehicle arrives after the primary vehicle has left, they will immediately return to their stations as they are not needed at the scene;
- if the expected time for the secondary vehicle to reach the patient exceeds the expected time for the primary vehicle to reach the patient then the secondary vehicle is not deployed, and so that transit job would be abandoned in the secondary simulation (although still seen by a primary vehicle in the initial simulation). We assume that secondary vehicles must still reach the pickup locations, due to potential difficulties in communication en-route.

Therefore job service times for secondary vehicles, $\tilde{T}_{pak}$, are given by:

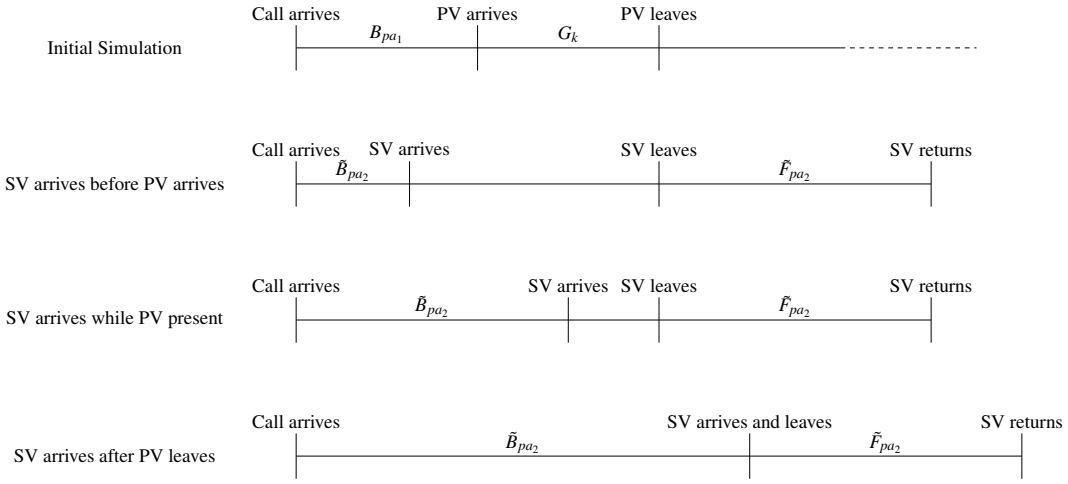


Figure 5: Visualising secondary vehicle (SV) logic reacting to the primary vehicle's (PV) actions.

$$\tilde{T}_{pak} = \max\left(\tilde{B}_{pa_2}, B_{pa_1} + G_k\right) + \tilde{F}_{pa_2} \tag{16}$$

5.5. Combined Model

Combining the logic presented in Sections 5.1 and 5.4 in a sequential manner, with the output of one determining the input of the other, gives the overall simulation logic for simulating both primary and secondary vehicle types. It is used to quantify the effectiveness of a given allocation $(Z_a, \tilde{Z}_a)$ for all $a \in \mathscr{A}$, by recording several useful KPIs. In this work, the KPIs of interest are: the average primary vehicle utilisation, the average secondary vehicle utilisation, the mean response time, the percentage of abandoned calls (that is a measure of the primary vehicle unavailability), and the expected overall survival. Survival is modelled using a combination of survival function curves and hard cut-offs, and is described in Section 4.1. The source code for the combined simulation, is available at [35] and development took place at `https://github.com/MarkTuson/ambulance_simulation`.

6. The Case of Jakarta

Jakarta, the capital of Indonesia, has a population of 10.5 million [7] residing within 664 km^2. The city is divided into five municipalities and one district, each of which is divided into sub districts and, in turn, neighbourhoods. There are in total 42 sub districts and 261 neighbourhoods within mainland Jakarta (excluding the Thousand Islands regency in the north) [6]; a map is given in Figure 6.

As described in Section 1, there are many challenges to calling for an ambulance in Jakarta, as captured in [13]. Based on this work the regional government of Jakarta invested in a new fleet of coordinated ambulances accessible via the single emergency number 119. This work, in collaboration with Ambulans 118 and the 119 Emergency Ambulance Service, finds potential fleet allocations using the optimisations methods given in Section 4, and evaluates these allocations' effectiveness using the simulation described in Section 5.

As of October 2022, the 119 service in Jakarta ran a fleet of 81 primary Emergency Ambulances (EAs) and 13 secondary motorbike Rapid Response Vehicles (RRVs), distributed across 67 ambulance stations throughout the city. Figure 7 summarise the allocations vehicle locations. Working with our ambulance partners, after careful consideration the following patient 'specialities' ($k \in \mathscr{K}$) were agreed based on clinical need (although other categories could readily be adopted and included as necessary):

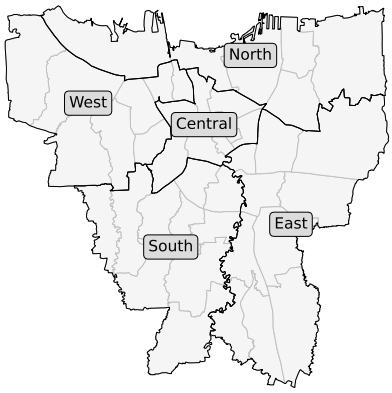


Figure 6: Map of Jakarta's Municipalities.

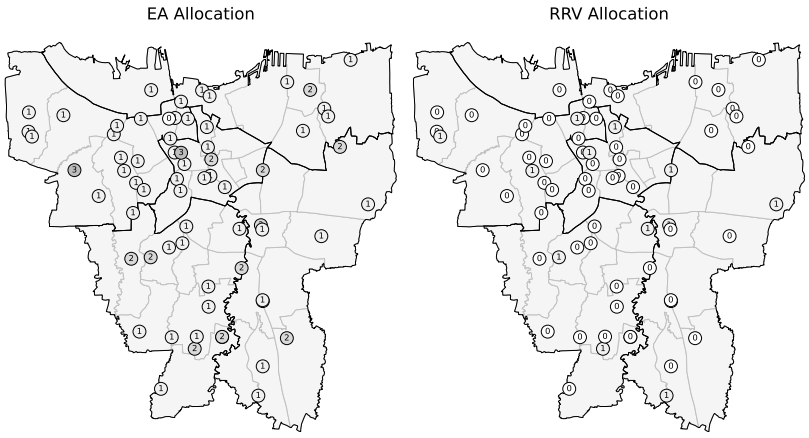


Figure 7: Map of current Emergency Ambulance (EA) and Rapid Response Vehicle (RRV) locations and allocations to ambulance bases across Jakarta.

