OPERATION OF AN AMBULANCE FLEET UNDER UNCERTAINTY

Vincent Guigues
School of Applied Mathematics, FGV
Praia de Botafogo, Rio de Janeiro, Brazil
vincent.guigues@fgv.br

Anton Kleywegt
Georgia Institute of Technology
Atlanta, Georgia 30332-0205, USA,
anton@isye.gatech.edu

Victor Hugo Nascimento
UFRJ
Rio de Janeiro, Brazil
victorhugo.vhrn@gmail.com

Abstract. We introduce two new optimization models for the dispatch of ambulances. These models are to our knowledge the first providing a full modelling of the operation of an ambulance fleet, taking into account all or almost all constraints of the problem. The first model, called the ambulance selection problem, is used when an emergency call arrives to decide whether an ambulance should be dispatched for that call, and if so, which ambulance should be dispatched, or whether the request should be put in a queue of waiting requests. The second model, called the ambulance reassignment problem, is used when an ambulance finishes its current task to decide whether the ambulance should be dispatched to a request waiting in queue, and if so, which request, or whether the ambulance should be dispatched to an ambulance staging location, and if so, which ambulance staging location. These decisions affect not only the emergency call and ambulance under consideration, but also the ability of the ambulance fleet to service future calls. There is uncertainty regarding the locations, arrival times, and types of future calls. To obtain decisions that are good not only for the current emergency call and ambulance under consideration, but also for future calls, we propose a rolling horizon approach that combines the current decisions to be made as first-stage decisions with second-stage models that represent the ability of the ambulance fleet to service future calls. The second-stage optimization problems can be formulated as large-scale deterministic integer linear programs. We propose a column generation algorithm to solve the continuous relaxation of these second-stage problems. The optimal objective values of these second-stage continuous relaxations are used to make approximately optimal first-stage decisions. We compare our resulting dispatch policy with the popular closest-available-ambulance dispatch rule (for the ambulance selection problem) and a combination of priority dispatch and the closest ambulance staging location rule (for the ambulance reassignment problem), for the Rio de Janeiro emergency medical service, based on data of more than 2 years of emergency calls for that emergency medical service. These tests show that our proposed policy results in smaller response times than the popular decision rules.

Key Words: stochastic programming, column generation, vehicule routing, ambulance allocation, health.

AMS subject classifications: 90C15, 90C90, 90C30.

1. Introduction

This paper proposes new optimization models for aiding ambulance dispatch decisions. The paper also presents a solution method for these problems, and tests the proposed approach with real emergency medical service data.

1.1. **Ambulance Fleet Operations.** We consider requests for emergency medical service that arrive at a call center. A call center telecommunicator receives the call and obtains data from the caller, including the nature of the emergency and the location of the emergency. (Often multiple calls are received related to the same emergency. It is an important question how to determine whether or not a particular call is related to an emergency that has already been reported. We do not address that question in this paper.) The telecommunicator records the data and decides whether to request that an ambulance be dispatched to the emergency. If so, a decision is made which ambulance should be dispatched to the emergency, or whether the request should be placed in a queue of requests waiting for an ambulance to be dispatched. Sometimes multiple ambulances are dispatched to an emergency. Here we consider the more typical situation in which a single ambulance is dispatched to an emergency. Usually, if an ambulance is dispatched to an emergency, it is an available ambulance, that is, an ambulance that is not on the way to an emergency, or busy at an emergency, or transporting patients from an emergency to a hospital. That is, an available ambulance is an ambulance that is either waiting at an ambulance staging location for an assignment, or on its way to an ambulance staging location after attending to an emergency. Some emergency medical services also consider dispatching an ambulance that is on the way to another emergency, that is, an ambulance that has already been dispatched to an (apparently less urgent) emergency is preempted and dispatched to the new (apparently more urgent) emergency. In some systems the telecommunicator makes the decision, and in other systems a separate ambulance dispatcher makes the decision.

If an ambulance is dispatched to the emergency, then it takes the ambulance some time to arrive at the location of the emergency. The amount of time that elapses from the moment the first call related to an emergency is received until the first ambulance personnel arrive at the patient(s) is called *the response time*. (In practice, distinction is made between different response times, for example, the elapsed time can start when the first call related to an emergency is received or when the telecommunicator requests that an ambulance be dispatched or when the ambulance personnel receive the dispatch instructions, and the elapsed time can end when the first ambulance arrives at the location of the emergency or when the first ambulance personnel arrive at the patient(s). In this paper we consider one of these times as the response time.) It has been found that mortality odds increase as the response time increases [8]. More specifically, the effectiveness of a treatment often depends on the response time. For example, it has been found that if an ambulance arrives 10 minutes after the onset of myocardial infarction (heart attack), then defibrillation reduces prehospital mortality from 6% to 2%, but if an ambulance arrives 60 minutes after the onset of infarction, then defibrillation reduces prehospital mortality from 13% to 12% [15]. In addition, response times of emergency medical services (EMSs) are relatively easy to measure, and most EMSs record and report (for example, to the National Emergency Medical Services Information System NEMSIS) response time data. Therefore many EMSs as well as academic papers put great emphasis on the response time performance metric. As summary statistic of response time data, EMSs often measure specific quantiles of the response time empirical distribution. For example, the 0.8 and 0.9 empirical quantiles are measured (often pooling all emergency data, that is, without distinction between different types of emergencies), and for performance to be acceptable, the considered quantiles have to be less than specified threshold values [34, 61, 51]. It has been pointed out that response time is not the only factor under control of emergency medical services that affects the survival probabilities of patients, and also that the impact of emergency medical care and response time depends on the type of emergency. For example, it has been found that advanced pre-hospital life-support has greater impact on mortality and morbidity of patients who had respiratory distress than patients who had cardiac arrest [68]. A potential concern with using the type of emergency in dispatch decisions is that emergencies are often misclassified by telecommunicators due to limited or incorrect data. Based on an analysis of medical emergency data in the UK under two systems for classifying emergencies into priority classes, [58] found that the benefits of priority dispatch outweighed the risk of misclassification. Therefore, it makes sense to take the characteristics of the emergency into account when deciding whether to immediately dispatch an ambulance to the emergency, and if so, which ambulance to dispatch, as opposed to considering only specific quantiles of the response time distribution irrespective of the nature of the emergencies.

Also, the capabilities of the ambulance and personnel can affect the survival probability of the patient, depending on the type of emergency. Although many academic papers consider all ambulances to be the same, typically ambulances are not the same. For example, some EMSs make a distinction between basic life support (BLS) ambulances and advanced life support (ALS) ambulances [64], some EMSs make a distinction between first responders staffed by emergency medical technicians and second responders staffed by paramedics, and some EMSs have stroke units in addition to BLS and ALS ambulances. At a sufficient level of detail, every ambulance is unique, because the crew members of different ambulances have different qualifications and experience. [59] compared a targeted EMS system that dispatches BLS or ALS ambulances according to the nature of the emergency and an EMS system that provides uniform service with all ALS ambulances. The study focused on witnessed ventricular fibrillation cardiac arrest emergencies, and compared system performance in terms of four outcomes: return of spontaneous circulation, survival to hospital admission, survival to hospital discharge, and survival to 1 year. All performance measures were better for the targeted system that dispatched BLS or ALS ambulances according to the nature of the emergency. For the reasons discussed above, our model distinguishes between different types of emergencies, and each individual ambulance is modeled.