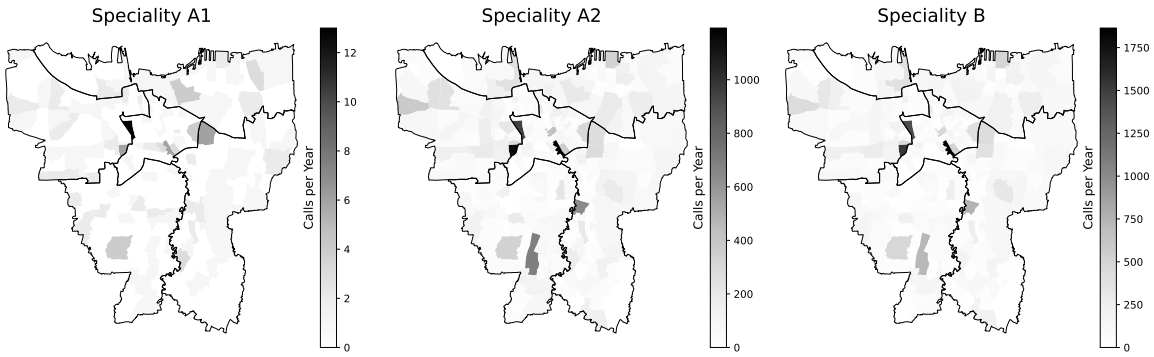


Figure 8: Number of calls received by speciality and neighbourhood.

- **A1 - High priority emergency patients**: critical patients who require immediate life-saving assistance with a target ambulance response time of 8 minutes. In our data set there were 168 identified calls that met this criterion.
- **A2 - Other emergency patients**: urgent patients that require assistance with a target response time of 15 minutes. In our data set there were 23,784 calls of this type.
- **B - Non-emergency patients**: patients that still benefit from an ambulance response but non-critical with a target response time of 60 minutes. There were 31,725 calls of this type in the data set.

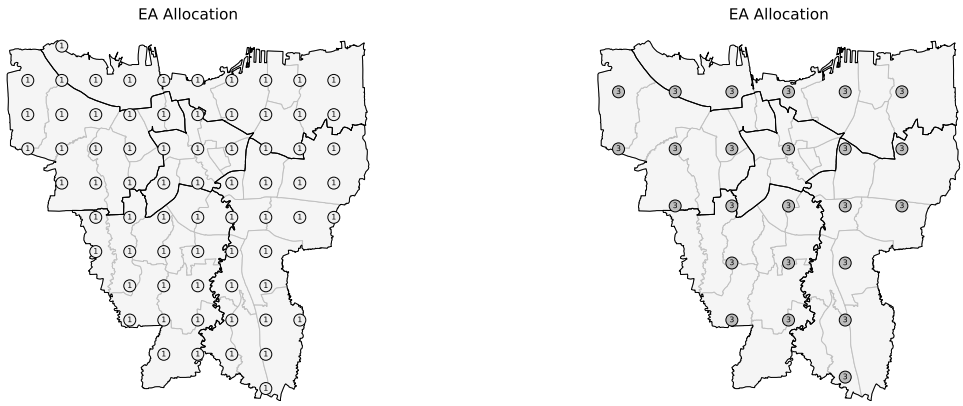
Secondary vehicles can be called to patients of speciality A1 and A2, therefore $\mathscr{K}_A = \{\text{A1}, \text{A2}\}$ and $\mathscr{K}_B = \{\text{B}\}$. The number of calls from each of these specialities varies by neighbourhood, as shown in Figure 8. It can be seen that many calls are highly concentrated in a handful of neighbourhoods rather than spread throughout the city. In this work, we approximate pick-up locations by the geographic centroid of each of the 261 neighbourhoods, each representing an ambulance pick-up that that occurred within that neighbourhood.

Firstly, in Section 6.1 we use the MESLMHPHF objective function of expected survival, and the simulation, to find KPIs that to compare the performance of two proposed grid allocations to the currently used allocation. An additional KPI is also calculated, the expected survival of A1 patients, found by taking Equation 5 and summing over $k = \text{A1}$ only. Then in Section 6.2 we outline four possible future demand scenarios, and in Section 6.3 we use the optimisation model to find and compare improved allocations for the current number of vehicles. Finally in Section 6.4 we investigate the number of vehicles required to meet the demand across the given scenarios.

All models are parameterised using demand data that covers all 261 neighbourhoods from 1 January to 31 December 2019, before the COVID-19 pandemic. Data from 2020-2021 was naturally heavily skewed by the pandemic, with significantly lower demand, and so was not considered as representative of a typical period for forecasting future needs, hence the decision to use the available 2019 demand. Appendix B provides further details on the parameterisation of the model, including how travel time estimates were obtained.

6.1. Evaluating 'Grid' Proposals

Discussions with senior staff at 119 identified that they were seriously considering a re-configuration their ambulance allocations into a grid structure, placing an ambulance station at regular intervals throughout the city and uniformly distributing the vehicles that they felt would ensure equitable coverage. Two possible grid allocations were being considered: one placing an EA every 3km across the city (giving a total of 70 vehicles),



(a) 3km grid.

(b) 5km grid.

Figure 9: Proposed grid allocations.

Allocation	Baseline	Grid 3km	Grid 5km
Number of EAs	81	70	71
Number of RRVs	13	0	0
Ambulance Utilisation	30.99%	36.58%	35.79%
RRV Utilisation	22.47%	-	-
Mean Response Time (mins)	17.69	22.29	24.05
MESLMHPHF Objective	98.34%	87.95%	85.64%
Expected A1 Survival	26.05%	13.64%	12.23%
Percent Abandoned	0.00%	0.52%	0.93%

Table 1: Calculated KPIs for the current and proposed grid allocations.

and another placing three EAs every 5km across the city (giving a total of 72 vehicles). Figure 9 show these proposed allocations.

Running the simulation and MESLMHPHF objective function of expected survival for both the current and the two proposed grid allocations gives the results shown in Table 1. The simulation was run over half a year, with a month warm up time, over eight replications. We observe that the two proposed grid allocations perform much worse than the current allocation as measured by mean response time, survival, and percentage of calls abandoned. This is expected as vehicle numbers are lower under the new proposals, as well as the allocations focussing on coverage rather than response times or survival.

6.2. Planning for Future Demand

Similarly to many other cities worldwide, ambulance services in Jakarta anticipate that demand for services will grow in future years, not least because there is now a coordinated service accessible via a single common number '119' to call. This should help raise awareness and visibility of an ambulance service. In the case of Jakarta, certainly increased use of 119 and therefore increased demand would be a desired outcome

Reason	% of emergency respondents
Used Ambulance	13.0
Too expensive	5.2
Not available	13.1
Would take too long	16.5
Not necessary	15.7
Not aware of service	33.1
Other	3.5

Table 2: Barriers to use of an ambulance service in Jakarta, from [13].

and is pro-actively supported by the Indonesian Government.

We consider four different demand scenarios, developed by considering the result of our cross-sectional study of patients attending EDs in Jakarta [13]. As part of the survey, randomly selected patients arriving at each emergency department were asked whether they had used an ambulance to attend, and reasons for not using an ambulance. For those patients who were categorised by the medical staff as emergency and for whom therefore an ambulance might have seemed a sensible option, the responses are given in Table 2.