After arriving at the location of the emergency, the ambulance crew treat the patients as the crew deem best, and decide whether to transport patients to a hospital. In some fraction of cases, it is decided not to transport any patients to a hospital, and thus the ambulance becomes available after treatment at the emergency site has been completed. In other cases, the ambulance transports patients to a chosen hospital, and becomes available thereafter. Sometimes ambulances are used for scheduled transportation of patients, for example, to transfer patients from one hospital to another. We model such transportation tasks as a specific type of "emergency". After patients have been transported in an ambulance, the ambulance has to be cleaned. The cleaning is regarded as part of the ambulance's work related to an emergency. After an ambulance has completed its work related to an emergency, a decision has to be made where to dispatch the newly available ambulance to. If any requests are waiting in queue, the ambulance can be dispatched to a chosen waiting emergency. The ambulance can also be used to preempt another ambulance that is on its way to the location of an emergency, for example, if the newly available ambulance is closer to the emergency than the previously dispatched ambulance. Otherwise the ambulance can be sent toward an ambulance staging location, and it is regarded as available during such a journey. There are various types of ambulance staging locations. Most EMSs have one or more facilities that make provision for ambulance maintenance, office space, and training. Such facilities can be used for ambulance staging, but typically a larger number of smaller facilities are also used for ambulance staging. These smaller facilities may make provision for ambulance personnel to relax during less busy times. During busy times, ambulances may use public facilities such as parking lots for ambulance staging. For the purpose of this paper, all ambulance staging locations will be called stations.

This paper proposes an optimization-based method to make two types of decisions mentioned above:

(1) When a request arrives, the decision whether an ambulance should be dispatched to the emergency, and if so, which ambulance to dispatch to the emergency, or whether to add the request to a queue of waiting requests. We will call this decision the *ambulance selection decision*.
(2) When an ambulance becomes available (after completing its service at the location of an emergency or at a hospital), the decision whether to dispatch the ambulance to an emergency waiting in queue, and if so, to which emergency to dispatch the ambulance, or whether to send the ambulance to a station, and if so, to which station to send the ambulance. We will call this decision the *ambulance reassignment decision*.

We do not include the two types of preemption alternatives mentioned above, that is, (1) when a request arrives, the alternative to preempt an ambulance already on its way to an emergency and to dispatch it to the newly arrived emergency, and (2) when an ambulance becomes available, the alternative to preempt an ambulance already on its way to an emergency and to dispatch the newly available ambulance to the emergency (called "diversions" by [51]). [50] simulated and compared four dispatch policies, obtained by switching each of the two alternatives mentioned above (the first alternative was called "Reroute Enabled Dispatch" and the second alternative was called "Free Ambulance Exploitation Dispatch") on and off. The results suggested that these alternatives add very little benefit. In addition, preempting ambulances have practical disadvantages, and thus it is questionable whether use of these alternatives is a wise decision.

1.2. **Related Literature.** There is a large literature on the location and relocation of emergency facilities, including stations and ambulances. Here we give a brief overview of this work with emphasis on the aspects that are relevant for ambulance operations. For surveys, see [12, 62, 9, 29, 49, 3]. A number of static problems have been proposed to choose locations for stations or ambulances. Many of these problems use the notion that a demand point is *covered* if a station or ambulance is located within a specified distance or travel time of the demand point. The location set covering problem (LSCP) was proposed by [71] to determine the minimum number of emergency service facilities (stations) that covers a given set of demand points. A related problem is the maximal covering location problem (MCLP), proposed by [14], to maximize the weighted set of demand points that is covered by a given number of facilities. Many variations of these two models have been proposed to capture aspects that are relevant for the location and relocation of ambulances. Many of these variations consider the policy of assigning ambulances to stations (called the ambulance's "home base" or "depot"), and when an ambulance becomes available and is sent to a station, to always send it to its home base. (In contrast with this policy, and similar to [65], we allow an ambulance to be sent to any station to improve coverage.) For example, [7] used a LSCP to locate stations, and simulation to determine the number of homogeneous ambulances to allocate to each station. [64] proposed two models to choose locations for multiple equipment types. [17] proposed an extension of LSCP that rewards the objective with the number of additional ambulances that can cover a demand point, to make provision for the possibility that some ambulances may be busy when an emergency call is received. Similarly, [36] proposed extensions of MCLP that also reward demand that is covered by more than one facility/ambulance. [22] proposed an extension of MCLP called the double standard model (DSM) that maximizes the demand covered by at least 2 ambulances, subject to 2 constraints on the coverage of all demand.

One shortcoming of the deterministic location problems mentioned above is that they do not explicitly model uncertainty regarding important problem parameters, such as when and where calls arrive, and where the available ambulances are when a call arrives. One of the first related quantities to be modeled as a random variable was whether an ambulance is busy or available when a call arrives. A very influential paper in this regard was [16], which proposed the maximum expected covering location problem (MEXCLP) that uses a binomial distribution for the number of busy ambulances (with the busy/available random variables of different ambulances being independent, with the same busy probability for all ambulances) to explicitly model the probability (or fraction of time) that a demand point is covered by different numbers of ambulances. [63] proposed two versions of the maximum availability location problem (MALP), version I with the same busy probability for all ambulances, and version II that allows different busy probabilities at different stations. The objective of MEXCLP is to maximize the expected demand that is covered by a given number of ambulances, whereas the objective of MALP is to maximize the demand-weighted set of points that is covered with probability at least equal to a specified quantity $\alpha \in (0, 1)$ by a given number of ambulances. [67] proposed a problem with the same objective as MEXCLP, but that allows different busy probabilities at different stations, similar to MALP II, and their simulation results demonstrated that the resulting model consistently produces better solutions than MEXCLP and MALP II. Various further extensions have been proposed. [60] proposed an extension of MEXCLP that models time-varying demand for ambulances. [37] proposed a heuristic for the MEXCLP with random delay times (prior to ambulance travel) and random travel times, and with different busy probabilities for different ambulances. [20] compared the solutions of the MCLP, the MEXCLP, the MCLP with random response times (MCLP+PR), the MEXCLP with random response times (MEXCLP+PR), and the MEXCLP+PR with different busy probabilities for different ambulances, in terms of the fraction of calls with response time less than a threshold, using data from Edmonton, Alberta, Canada.

A few papers proposed ambulance location models that did not use the notion of coverage. For example, [74] proposed a method to locate a given number of ambulances to minimize expected response time. [69] proposed a branch-and-bound approach to locate stations that uses simulation to evaluate the expected response time objective. [70] used simulation and a heuristic, also to minimize expected response time. [21] developed an approximation based on an $M/G/\infty$ queue for the probability that an ambulance is busy, used it to search for ambulance locations that minimize expected response time, and applied it to choose ambulance locations for the city center of Los Angeles.