The survey confirmed three major barriers to use, namely: the service's visibility (33.1% of respondents were unaware of the service), the service's reliability (13.1% of respondents reported that there was no ambulance available and 16.5% of respondents believed it would take too long), and the service's cost (5.2% of respondents cited cost as a deterrent). Therefore, four demand scenarios were considered, corresponding to addressing each of these barriers in turn:

- **D13**: representing the current situation where approximately 13% of emergency patients (specialities A1 and A2) do use an ambulance.
- **D19**: representing the situation where visibility is addressed. Here, we distribute the 33.1% of the respondents who were unaware of the service proportionally between using the ambulance and amongst the remaining issues. Using this methodology we would expect 19.4% of emergency patients to now use an ambulance.
- **D34**: representing the situation where reliability and visibility are addressed. Using the same methodology we would expect 34.8% of emergency patients to now use an ambulance.
- **D45**: representing the situation where cost, reliability and visibility are all addressed. Using the same methodology we would expect 45.4% of emergency patients to now use an ambulance.

Guided by our findings in [13], recent changes mean that the 119 services is now free to use for Jakarta residents, so certainly the highest demand (**D45**) scenario is plausible in the near future. After discussions with our partners, it was agreed that for each of the scenarios we should assume that non-emergency demand (speciality B) remains unchanged.

6.3. Improving the Current Allocation

Applying the optimisation metaheuristic from Section 4 to the current allocation re-allocates the 81 EAs and 13 RRVs. We do this for each of the four scenarios described in Section 6.2. The hyper-parameters used for this optimisation algorithm throughout this study are given in Table 3, which include approximations

Parameter	Value Used	Explanation
μ	$\frac{1}{3.886\times60}$	Average primary vehicle service rate (hr^{-1}).
$\tilde{\mu}$	$\frac{1}{1.038\times60}$	Average secondary vehicle service rate (hr^{-1}).
N	100	Population size.
κ	20	Number of solutions to keep per generation.
m_0	6	Initial number of mutations.
c	0.1	Cooling rate, rate at which number of mutations decreases.
H	200	Number of generations.

Table 3: Hyper-parameters used in the optimisation for all experiments.

Demand	Allocation	Expected Survival	Expected A1 Survival	Primary Utilisation	Secondary Utilisation	Mean Response Time (mins)	Abandoned
D13	Current	98.34%	26.05%	30.99%	22.47%	17.69	0.00%
	Optimised	99.10%	30.99%	31.06%	18.29%	17.86	0.00%
D19	Current	97.74%	25.17%	38.11%	32.24%	18.48	0.44%
	Optimised	99.64%	29.85%	38.12%	27.39%	18.52	0.46%
D34	Current	95.78%	22.75%	54.07%	52.28%	22.21	3.44%
	Optimised	99.61%	29.59%	54.16%	49.07%	22.13	3.79%
D45	Current	92.88%	20.67%	61.79%	60.89%	23.99	9.69%
	Optimised	99.54%	29.36%	61.62%	56.80%	23.18	9.27%

Table 4: Calculated KPIs for the current and improved allocations under the four possible demand scenarios.

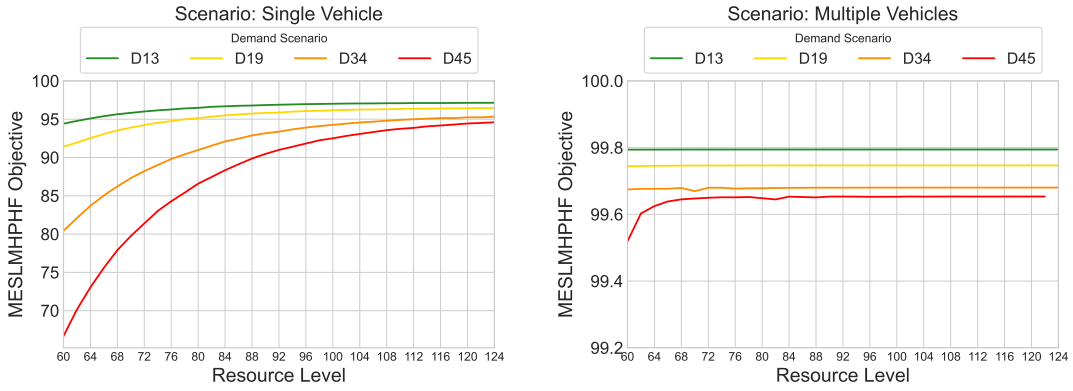
for the average service times of primary and secondary vehicles derived from the initial simulation model. Appendix C shows some initial explorations on the hyper-parameter choices, giving confidence that the chosen parameters, with a large enough number of iterations N, are sufficient to find allocations that perform well.

Table 4 compares the MESLMHPHF objective function value and simulation derived KPIs between the current and the optimised allocations, for each demand scenario. It can be seen that, as expected, as demand increases then utilisation of primary and secondary vehicles increase, as the ambulance service is busier. Similarly, mean response time increases with demand, this is due to the vehicles being busier, and therefore less chance that the most optimally placed vehicle is dispatched. This is confirmed by the percentage of calls abandoned increasing with demand, showing ambulance unavailability. Looking at the expected survival, that is the value of the MESLMHPHF objective function, we see that using an better allocation does increase the survival, and that the drop in survival due to demand increases is not as severe when using an improved allocation in comparison to the current allocation. We can also see that in particular A1 patients are benefiting from the improved allocations, as these are the patients most sensitive to small fluctuations in response time, due to their survival function Equation 3. It is worth noting that for the simulation based KPIs (mean response times and percent abandoned), there is not too much difference between the currently used allocation and the better allocation, showing that these KPIs may not be good approximations of survival.

6.4. *Finding Better Allocations for Any Resource Level*

The metaheuristic algorithm described in Section 4.5 finds allocation of ambulances across the 67 current ambulance stations for a given number of primary and secondary vehicles. Information supplied by one of the ambulance operators suggested that in terms of total running costs one EA (primary vehicle) was approximately equivalent to three RRVs (secondary vehicles), and so we consider three RRVs to be one resource; that is, a resource level of 60 could represent 60 primary and 0 secondary vehicles, or 59 primary and 3 secondary vehicles, or 58 primary and 6 secondary vehicles, and so on. For a given resource level we run the optimisation algorithm allowing for one primary to be swapped for three secondary vehicles, and vice versa, and we report KPIs for the best performing, in terms of maximal survival, combinations, as well as for the case when the number of secondary vehicles is fixed at zero. We call these the multiple vehicle type and single vehicle type scenarios.

Figures 10a-15b display the obtained KPIs for each of these allocations. Confidence intervals over the eight replications were too small to display on the plots. It can immediately be seen that increasing the resource level has a positive effect on all KPIs, with vehicle utilisations, mean response times, and percentage abandoned decreasing, and overall survival increasing. Similarly, as expected increasing the demand of emergency calls has a negative impact on all KPIs.



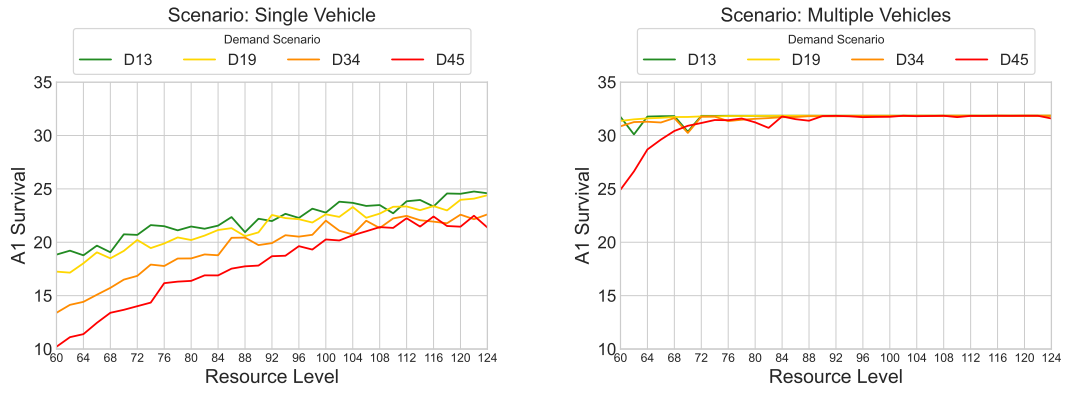
(a) Single vehicle type scenario.

(b) Multiple vehicle type scenario.

Figure 10: MESLMHPHF objective expected survival results.

It is particularly interesting to compare the scenario in which single vehicle types (only emergency ambulances) were allocated against the scenario where multiple vehicle types (both emergency ambulances and rapid response vehicles) were allocated. For equivalent resource levels, introducing secondary vehicles increases ambulance utilisation, and so decreases availability, resulting in an increase in abandoned calls. However, introducing secondary vehicles also results in a decrease in the mean response time, and a large increase in the expected survival, especially those of speciality A1.

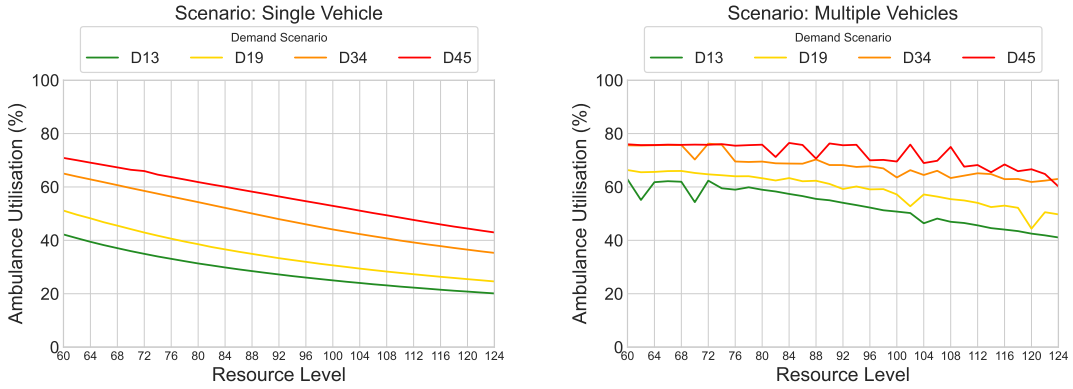
An interesting phenomenon occurs when looking at Figure 14b. Here it seems that, when using multiple vehicles, mean response time decreases as demand increases, which is counter-intuitive. This is however due to a peculiar interplay between demand and vehicle behaviours: as stated in Section 6.2 only demand for emergency patients are increased (specialities A1 and A2), which are also the patients that can be seen by secondary vehicles, and so would have a lower response time than non-emergency patients (speciality B). Thus, demand scenario **D45** receives a larger proportion emergency patients than scenario **D13**, and so a larger proportion of patients with lower response times due to being seen by secondary vehicles; bringing the overall



(a) Single vehicle type scenario.

(b) Multiple vehicle type scenario.

Figure 11: Expected survival of A1 patients results.



(a) Single vehicle type scenario.

(b) Multiple vehicle type scenario.

Figure 12: Ambulance utilisation results.

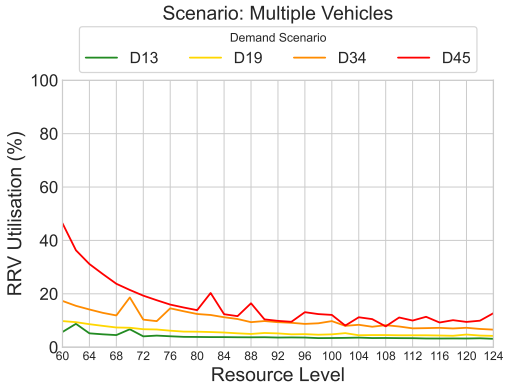
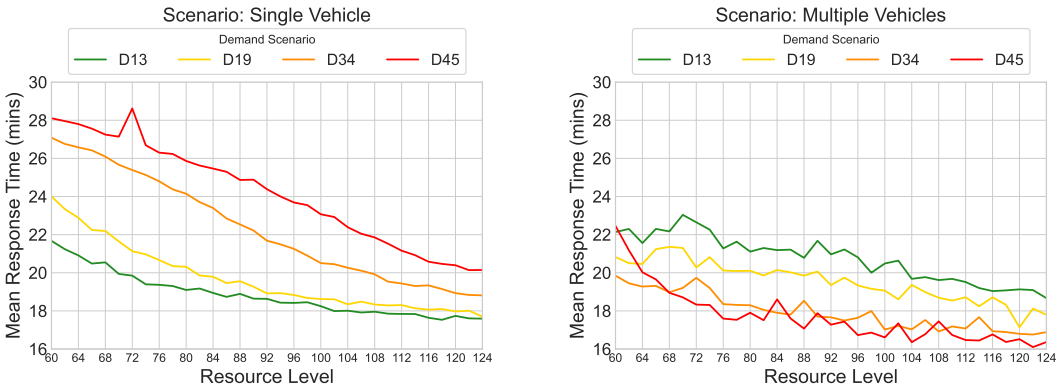


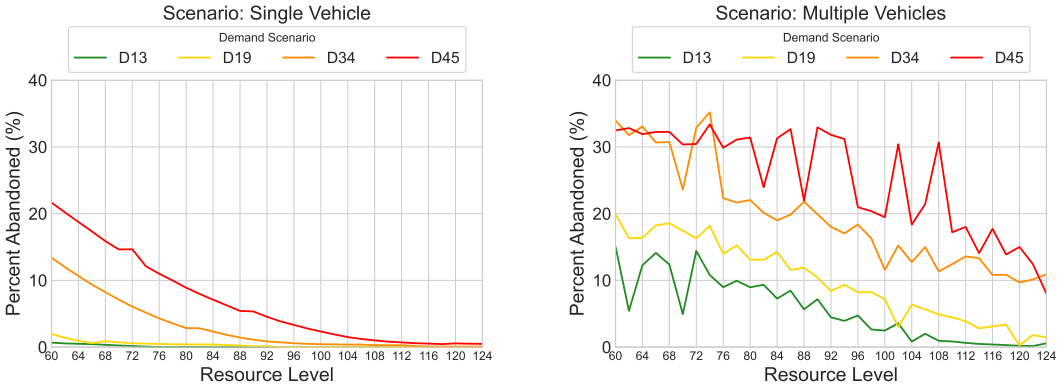
Figure 13: RRV utilisation results.



(a) Single vehicle type scenario.

(b) Multiple vehicle type scenario.

Figure 14: Mean response time results.



(a) Single vehicle type scenario.

(b) Multiple vehicle type scenario.