[43] proposed a continuous-time Markov process called the hypercube queueing model that can be used for performance modeling of emergency operations. In the model there are N distinct servers and multiple demand locations. [43] proposed an algorithm to compute the transition rates for any given fixed preference policy, that is, a policy that specifies for every demand point a preference list of all the servers from most preferred to least preferred (with ties allowed) independent of the state of the process. Then, when a call arrives from a demand point, the most preferred available server in the preference list for that demand point (one of the most preferred available servers in case of ties) is dispatched to serve the call. The algorithm exploits the similarity of the most preferred available servers for adjacent states of the Markov process, that is, states that differ in the availability of only one server, to reduce the effort to compute the transition rates. After the transition rates have been computed, a system of 2^N linear equations can be solved to compute the stationary probabilities, and then various long-run average performance metrics can be computed. [44]

proposed an approximation procedure to compute busy probabilities for the servers of the hypercube queueing model. [35] applied the hypercube queueing model for ambulance location and design of ambulance response districts in Boston. [41] proposed a probability model that uses an approximation procedure similar to that of [44], but unlike the hypercube queueing model, allows general service time distributions that depend on both the station as well as the emergency location. [26, 27] also proposed a probability model that uses an approximation procedure similar to that of [44], but unlike the hypercube queueing model, includes travel time from station to the emergency location as well as on-site service time that is allowed to depend on the emergency location (and, in the case of [26], is also allowed to depend on the emergency type). Then the ambulance location problem was formulated as a mixed integer nonlinear program. [28] proposed and compared methods to solve the system of N nonlinear equations. The hypercube queueing model of [43] can accommodate ties in the given fixed preference policy by enumerating all permutations of the tied servers. However, this is inefficient, and therefore [10] proposed a more efficient model that explicitly formulates the balance equations allowing for ties. [61] proposed two models that use the Erlang loss formula to approximate the fraction of calls with response time more than a threshold and to allocate ambulances to stations.

[19] proposed to use patient survival probability as objective, and they compared, in terms of the expected number of survivors, the solutions of the problems considered in [20] with the solutions of modifications of these problems in which the objective is to maximize patient survival probability. [42] proposed to use different functions of patient survival probability as a function of response time for different patient types (unlike [19] that used the same patient survival probability function for all patients), and they allocated ambulances to stations to maximize the expected total number of survivors. [66] extended the DSM to a multistage model with time-dependent travel times and time-dependent ambulance location decisions. [13] solved the problem of location of both trauma centers and ambulances (in their case, helicopters).

Compared with the work on static location problems for stations and/or ambulances, relatively little work has been done to optimize ambulance operations. In ambulance operations, two types of dispatch decisions mentioned in Section 1.1 are important: (1) the ambulance selection decision, and (2) the ambulance reassignment decision. The solutions of the static location problems mentioned above are sometimes used for ambulance reassignment decisions by assigning a home base to each ambulance, and then when an ambulance becomes available and is not dispatched to a request waiting in queue, the ambulance is sent to its home base [26, 34, 61, 4, 42, 51, 56, 5]. An extension of this approach for ambulance reassignment is to choose which station to send an ambulance to when an ambulance becomes available and is not dispatched to a request waiting in queue. This way ambulances can be dynamically positioned when the ambulances become available to better cover demand points. This approach is followed in [65] and in this paper. A further extension is to also allow the decision to send an ambulance at any time from one station to another station to improve the configuration of available ambulances. Various terms are used for such decisions, including ambulance redeployment, repositioning, relocation, move-up, or system-status management. One approach to ambulance redeployment is to solve, up-front, an ambulance location problem for each possible number of available ambulances, and to store the solutions in a compliance table. Then, whenever the number of available ambulances changes, the available ambulances are redeployed according to the solutions stored in the compliance table [1]. This approach has some shortcomings — it ignores the cost of frequently redeploying the ambulances, which can be considerable [6], and it does not look into the future, for example to take into account ambulances that should become available soon. Many researchers have studied various ambulance redeployment problems, including [23, 9, 24, 2, 57, 55, 65, 53, 54, 18, 38, 6, 72, 73]. As pointed out by [32, 2, 65, 4, 56, 5, 6], there are practical problems associated with redeploying ambulances from one station to another. For example, [65] mentioned that it is illegal in Austria to redeploy ambulances from one station to another. Also, [45] showed, with both simple models and numerical examples, that the mean response time is not very sensitive to the location of ambulances. More specifically, it was shown that the mean travel time obtained by locating emergency facilities randomly is close to the mean travel time obtained by locating emergency facilities optimally. For the reasons mentioned above, when an ambulance becomes available we dynamically choose to which station to send the ambulance, but we do not send an ambulance that is already at one station to another station.

For the ambulance selection decision, the closest available ambulance rule is simple and popular [7, 25, 33, 34, 52, 55, 53, 1]. Some emergency services partition the service region into response areas or districts, and apply some fixed preference policy to make ambulance selection decisions. For example, for each district, a preference list of districts is chosen in advance. Typically, for each district, that district appears first in its preference list. Then, when an emergency call located in a district arrives, the first district in the emergency district's priority list with an available ambulance is determined, and the ambulance in that district closest to the emergency location is dispatched. [11] conducted a detailed study for a setting with 2 ambulances, and characterized the optimal response area for each ambulance. [69] compared two dispatch rules, the closest available ambulance rule, and a service district rule that works as follows: Each ambulance is assigned to a service district, with one ambulance assigned to each district. When an emergency call arrives in a district, if the ambulance assigned to that district is available, then it is dispatched to the call, even if it is temporarily outside its district; otherwise, if the ambulance assigned to that district is not available, then the closest available ambulance is dispatched to the call. [42] proposed a dispatch rule based on a static preference matrix ρ. The service region is partitioned into "demand nodes", and each available ambulance is at a station. Then $\rho_{i,j}$ denotes the jth most preferred station to use for an emergency at demand node i. An ambulance is dispatched from station $\rho_{i,j}$ to an emergency at demand node i if and only if there is no available ambulance at stations $\rho_{i,1}, \ldots, \rho_{i,j-1}$ and there is at least one available ambulance at station $\rho_{i,j}$.

[56] proposed a heuristic to partition the service region into districts, with a number of ambulances in each district. They used simulation to compare the performance of four dispatch policies for a setting with two priority levels; two types of policies specifying ambulance selection decisions if there is an ambulance available in the same district as the emergency, combined with two types of policies specifying ambulance selection decisions if there is no ambulance available in the same district. If there is an ambulance available in the same district as the emergency, then the first type of policy evaluated dispatches the closest available ambulance within the same district, and the second type of policy evaluated applies a heuristic ambulance selection rule to each district. If there is no ambulance available in the same district as the emergency, then the first type of policy evaluated assumes that an alternate emergency response, for example provided by the fire department, is automatically dispatched within the same district, and the second type of policy evaluated dispatches an ambulance from another district using a preference list of ambulances. Under both policies, if an emergency call arrives and all ambulances are busy, then the emergency is handled by an outside service, that is, there is no queue of waiting requests in the simulation. Also, the simulation returns an ambulance to its home base when it becomes available. Therefore, the simulation includes an ambulance selection decision and a very restrictive ambulance reassignment decision.