Figure 15: Percent of abandoned calls results.

mean response time down. This does not mean however that emergency patients in scenario **D45** have lower response times than those in scenario **D13**.

It is also noticeable that the increases or decreases in the simulation derived KPIs are not necessarily monotonic with increases in resource level. This is due to the derived allocations being based on the MESLMHPHF score only, which is a measure of survival, while the simulation derived KPIs are indirect measures of performance. Again, this may indicate that these indirect measures of performance, such as mean response times or percentage within target, are not good indicators of survival.

Crucially, the plots show that the allocations produced by the optimisation algorithm perform better than the current allocation for the same demand and resource levels. The current levels include 81 EAs and 13 RRVs, so a resource level of 85.33, with multiple vehicles. Comparing the derived allocations for this level in Figures 10a-15b to the results given in Table 1, the derived allocations give lower ambulance and RRV utilisations, lower mean response times, and a higher percentage of patients seen within target.

7. Discussion & Conclusions

This paper has described the development of EMS demand and capacity models and their application to the city of Jakarta, Indonesia. To our knowledge, this is the first such study of analysing emergency ambulance services in Jakarta, and the work has directly assisted and guided providers and the government to make investment decisions for a coordinated and free to use EMS.

Firstly, we have considered patient survival and outcomes within a developed MESLMHPHF model. By considering geospatial travel times and accounting for the varying needs of different patient types through categorising patients by speciality and through using different survival function profiles, an expected overall survival was given. This was used within a optimisation processes, with an evolutionary metaheuristic, to find vehicle fleet allocations that maximised expected survival. This expected survival function extends previous work in [24] to include more than one vehicle type. A particular novelty here was numerically solving implicit utilisation relationships, equations 13 and 14, to calculate the expected survival, as an alternative to busy fractions and hypercube models.

Secondly, a discrete event simulation model has been used to evaluate existing and potential heterogeneous vehicle ambulance feet allocations in terms of key performance measures, such as response times, survival, and vehicle utilisation rates. This takes into consideration geospatial demand and travel, and temporal variation in demand and traffic levels. A novel feature of our approach is that the model comprises of sequential simulations, feeding data directly from one into the other in order to simulate primary and secondary vehicles separately while maintaining synchronicity, with the overall aim of reducing the complexity of the simulation logic and maintaining modularity.

Both models investigate ambulance service activity in the case of heterogeneous patients and heterogeneous fleets. Heterogeneous patient groups consider those with distinct demand profiles, priorities, and survival functions. Heterogeneous fleets concern different types of vehicles, in our case emergency ambulances (EAs) which respond to every patient, and Rapid Response Vehicles (RRVs) which can be utilised to reach patients faster, despite being unable to transport patients themselves.

Using a combination of models has permitted an approach that can capture performance measures or take into account factors that each model in isolation can not. For example, the maximum expected survival model is a lot more appropriate for use within an optimisation process as the runtimes are a far more reasonable than the simulation. On the other hand, the simulation allows for greater complexity such as temporal demand, and is able to capture a greater range of KPIs. Crucially, using mixed methodology such as this allowed for independent evaluation of the output of one using the other; but also insights from an initial run of the simulation gave helped parametrise and justify assumptions of the optimisation.

A key feature of both models, and their novelty, is the consideration of heterogeneous fleets. In a highly populated area such as Jakarta (around 16,000 population per km^2), the inclusion of RRVs such as paramedics on motorbikes is very important, given RRVs can access areas that cannot easily or quickly be accessed by ambulances. Results from both the optimisation and simulation research showed that the RRVs can help increase the overall survival and reduce the response times, especially crucial when responding to life threatening events such as cardiac arrests [20].

The results of our research also suggest that allocation strategies may not be intuitive to ambulance service managers, further emphasising the value of a modelling approach. For example, senior managers suggested the use of a grid allocation to maximise geographic coverage, with ambulances equally spread across the city. This was shown to be sub-optimal because neighbourhoods have different population densities and characteristics. For example, the number of daily commuters in Jakarta is considerably high during day time due to work and education [5]. Municipalities where trading, educational institutions, and offices are concentrated may become more populated during day time. The model has quantified the deterioration in

response times and patient outcomes should a fixed grid system be implemented, leading to 119 to drop this consideration.

The ambulance posts in Jakarta depend on the service provider. Those ambulances provided by the government are in various locations including government buildings as well as community clinics and sub-district hospitals. In our case, we used current ambulance posts as the locations for allocating the resources. As demands for ambulances may change from time to time, future work could evaluate different, and perhaps dynamic, ambulance posts that depend on changing demand volumes in neighbourhoods by time of the day.

The data used for ambulance demand covered one year from 1 January to 31 December 2019, prior the COVID-19 pandemic. Recent studies related to ambulance demand indicate that the pandemic has severely impacted on the utilisation of EMS, for examples in call volumes [39] and in specific medical conditions such as trauma [4], and possibly still continue to affect demand. Ambulance providers, such as 118 and 119, may use our modelling tools to incorporate future demand data and re-evaluate resource needs. To aid this, we are currently working on interfacing the simulation and optimisation into a single easy to use decision support tool with a dashboard (data visualiser).

Studies have shown that the quality in pre-hospital data collection varies considerably in Indonesia [21]. We encountered similar challenges and have suggested to 119 and the Indonesian Government that they should continue to strive to improve the coverage and quality of data collection. Future work could explore in more depth unmet demand that is not recorded in the ED or in the ambulance data.

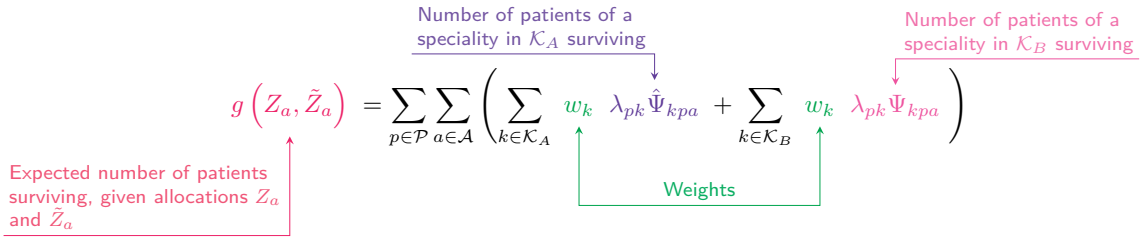
Indonesia is a vast country with an uneven population density and varying quality of available healthcare resources. The models have been parameterised with data from a high population, high density urban area which is also relatively well resourced with healthcare facilities compared to other regions in the country. Ambulance services in rural areas of Indonesia may not have the same quality in management and organisation compared to the capital, Jakarta. We intend to conduct future studies to apply the developed models to analyse ambulance demand and allocations in different areas of Indonesia, including in rural regions.