Various dispatch policies that require real-time computation have been proposed. [2] proposed heuristics for ambulance dispatch and relocation for a system with three priority levels, based on a measure of "preparedness" for each zone. For priority 1 calls, they dispatch the closest available ambulance. They allow preemption of ambulances already dispatched to lower priority calls. For priority 2 and 3 calls, they dispatch the ambulance with expected travel time less than a specified threshold that will result in the least decrease in the minimum preparedness measure over all zones. For the available ambulance relocation decision, they proposed an integer nonlinear program that minimizes the relocation time subject to constraints that specify the maximum number of ambulances that may be relocated and the minimum preparedness measure after relocation. They did not specify how they select towards which request waiting in queue to dispatch a newly available ambulance. [46] considered the same preparedness measure as [2], and showed that it resulted in worse performance than the closest-available-ambulance rule. Then two modifications of the preparedness-based dispatching rule were proposed. The first modification dispatches the available ambulance that maximizes the minimum preparedness measure over all zones divided by the travel time from the ambulance to the emergency location. The second modification replaces the minimum preparedness measure over all zones in the calculations with other aggregates of the preparedness measures of different zones. If an ambulance becomes available and there are requests waiting in queue, then the newly available ambulance is dispatched to the request in queue closest to the ambulance. [47] proposed a rule for the ambulance reassignment decision that takes into account both the distances or times between the newly available ambulance and requests waiting in queue, as well as a centrality measure of each request waiting in queue. It was not specified how to select a station to which to send the newly available ambulance if there are no requests waiting in queue, or how the ambulance selection decision is made. [65] proposes a dynamic programming formulation for the ambulance selection decision and the ambulance reassignment decision, with the restriction that if an ambulance becomes available and there are requests waiting in queue, then the newly available ambulance is dispatched to the next request in first-come-first-served order. An approximate dynamic programming method is used to produce solutions, and it is shown that after sufficient training these solutions outperform policies that combine the closest-available-ambulance dispatching rule with the rule to send a newly available ambulance to its home base (or to a random station) if there are no requests waiting in queue. [4] formulated an ambulance dispatching problem with two priority levels and exponentially distributed service times as a continuous-time Markov Decision Process (MDP), and showed (for a sufficiently small number of ambulances to enable solving the MDP) that the closest-available-ambulance dispatching policy is suboptimal. It was assumed that if an emergency call arrives and all ambulances are busy, then the emergency is handled by an outside service, that is, there is no queue of waiting requests in the model. It was also assumed that an ambulance returns to its home base when it becomes available. Therefore, the MDP includes an ambulance selection decision and a very restrictive ambulance reassignment decision. Similarly, [5] proposed an ambulance dispatching heuristic that takes emergency priorities into account, and uses simulation to compare the performance of the heuristic and the closest-available-ambulance dispatching rule for a setting with two priority levels. The same assumptions as in [4] are made, and therefore the model also includes an ambulance selection decision and a very restrictive ambulance reassignment decision. [48] used simulation to compare two dispatch policies, a policy that dispatches the closest available ambulance to all calls, and a policy that dispatches the closest available ambulance to priority 1 calls, and the ambulance within a specified response time radius which has the least utilization to priority 2 and 3 calls. It was assumed that if an emergency call arrives and all ambulances are busy, then the call is lost. It was also assumed that an ambulance returns to its home base when it becomes available. Therefore, the simulation also includes an ambulance selection decision and a very restrictive ambulance reassignment decision. [39] compared two dispatch policies with the closest-available-ambulance policy. In all policies, ambulance reassignment decisions are made according to the first-come first-served rule: if an ambulance becomes available and there are calls waiting in queue, then the ambulance is dispatched to the waiting call that first entered the queue. If an ambulance becomes available and there is no call waiting in queue, then the ambulance returns to its home base. One dispatch policy is based on a Markov decision process with a simplified state, that represents the location of the currently considered emergency, and the set of available ambulances, assuming that each available ambulance is at its home base. The second dispatch policy, called DMEXCLP, works as follows: for each available ambulance that can reach the emergency location within the threshold time, the objective value of the MEXCLP with the available ambulances excluding that ambulance is computed. Then the available ambulance that can reach the emergency location within the threshold time with the largest objective value of the MEXCLP without that ambulance is dispatched. If no available ambulance can reach the emergency location within the threshold time, then the available ambulance, irrespective of travel time to the emergency location, with the largest objective value of the MEXCLP without that ambulance is dispatched. Simulation results showed that the heuristic has a much lower fraction of late arrivals than the closest-available-ambulance policy, but that the heuristic also has a much greater mean response time than the closest-available-ambulance policy. [40] assumed that all ambulances are dispatched from their home bases, that there is always an ambulance available at its home base when an emergency call arrives, and that an ambulance must always be dispatched immediately to an emergency. They also assumed that after service each ambulance returns to its home base, and that the time that elapses from the moment the ambulance arrives at the emergency location until the ambulance is back at its home base is deterministic and is the same for all emergencies and all ambulances, that is, the elapsed time does not depend on the emergency location or the ambulance home base location. Thus they considered a very restrictive ambulance reassignment decision. They showed with an example that for any online policy the ratio of fraction of late arrivals to the offline minimum fraction of late arrivals can be arbitrarily large, and thus no online policy can have a finite competitive ratio. They also used simulation to compare the expected performance ratios of the closest-available-ambulance policy and of DMEXCLP, and showed that the expected performance ratio of DMEXCLP is better than that of the closest-available-ambulance policy.

1.3. **Contributions.** The contributions of this paper are as follows.

(A) Modeling the operation of an ambulance fleet under uncertainty. We propose a model for optimizing ambulance dispatch decisions, including ambulance selection decisions and ambulance reassignment decisions. In particular, the proposed model has the following features:

(1) The model takes into account the type of each emergency. The type of emergency affects the type of ambulance and crew needed, the marginal value of response time, and the set of appropriate hospitals for the emergency. As pointed out in Section 1.2, most existing models consider only one emergency type, that is, the effects of emergency type mentioned above are ignored. Some existing models distinguish a small number (2 or 3) priority levels. However, as pointed out above, the emergency type has more dimensions than just priority level.
(2) The model considers each ambulance and crew as unique. As pointed out in Section 1.1, each ambulance and crew has unique capabilities and skills. For example, some crews may have special training or experience in the handling of stroke victims, and some crews may be skilled in coping in dangerous situations such as rioting. As pointed out in Section 1.2, most existing models consider ambulances as interchangeable, that is, the only attributes of ambulances and crews taken into account are the availability and location of the ambulance. A small number of existing models make a distinction between BLS and ALS ambulances.
(3) The model allows the hospital for each emergency to be chosen, taking into account the emergency type and the location of the emergency relative to hospitals. Most existing models ignore choices among multiple hospitals; in fact, most existing models either have no hospital entity in the model, or have a single "hospital" entity irrespective of the type and location of the emergency.
(4) The model makes provision for a queue of waiting emergencies. This is both necessary if all ambulances are busy, and desirable if the emergency is not urgent and few ambulances are available. Many existing models ignore the possibility of a queue of waiting emergencies. Instead, it is assumed that if all ambulances are busy then some unlimited outside service will take care of the emergency.
(5) The model makes provision for both ambulance selection decisions as well as nontrivial ambulance reassignment decisions. When an ambulance becomes available, it can be assigned to an emergency in queue, and it can also be sent to a chosen station. Since many existing models ignore the possibility of a queue of waiting emergencies, such models cannot accommodate decisions to assign available ambulances to emergencies in queue. Also, many existing models assume that when an ambulance becomes available, it will go to its home base, even if sending it to a different station would improve coverage greatly.
(6) The model allows ambulances on their way to a station to be dispatched to an emergency. Because of the triangle inequality, that would result in a smaller response time than waiting until the ambulance reaches the station before dispatching the ambulance to the emergency. Also, with modern communication technology dispatchers are in constant contact with ambulances, and such en-route dispatching is easy to execute. As pointed out in Section 1.2, most existing models allow only ambulances at stations to be dispatched.
(7) Dispatch decisions have consequences not only for the emergency and the ambulance under consideration, but also for future emergencies and for other ambulances that have to take care of future emergencies. It is challenging to take these future consequences of dispatch decisions into account, because future consequences are a complicated function of current decisions, and because the future consequences are uncertain. The model takes into account that future consequences are uncertain. Most existing models either do not take uncertainty into account, or incorporate uncertainty into a simulation model.

(B) Solution method for solving the optimization problems to dispatch ambulances under uncertainty. In principle, the problem of optimizing ambulance operations can be formulated as a Markov decision process or as a multistage stochastic integer program. However, these problems would be intractable. Therefore, we propose to use the following rolling horizon approach (see for instance [31], [30]). Each time an ambulance selection decision or an ambulance reassignment decision has to be made, a two-stage stochastic optimization problem is formulated and solved. The first-stage decisions are either ambulance selection decisions or ambulance reassignment decisions, as appropriate for the decision at hand. The second-stage decisions are sequences of ambulance selection decisions and ambulance reassignment decisions over a considered time horizon. For each first-stage decision, the setting for the decision is known. For example, if an ambulance selection decision has to be made then the type and location of the newly arrived emergency is known, or if an ambulance reassignment decision has to be made then the location of the newly available ambulance is known, and in both cases the current state of the system is known. In contrast, the settings

for second-stage decisions are not known yet, and the uncertainty is represented with a set of second-stage scenarios. The number of first-stage alternatives is relatively small. If an ambulance selection decision has to be made, then the first-stage alternatives correspond to the available compatible ambulances combined with the candidate hospitals, as well as the alternative to add the emergency to the queue. If an ambulance reassignment decision has to be made, then the first-stage alternatives correspond to the emergencies in queue as well as the stations. Therefore, each two-stage problem can be solved by computing the expected second-stage cost for each feasible first-stage alternative, and then choosing the first-stage alternative with the best combination of first-stage and second-stage cost. Therefore, to solve the two-stage problems fast enough, we need to be able to quickly compute the expected second-stage cost for every feasible first stage decision. The second-stage cost is given by the mean cost over a finite set of scenarios, and its computation requires solving a sequence of ambulance selection problems and ambulance reassignment problems for each scenario. That is, the second-stage cost can be computed by separately solving a deterministic sequence of ambulance selection problems and ambulance reassignment problems for each scenario. To facilitate faster selection of first-stage decisions, we consider a continuous relaxation of the second-stage problem for each scenario. In addition, due to the huge number of second-stage decision variables for each scenario, we solve the continuous relaxation of the second-stage problem with column generation, and we devise an approach to quickly solve the column generation subproblems to find decision variables with negative reduced cost.
(C) Numerical tests with real emergency medical service data. We used more than 2 years of emergency call data of the Rio de Janeiro emergency medical service to calibrate models of emergency call arrivals. We used the resulting models to compare the performance of the proposed ambulance dispatch policies with the closest-available-ambulance rule (for the ambulance selection problem) and the closest-station rule (for the ambulance reassignment problem). The results showed that the proposed dispatch policy yields smaller response times. The results also showed the benefit of the proposed column generation algorithm to solve the two-stage stochastic problems. In addition, it was observed that reasonable response times can be obtained even with quite a small number of ambulances and stations.

The rest of this paper is organized as follows. In Section 2, we specify the optimization models. In Section 3, we explain how these problems are solved. The numerical results for the Rio de Janeiro EMS are presented in Section 4.

2. Optimization Models

In this section we describe the two optimization problems to be solved in rolling horizon fashion to make ambulance dispatch decisions:

(1) The ambulance selection problem: When a request arrives, the problem to decide whether an ambulance should be dispatched to the emergency, and if so, which ambulance to dispatch to the emergency, or whether to add the request to a queue of waiting requests.
(2) The ambulance reassignment problem: When an ambulance becomes available, the problem to decide whether to dispatch the ambulance to an emergency waiting in queue, and if so, to which emergency to dispatch the ambulance, or whether to send the ambulance to a station, and if so, to which station to send the ambulance.

Both models "look ahead" until the end of a chosen time horizon (for example, a few hours or until the end of the day) to approximate the impact of current decisions on the objective function in the future. Both problems minimize a combination of the cost of the immediate decision and the expected future costs affected by the immediate decision over the planning horizon.

First we present a deterministic formulation of the problem, as though the arrival times, locations, and types of the emergency calls over the planning horizon are known. In Section 2.2 we describe the extension of this model to incorporate random arrival times, locations, and types of emergency calls.

2.1. Deterministic Models.

Ambulances can be dispatched from stations, from emergency locations (if the ambulance is not needed to transport patients to a hospital), from hospitals, and from intermediate locations while traveling towards a station. The models do not allow ambulances to be dispatched while busy with service — while traveling towards an emergency location, or while providing on-site emergency care, or while traveling with patient(s) towards a hospital. That is, we do not model preemption or "forward" dispatching of ambulances. An ambulance that is not busy with service must either be at a station, or traveling toward

Number of time steps in the time horizon	T
Set of locations	$\mathcal{L}$
Set of emergency call types	$\mathcal{C}$
Set of ambulance types	$\mathcal{A}$
Set of stations	$\mathcal{B}$
Set of hospitals	$\mathcal{H}$
Set of ambulance types serving call type $c \in \mathcal{C}$	$\mathcal{A}(c)$
Set of candidate hospitals for call type $c \in \mathcal{C}$ at location $\ell \in \mathcal{L}$	$\mathcal{H}(c, \ell)$
Set of emergency call types that can be served by ambulance type $a \in \mathcal{A}$	$\mathcal{C}(a)$
Location of the emergency call at $t = 0$	ℓ_0
Type of the emergency call at $t = 0$	c_0
Type of the ambulance which becomes available at $t = 0$	a_0
Hospital where the ambulance becomes available at $t = 0$	h_0
Number of calls of type c at location ℓ in queue at $t = 0$	$C_0(c, \ell)$
Number of ambulances of type a available at station b at $t = 0$	$A_0(a, b)$
Number of ambulances of type a available at hospital h at $t = 0$	$A_0(a, h)$
Location at time $t + 1$ of an ambulance of type a which is at ℓ_1 at time t on its way to ℓ_2	$L(t, a, \ell_1, \ell_2)$
Set of intermediate locations that can be reached at time t by an ambulance of type a going from a hospital to station b	$\mathcal{L}(t, a, b)$
Number of ambulances of type a at location ℓ_1 at $t = 0$ going to station b	$A_0(a, \ell_1, b)$
Number of ambulances of type a at location ℓ_1 at $t = 0$ going to an emergency of type c at location ℓ and then going to hospital h	$A_0(c, a, \ell_1, \ell, h)$
Number of ambulances of type a at location ℓ_1 at $t = 0$ traveling to hospital h with emergency type c patient after on-site emergency care has been provided	$A_0(c, a, \ell_1, h)$
Number of emergency calls of type c for period t and location ℓ	$\lambda(t, c, \ell)$
Time for an ambulance of type a to go from location ℓ_1 at time t to location ℓ, provide on-site emergency care for an emergency of type c at location ℓ, travel from location ℓ to hospital h and deliver the patient to hospital h	$\tau(t, c, a, \ell_1, \ell, h)$
Time for an ambulance of type a to travel with a patient of type c from location ℓ_1 at time $t = 0$, after on-site emergency care has been provided, to hospital h and deliver the patient at hospital h	$\tau_0(c, a, \ell_1, h)$
Maximum number of ambulances at station b	$A_{\max}(b)$

TABLE 1. Problem input parameters

a station. If ambulances are allowed to wait at a hospital for a dispatch, then the hospital is also a station in the model.

2.1.1. *Problem input parameters.* In this section we describe the input parameters of the models. These parameters are summarized in Table 1.

Basic problem parameters. Both time and space are discretized for the model. Let $t = 0$ denote the current time (which can be any time instant during the current day) and let $t = 1, \ldots, T$ denote the time steps until the end T of the time horizon. Let $\mathcal{L}$ denote the set of discrete locations, used for representing emergency call locations as well as ambulance locations. Each emergency call is characterized by its arrival time, its location, and its type. Let $\mathcal{C}$ denote the set of call types, let $\mathcal{A}$ denote the set of ambulance types, let $\mathcal{B}$ denote the set of ambulance stations, and let $\mathcal{H}$ denote the set of hospitals. A call of type $c \in \mathcal{C}$ can be served by a subset $\mathcal{A}(c) \subset \mathcal{A}$ of ambulance types, and a call of type $c \in \mathcal{C}$ at location $\ell \in \mathcal{L}$ can be sent to a subset $\mathcal{H}(c, \ell) \subset \mathcal{H}$ of hospitals. For each ambulance type $a \in \mathcal{A}$, let $\mathcal{C}(a) := \{c \in \mathcal{C} \; : \; a \in \mathcal{A}(c)\}$ denote the set of call types that can be served by ambulance type a.

Initial conditions. Other input includes the initial conditions for the problem. If a call arrives at time 0, then the location and the type of the call that has just arrived are denoted by ℓ_0 and c_0 respectively. If an ambulance completes service at a hospital at time 0, then the ambulance type and the hospital are denoted by a_0 and h_0 respectively. For each emergency type $c \in \mathcal{C}$ and emergency location $\ell \in \mathcal{L}$, let $C_0(c, \ell)$ denote the number of calls at time 0 waiting in queue for an ambulance to be dispatched to serve the call. This includes the call that has just arrived at $t = 0$ at location ℓ_0.

For each ambulance type $a \in \mathcal{A}$ and ambulance station $b \in \mathcal{B}$, let $A_0(a, b)$ denote the number of ambulances of type a at b available for dispatch just before the dispatch of ambulances at time 0. Similarly, for each ambulance type $a \in \mathcal{A}$ and hospital $h \in \mathcal{H}$, let $A_0(a, h)$ denote the number of ambulances of type a at h available for dispatch just before the dispatch of ambulances at time 0, including the ambulance of type a_0 at hospital h_0 that just became available.

Keeping track of ambulance locations. As mentioned above, ambulances can also be dispatched while traveling towards a station. To model this, and in general to keep track of ambulance locations both while stationary and while moving, it is useful to determine where ambulances can be while traveling to specific destinations. First, for each time $t \in \{0, 1, \ldots, T-1\}$, ambulance type $a \in \mathcal{A}$, hospital $h \in \mathcal{H}$, and destination station $b \in \mathcal{B}$, let $L(t, a, h, b) \in \mathcal{L}$ denote the forecasted location of the ambulance at time $t+1$ if the ambulance starts from h at t to travel towards b. In addition, for each current ambulance location $\ell_1 \in \mathcal{L}$ at time t, and destination station $b \in \mathcal{B}$, let $L(t, a, \ell_1, b) \in \mathcal{L}$ denote the forecasted location of the ambulance at time $t+1$ if the ambulance continues to travel towards b. Next, for each time $t \in \{0, 1, \ldots, T\}$, let

$$\mathcal{L}_1(t, a, b) \quad := \quad \bigcup_{h \in \mathcal{H}} \{L(t-1, a, h, b)\} \setminus \{b\}$$

denote the set of all intermediate locations that can be reached by an ambulance at time t if the ambulance starts traveling at time $t-1$ from some hospital towards b. By induction, for each $\tau \geq 2$, let

$$\mathcal{L}_\tau(t, a, b) \quad := \quad \bigcup_{\ell_1 \in \mathcal{L}_{\tau-1}(t-1, a, b)} \{L(t-1, a, \ell_1, b)\} \setminus \{b\}$$

denote the set of all intermediate locations that can be reached by an ambulance at time t if the ambulance starts traveling at time $t - \tau$ from some hospital towards b. Then, for each time $t \in \{0, 1, \ldots, T\}$, ambulance type $a \in \mathcal{A}$, and destination station $b \in \mathcal{B}$, let

$$\mathcal{L}(t, a, b) \quad := \quad \bigcup_{\tau \geq 1} \mathcal{L}_\tau(t, a, b)$$

denote the set of all intermediate locations that can be reached by an ambulance at time t if the ambulance travels from some hospital towards b.

For each ambulance type $a \in \mathcal{A}$, station $b \in \mathcal{B}$, and initial ambulance location $\ell_1 \in \mathcal{L}(0, a, b)$, let $A_0(a, \ell_1, b)$ denote the number of ambulances of type a at location ℓ_1 traveling towards b available for dispatch just before the dispatch of ambulances at time 0. In addition, for each call type $c \in \mathcal{C}$, ambulance type $a \in \mathcal{A}(c)$, initial ambulance location $\ell_1 \in \mathcal{L}$, emergency location $\ell \in \mathcal{L}$, and hospital $h \in \mathcal{H}(c, \ell)$, let $A_0(c, a, \ell_1, \ell, h)$ denote the number of ambulances of type a at location ℓ_1 at time 0 traveling to an emergency type c at location ℓ and from there to hospital h. Also, let $A_0(c, a, \ell_1, h)$ denote the number of ambulances of type $a \in \mathcal{A}(c)$ at location $\ell_1 \in \mathcal{L}$ at time 0 traveling with emergency type $c \in \mathcal{C}$ patient(s) after on-site emergency care has already been provided, to hospital $h \in \cup_{\ell \in \mathcal{L}} \mathcal{H}(c, \ell)$.