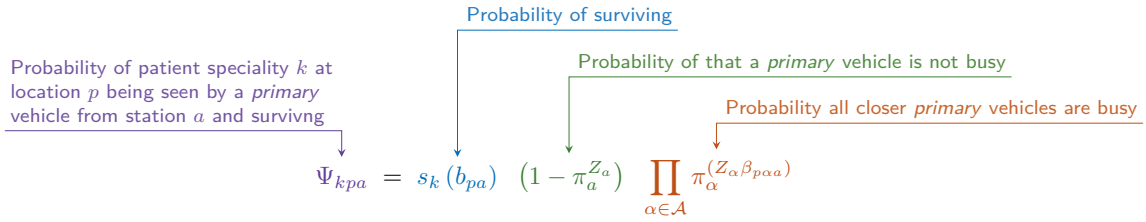
In conclusion, the developed models have demonstrated a novel approach in modelling ambulance allocations that incorporate multiple vehicle types and health conditions, whilst also capturing patient survival. Our work has already informed major decisions on the design of a free to use and coordinated EMS system for Jakarta. Ongoing collaboration will continue to assist ambulance providers and the Indonesian Government in providing evidence-based decision making for the benefit of patients and the population they serve, including the exploration of the roll-out of 119 beyond Jakarta to other regions of Indonesia.

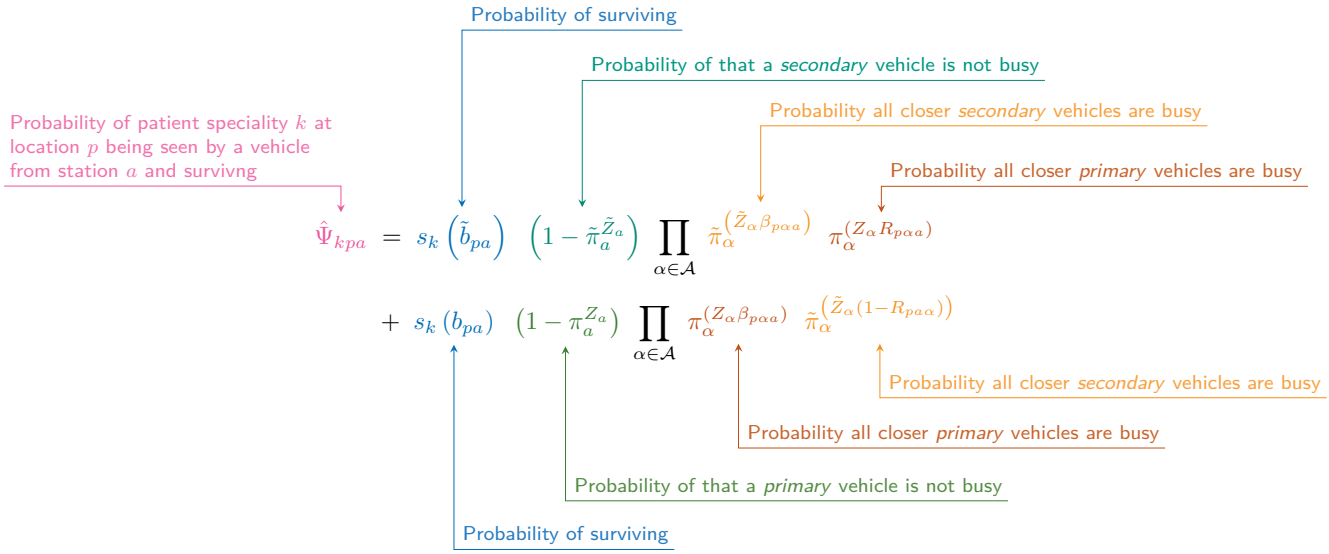
Acknowledgements

The study was funded by EPSRC with grant no: EP/T003197/1. We would like to express our sincere gratitude to Indonesian emergency ambulance providers including 118 and 119 for their support, insights and provision of data. In particular we wish to thank Professor Aryono Djuned Pusponegoro and Ms Asti Puspita Rini, Founder and Director of 118 Ambulance Service Foundation and Dr Winarto, Head of the 119 Ambulance Service in Jakarta.

Appendix A. Annotated Objective Function







Appendix B. Model Parameters

Call arrival rates λ_{pk} are derived from 2019 demand data split by municipality and speciality, and time of day. Probabilities q_{pky} are similarly derived from the 2019 data of transit journeys. All traffic-free travel times are found using Google Maps API, while time-dependent traffic delays are found from the TomTom website and given in Table B.5.

h	0000-0500	0500-1500	1500-1800	1800-0000
d_h	0.98	0.66	0.59	0.77
$\tilde{d}_h$	1.96	0.91	0.83	1.16

Table B.5: Primary and secondary delay factors.

The other models are parameterised in the following way:

- From a sample of calls the time at site G_k was found to follow a lognormal distribution with parameters $\mu = -0.6219$ and $\sigma = 0.8048$. For the case of Jakarta this was modelled identically for all specialities k. Comparison between the lognormal fit and the sampled times are given in Figure B.16.
- From discussions with staff at the ambulance service in Jakarta, the time at hospital J_k was modelled as a Uniform distribution between 40 and 60 minutes for the emergency specialities A1 and A2, and between 20 and 30 minutes for non-emergency speciality B.
- From discussions with staff at the ambulance service in Jakarta, the refill time Θ is taken to be 60 minutes for an emergency ambulance, and 15 minutes for an RRV.

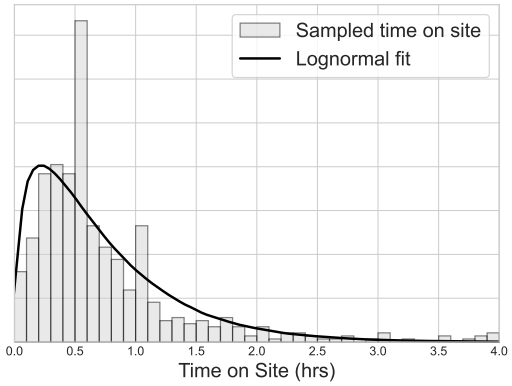


Figure B.16: Comparison between the sampled time on site and the lognormal fit.

Appendix C. Exploration of Optimisation Hyperparameters

Figure C.17 shows the performance of the evolutionary metaheuristic algorithm under different values of the hyperparameters N, κ, m_0, and c.

Appendix D. Improved Allocations for Current Vehicle Numbers

Figures D.18, D.19, D.20 and D.21 show the improved allocations for 81 primary and 13 secondary vehicles under demand scenarios **D13**, **D19**, **D34**, and **D45** respectively.

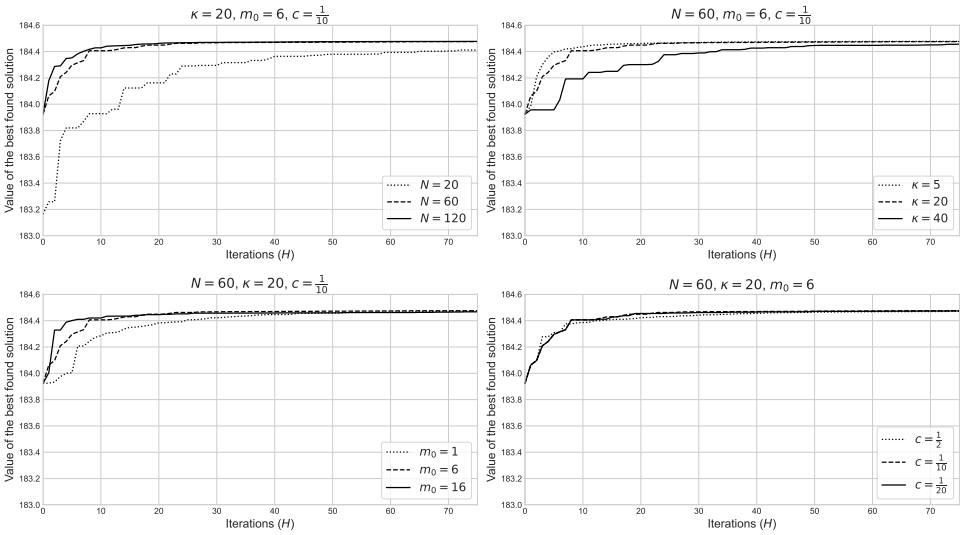


Figure C.17: Comparison between optimisation performance for low, medium, and high values of N, κ, m_0, and c. Optimisation is run on demand scenario **D19**, for a resource level of 75.

References

[1] Aboueljinane, L., Sahin, E., and Jemai, Z. (2013). A review on simulation models applied to emergency medical service operations. *Computers & Industrial Engineering*, 66(4):734–750.

[2] Amorim, M., Antunes, F., Ferreira, s., and Couto, A. (2019). An integrated approach for strategic and tactical decisions for the emergency medical service: Exploring optimization and metamodel-based simulation for vehicle location. *Computers & Industrial Engineering*, 137:106057.

[3] Aringhieri, R., Bruni, M., Khodaparasti, S., and van Essen, J. (2017). Emergency medical services and beyond: Addressing new challenges through a wide literature review. *Computers & Operations Research*, 78:349–368.

[4] Azbel, M., Heinänen, M., Lääperi, M., and Kuisma, M. (2021). Effects of the covid-19 pandemic on trauma-related emergency medical service calls: a retrospective cohort study. *BMC emergency medicine*, 21(1):1–10.

[5] Badan Pusat Statistik Provinsi DKI Jakarta (2001). *Migrasi Penduduk JABOTABEK 2001*. Badan Pusat Statistik provinsi DKI Jakarta.

[6] Badan Pusat Statistik Provinsi DKI Jakarta (2020). *DKI Jakarta Province in Figures*. Badan Pusat Statistik provinsi DKI Jakarta.

[7] Badan Pusat Statistik Provinsi DKI Jakarta (2023). *Analisis Profil Penduduk Provinsi DKI Jakarta: mendeskripsikan Peran Penduduk dalam Pembangunan*. Badan Pusat Statistik provinsi DKI Jakarta.

[8] Bélanger, V., Lanzarone, E., Nicoletta, V., Ruiz, A., and Soriano, P. (2020). A recursive simulation-optimization framework for the ambulance location and dispatching problem. *European Journal of Operational Research*, 286(2):713–725.

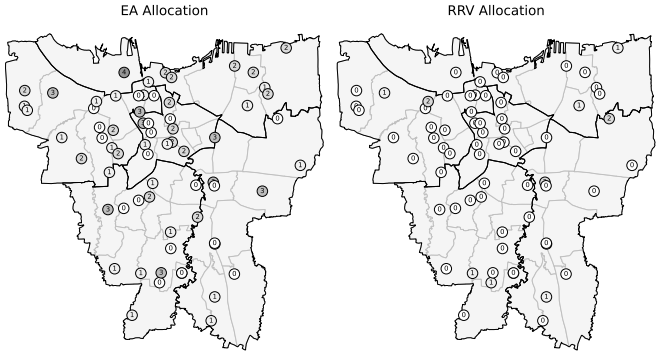


Figure D.18: Improved allocation of 81 EAs and 13 RRVs, under scenario **D13**.

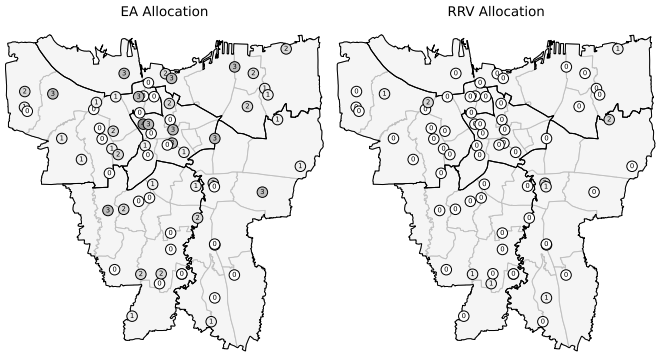


Figure D.19: Improved allocation of 81 EAs and 13 RRVs, under scenario **D19**.

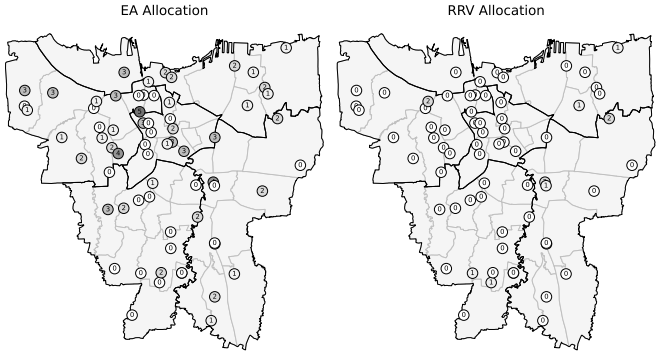


Figure D.20: Improved allocation of 81 EAs and 13 RRVs, under scenario **D34**.

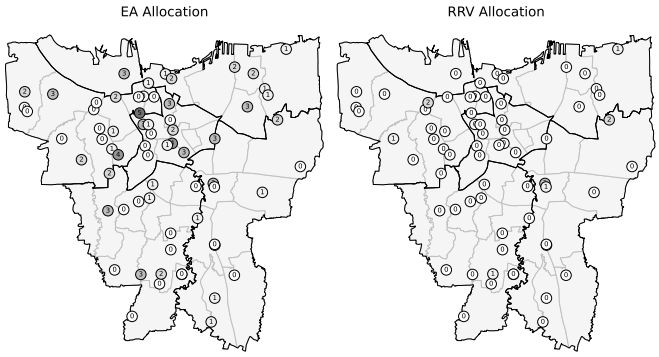


Figure D.21: Improved allocation of 81 EAs and 13 RRVs, under scenario **D45**.

[9] Bélanger, V., Ruiz, A., and Soriano, P. (2019). Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. *European Journal of Operational Research*, 272(1):1–23.

[10] Birmingham, L., Arens, A., Longinaker, N., and Kummet, C. (2021). Trends in ambulance transports and costs among medicare beneficiaries, 2007–2018. *The American Journal of Emergency Medicine*, 47:205–212.

[11] Boutilier, J. and Chan, T. (2022). Drone network design for cardiac arrest response. *Manufacturing & Service Operations Management*, 24(5):2407–2424.

[12] Brailsford, S., Eldabi, T., Kunc, M., Mustafee, N., and Osorio, A. (2019). Hybrid simulation modelling in operational research: A state-of-the-art review. *European Journal of Operational Research*, 278(3):721–737.