Forecasts of emergency calls, service times, and travel times For each time $t \in \{1, \ldots, T\}$, call type $c \in \mathcal{C}$, and emergency location $\ell \in \mathcal{L}$, let $\lambda(t, c, \ell)$ denote the forecasted number of calls of type c at location ℓ in time period t. For each dispatch time $t \in \{0, 1, \ldots, T\}$, call type $c \in \mathcal{C}$, ambulance type $a \in \mathcal{A}(c)$, initial ambulance location $\ell_1 \in \cup_{b \in \mathcal{B}} \mathcal{L}(t, a, b) \cup \mathcal{B} \cup \mathcal{H}$, emergency location $\ell \in \mathcal{L}$, and hospital $h \in \mathcal{H}(c, \ell)$, let $\tau(t, c, a, \ell_1, \ell, h)$ denote the forecasted time for ambulance type a to travel from ℓ_1 at time t to ℓ, provide on-site emergency care for call type c at ℓ, travel with patient(s) from ℓ to hospital h, and deliver the patient(s) at h. Also, for each call type $c \in \mathcal{C}$, ambulance type $a \in \mathcal{A}(c)$, initial ambulance location $\ell_1 \in \mathcal{L}$, and hospital $h \in \cup_{\ell \in \mathcal{L}} \mathcal{H}(c, \ell)$, let $\tau_0(c, a, \ell_1, h)$ denote the forecasted time for ambulance type a to travel with patient(s) from ℓ_1 at time 0 after on-site emergency care has already been provided, to hospital h, and deliver the patient(s) of type c at h.

2.1.2. *Decision variables for the ambulance selection problem.* The following first-stage (for $t = 0$) decision variables are used for the ambulance selection problem:

- $x_0(c, a, b, \ell, h)$ = the number of ambulances of type $a \in \mathcal{A}(c)$ dispatched at time 0 from station $b \in \mathcal{B}$ to serve calls of type $c \in \mathcal{C}$ at location $\ell \in \mathcal{L}$, and transport them to hospital $h \in \mathcal{H}(c, \ell)$; this includes both calls in queue as well as the call that arrived at time 0;
- $x_0(c, a, \ell_1, b, \ell, h)$ = the number of ambulances of type $a \in \mathcal{A}(c)$ at location $\ell_1 \in \mathcal{L}(0, a, b)$ at time 0 traveling toward station $b \in \mathcal{B}$ dispatched to serve calls of type $c \in \mathcal{C}$ that arrived at location $\ell \in \mathcal{L}$ at time 0, and transport them to hospital $h \in \mathcal{H}(c, \ell)$.

The following second-stage (for $t = 1, \ldots, T$) decision variables are used for the ambulance selection problem:

- $x_t(c, a, b, \ell, h)$ = the number of ambulances of type $a \in \mathcal{A}(c)$ dispatched at time t from station $b \in \mathcal{B}$ to serve calls of type $c \in \mathcal{C}$ at location $\ell \in \mathcal{L}$, and transport them to hospital $h \in \mathcal{H}(c, \ell)$; this includes both calls in queue as well as the calls that arrived at time t;
- $x_t(c, a, \ell_1, b, \ell, h)$ = the number of ambulances of type $a \in \mathcal{A}(c)$ at location $\ell_1 \in \mathcal{L}(t, a, b)$ at time t traveling toward station $b \in \mathcal{B}$ dispatched to serve calls of type $c \in \mathcal{C}$ that arrived at location $\ell \in \mathcal{L}$ and transport them to hospital $h \in \mathcal{H}(c, \ell)$;
- $x_t(c, a, h', \ell, h)$ = the number of ambulances of type $a \in \mathcal{A}(c)$ dispatched at time t from hospital $h' \in \mathcal{H}$ to serve calls of type $c \in \mathcal{C}$ at location $\ell \in \mathcal{L}$, and transport them to hospital $h \in \mathcal{H}(c, \ell)$;
- $y_t(a, h, b)$ = the number of ambulances of type $a \in \mathcal{A}$ instructed at time t to move from hospital $h \in \mathcal{H}$ towards station $b \in \mathcal{B}$;
- $C_t(c, \ell)$ = the number of calls of type $c \in \mathcal{C}$ waiting in queue at location $\ell \in \mathcal{L}$ at the beginning of time t;
- $A_t(a, b)$ = the number of ambulances of type $a \in \mathcal{A}$ at station $b \in \mathcal{B}$ at the beginning of time t;
- $A_t(a, \ell_1, b)$ = the number of ambulances of type $a \in \mathcal{A}$ at location $\ell_1 \in \mathcal{L}(t, a, b)$ moving towards station $b \in \mathcal{B}$ at the beginning of time t.

Next we make a few remarks regarding these decision variables.

Remark 2.1. *First, recall that the decision variables above are written for a deterministic problem, but these variables will be used as part of a stochastic problem. If the realized emergency calls during the time horizon coincided with the forecasted calls, then an optimal solution of the deterministic optimization problem would give optimal dispatch decisions for the entire time horizon. However, typically the realized calls will not coincide with the forecasted calls, in which case the optimal first-stage decision variables give a useful immediate dispatch decision but the optimal second-stage decision variables may not be useful as decisions — the purpose of the second-stage model is to approximate the effect of the first-stage decisions on future costs, and not to fix useful decisions for the future.*

Second, recall that times $t = 1, 2, \ldots, T$, correspond to a discretization of the time horizon, but that $t = 0$ can be any time when an emergency call arrives (for the ambulance selection problem) or when an ambulance becomes available (for the ambulance reassignment problem). Therefore, we assume that it does not happen simultaneously that an emergency call arrives and an ambulance becomes available. This explains why at (each) time $t = 0$, either an ambulance selection problem or an ambulance reassignment problem is considered.

Third, first-stage decision variables $x_0(c, a, b, \ell, h)$ and $x_0(c, a, \ell_1, b, \ell, h)$ can be restricted to pairs (c, ℓ) corresponding to calls in queue at $t = 0$. Although these restrictions are implemented in code, for simplicity of exposition we do not introduce notation for such restrictions. When there is no call in queue for a given pair (c, ℓ), then the constraints below will imply that the corresponding variables $x_0(c, a, b, \ell, h)$ and $x_0(c, a, \ell_1, b, \ell, h)$ are zero.

2.1.3. *Decision variables for the ambulance reassignment problem.* The following first-stage decision variables are used for the ambulance reassignment problem:

- $x_0(c, a, h, \ell, h')$ = the number of ambulances of type $a \in \mathcal{A}(c)$ dispatched at time 0 from hospital $h \in \mathcal{H}$ to serve calls of type $c \in \mathcal{C}$ at location $\ell \in \mathcal{L}$, and transport them to hospital $h' \in \mathcal{H}(c, \ell)$;
- $y_0(a, h, b)$ = the number of ambulances of type $a \in \mathcal{A}$ instructed at time 0 to move from hospital $h \in \mathcal{H}$ towards station $b \in \mathcal{B}$.