[13] Brice, S., Boutilier, J., Gartner, D., Harper, P., Knight, V., Lloyd, J., Pusponegoro, A., Rini, A., Turnbull-Ross, J., and Tuson, M. (2022). Emergency services utilization in Jakarta (Indonesia): a cross-sectional study of patients attending hospital emergency departments. *BMC health services research*, 22(1):639–639.

[14] Chang, C., Abujaber, S., Reynolds, T., Camargo, C. J., and Z, O. (2016). Burden of emergency conditions and emergency care utilization: New estimates from 40 countries. *Emergency Medical Journal*, 33:794–800.

[15] Daskin, M. (1983). A maximum expected covering location model: formulation, properties and heuristic solution. *Transportation science*, 17(1):48–70.

[16] Erkut, E., Ingolfsson, A., and Erdogan, G. (2008). Ambulance location for maximum survival. *Naval Research Logistics*, 55(1):42–58.

[17] Farahani, R., Fallah, S., Ruiz, R., Hosseini, S., and Asgari, N. (2019). Or models in urban service facility location: A critical review of applications and future developments. *European journal of operational research*, 276(1):1–27.

[18] Fraser, A., Newberry Le Vay, J., Byass, P., et al. (2020). Time-critical conditions: assessment of burden and access to care using verbal autopsy in agincourt, south africa. *BMJ Global Health*, 5:e002289.

[19] Henderson, S. and Mason, A. (2004). Ambulance service planning: simulation and data visualisation. *Operations research and health care: a handbook of methods and applications*, pages 77–102.

[20] Holmén, J., Herlitz, J., Ricksten, S., Strömsöe, A., Hagberg, E., Axelsson, C., and Rawshani, A. (2020). Shortening ambulance response time increases survival in out-of-hospital cardiac arrest. *Journal of the American Heart Association*, 9(21):e017048.

[21] Hooper, C., Ranse, J., and Hutton, A. (2019). How is ambulance patient care and response time data collected and reported in malaysia and indonesia? *Australasian Journal of Paramedicine*, 16:1–8.

[22] Horn, M. (2000). Efficient modeling of travel in networks with time-varying link speeds. *Networks: An International Journal*, 36(2):80–90.

[23] Kergosien, Y., Bélanger, V., Soriano, P., Gendreau, M., and Ruiz, A. (2015). A generic and flexible simulation-based analysis tool for ems management. *International Journal of Production Research*, 53(24):7299–7316.

[24] Knight, V., Harper, P., and Smith, L. (2012). Ambulance allocation for maximal survival with heterogeneous outcome measures. *Omega*, 40(6):918–926.

[25] Knight, V. and Palmer, G. (2023). drvinceknight/heterogeneousambulancefleetallocations: v1.0.0.

[26] Lewis, P. and Shedler, G. (1979). Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413.

[27] Li, X., Zhao, Z., Zhu, X., and Wyatt, T. (2011). Covering models and optimization techniques for emergency response facility location and planning: a review. *Mathematical Methods of Operations Research*, 74:281–310.

[28] Liu, Y., Yuan, Y., Shen, J., and Gao, W. (2021). Emergency response facility location in transportation networks: A literature review. *Journal of traffic and transportation engineering (English edition)*, 8(2):153–169.

[29] Lowthian, J., Cameron, P., Stoelwinder, J., Curtis, A., Currell, A., Cooke, M., and McNeil, J. (2011). Increasing utilisation of emergency ambulances. *Australian Health Review*, 35(1):63–69.

[30] McCormack, R. and Coates, G. (2015). A simulation model to enable the optimization of ambulance fleet allocation and base station location for increased patient survival. *European Journal of Operational Research*, 247(1):294–309.

[31] McLay, L. A. and Mayorga, M. E. (2010). Evaluating emergency medical service performance measures. *Health care management science*, 13:124–136.

[32] Morgan, J., Howick, S., and Belton, V. (2017). A toolkit of designs for mixing discrete event simulation and system dynamics. *European Journal of Operational Research*, 257(3):907–918.

[33] Mukhopadhyay, A., Pettet, G., Vazirizade, S., Lu, D., Jaimes, A., El Said, S., Baroud, H., Vorobeychik, Y., Kochenderfer, M., and Dubey, A. (2022). A review of incident prediction, resource allocation, and dispatch models for emergency management. *Accident Analysis & Prevention*, 165:106501.

[34] Palmer, G., Knight, V., Harper, P., and Hawa, A. (2019). Ciw: An open-source discrete event simulation library. *Journal of Simulation*, 13(1):68–82.

[35] Palmer, G. and Tuson, M. (2023). Marktuson/ambulance_simulation: v1.0.0.

[36] Pidd, M. (2004). *Computer simulation in management science*. Number 5th. John Wiley and Sons Ltd.

[37] Pusponegoro, A. (2003). Terrorism in indonesia. *Prehospital and disaster medicine*, 18(2):100–105.

[38] Reuter-Oppermann, Mand van den Berg, P. and Vile, J. (2017). Logistics for emergency medical service systems. *Health Systems*, 6(3):187–208.

[39] Şan, İ., Usul, E., Bekgöz, B., and Korkut, S. (2021). Effects of covid-19 pandemic on emergency medical services. *International Journal of Clinical Practice*, 75(5):e13885.

[40] Schmid, V. and Doerner, K. (2010). Ambulance location and relocation problems with time-dependent travel times. *European journal of operational research*, 207(3):1293–1303.

[41] Suryanto, M., Plummer, V., Boyle, M., et al. (2017). EMS systems in lower-middle income countries: a literature review. *Prehospital and disaster medicine*, 32(1):64–70.

[42] The Ciw library developers. (2022). Ciwpython/ciw: v2.3.3.

[43] Toro-Díaz, H., Mayorga, M., Chanta, S., and McLay, L. (2013). Joint location and dispatching decisions for emergency medical services. *Computers & industrial engineering*, 64(4):917–928.

[44] Toro-Díaz, H., Mayorga, M., McLay, L., Rajagopalan, H., and Saydam, C. (2015). Reducing disparities in large-scale emergency medical service systems. *Journal of the Operational Research Society*, 66(7):1169–1181.

[45] Valenzuela, T., Roe, D., Nichol, G., Clark, L., Spaite, D., and Hardman, R. (2000). Outcomes of rapid defibrillation by security officers after cardiac arrest in casinos. *New England Journal of Medicine*, 343(17):1206–1209.

[46] Veser, A., Sieber, F., Groß, S., and Prückner, S. (2020). The demographic impact on the demand for emergency medical services in the urban and rural regions of bavaria, 2012–2032. *Journal of Public Health*, 23:181–188.

[47] Virtanen, P., Gommers, R., Oliphant, T., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S., Brett, M., Wilson, J., Millman, K., Mayorov, N., Nelson, A., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E., Harris, C., Archibald, A., Ribeiro, A., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

[48] Wang, W., Wu, S., Wang, S., Zhen, L., and Qu, X. (2021). Emergency facility location problems in logistics: Status and perspectives. *Transportation research part E: logistics and transportation review*, 154:102465.

[49] Yusvirazi, L., Ramlan, A., and Hou, P. (2018). State of emergency medicine in Indonesia. *Emergency Medicine Australasia*, 30(6):820–826.