The second-stage (for $t = 1, \ldots, T$) decision variables for the ambulance reassignment problem as the same as the decision variables for the ambulance selection problem.

Next we make a few remarks regarding the decision variables for the ambulance reassignment problem.

Remark 2.2. *First-stage decision variables $x_0(c, a, h, \ell, h')$ and $y_0(a, h, b)$ are needed for $a = a_0$ and $h = h_0$ only. In addition, similar to the third point of Remark 2.1, first-stage decision variables $x_0(c, a, h, \ell, h')$ can be restricted to pairs (c, ℓ) corresponding to calls in queue at $t = 0$. Although these restrictions are implemented in code, for simplicity of exposition we do not introduce notation for such restrictions. When $a \neq a_0$ or $h \neq h_0$, then $A_0(a, h) = 0$ and the constraints below will imply that the corresponding variables $x_0(c, a, h, \ell, h')$ and $y_0(a, h, b)$ are zero, and when there is no call in queue for a given pair (c, ℓ), then the constraints below will imply that the corresponding variables $x_0(c, a, h, \ell, h')$ are zero.*

2.1.4. *Constraints for the ambulance selection problem.* The following five sets of constraints apply to the first-stage variables for the ambulance selection problem only:

(S1) Flow balance equations at the stations: For each $a \in \mathcal{A}$, $b \in \mathcal{B}$,

$$
\begin{aligned}
A_1(a, b) \quad &= \quad A_0(a, b) - \sum_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c, \ell)} x_0(c, a, b, \ell, h) \\
&\quad + \sum_{\{\ell_1 \in \mathcal{L}(0, a, b) \,:\, L(0, a, \ell_1, b) = b\}} \left[A_0(a, \ell_1, b) - \sum_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c, \ell)} x_0(c, a, \ell_1, b, \ell, h) \right].
\end{aligned} \tag{2.1}
$$

The left side in (2.1) is the number of ambulances of type a at station b at the beginning of period $t = 1$. This is equal to the initial number $A_0(a, b)$ of ambulances of type a at station b available for dispatch minus the number of ambulances of type a that leave station b in the first stage plus the number of ambulances of type a that arrive at station b during the first stage. This latter is the number of ambulances of type a that would have arrived at station b during the first stage based on their status at $t = 0$ minus the number of these ambulances that are dispatched while en-route to station b to attend a call. The remaining flow constraints follow the same logic and therefore are given without detailed explanation.

(S2) Flow balance equations at the locations between hospitals and stations: For each $a \in \mathcal{A}$, $b \in \mathcal{B}$, $\ell_1 \in \mathcal{L}(0, a, b)$,

$$
A_1(a, \ell_1, b) \quad = \quad \sum_{\{\ell_1' \in \mathcal{L}(0, a, b) \,:\, L(0, a, \ell_1', b) = \ell_1\}} \left[A_0(a, \ell_1', b) - \sum_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{L}} \sum_{h \in H(c, \ell)} x_0(c, a, \ell_1', b, \ell, h) \right]. \tag{2.2}
$$

(S3) Flow balance equations for the queues: For each $c \in \mathcal{C}$, $\ell \in \mathcal{L}$,

$$
\begin{aligned}
C_1(c, \ell) \quad &= \quad C_0(c, \ell) + \lambda(0, c, \ell) - \sum_{a \in \mathcal{A}(c)} \sum_{b \in \mathcal{B}} \sum_{h \in \mathcal{H}(c, \ell)} x_0(c, a, b, \ell, h) \\
&\quad - \sum_{a \in \mathcal{A}(c)} \sum_{b \in \mathcal{B}} \sum_{\ell_1 \in \mathcal{L}(0, a, b)} \sum_{h \in H(c, \ell)} x_0(c, a, \ell_1, b, \ell, h).
\end{aligned} \tag{2.3}
$$

(S4) Initial ambulance supply at locations constraints: For each $a \in \mathcal{A}$, $b \in \mathcal{B}$, $\ell_1 \in \mathcal{L}(0, a, b)$,

$$
\sum_{c \in \mathcal{C}(a)} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c, \ell)} x_0(c, a, \ell_1, b, \ell, h) \quad \leq \quad A_0(a, \ell_1, b). \tag{2.4}
$$

(S5) Initial ambulance supply at stations constraints: For each $a \in \mathcal{A}$, $b \in \mathcal{B}$,

$$
\sum_{c \in \mathcal{C}(a)} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c, \ell)} x_0(c, a, b, \ell, h) \quad \leq \quad A_0(a, b). \tag{2.5}
$$

For $t = 1, \ldots, T$, the following constraints apply:

(At1) Flow balance equations at the stations: For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $b \in \mathcal{B}$,

$$
\begin{aligned}
A_{t+1}(a,b) &= A_t(a,b) + \sum_{\{h \in \mathcal{H} \,:\, L(t,a,h,b)=b\}} y_t(a,h,b) \\
&\quad + \sum_{\{\ell_1 \in \mathcal{L}(t,a,b) \,:\, L(t,a,\ell_1,b)=b\}} \left[A_t(a,\ell_1,b) - \sum_{c \in \mathcal{C}(a)} \sum_{\ell \in \mathcal{L}} \sum_{h \in H(c,\ell)} x_t(c,a,\ell_1,b,\ell,h) \right] \\
&\quad - \sum_{c \in \mathcal{C}(a)} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c,\ell)} x_t(c,a,b,\ell,h).
\end{aligned} \tag{2.6}
$$

(At2) Flow balance equations at the hospitals: For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $h \in \mathcal{H}$,

$$
\begin{aligned}
& \sum_{c \in \mathcal{C}(a)} \sum_{\ell \in \mathcal{L}} \sum_{h' \in \mathcal{H}(c,\ell)} x_t(c,a,h,\ell,h') + \sum_{b \in \mathcal{B}} y_t(a,h,b) \\
= & \sum_{c \in \mathcal{C}(a)} \sum_{\ell_1 \in \mathcal{L}} \sum_{\{\ell \in \mathcal{L} \,:\, \tau(0,c,a,\ell_1,\ell,h)=t\}} A_0(c,a,\ell_1,\ell,h) \\
& + \sum_{c \in \mathcal{C}(a)} \sum_{\{\ell_1 \in \mathcal{L} \,:\, \tau_0(c,a,\ell_1,h)=t\}} A_0(c,a,\ell_1,h) \\
& + \sum_{c \in \mathcal{C}(a)} \sum_{\{\ell \in \mathcal{L} \,:\, h \in \mathcal{H}(c,\ell)\}} \sum_{b \in \mathcal{B}} \sum_{\{t' \in \{0,\ldots,t-1\} \,:\, t'+\tau(t',c,a,b,\ell,h)=t\}} x_{t'}(c,a,b,\ell,h) \\
& + \sum_{c \in \mathcal{C}(a)} \sum_{\{\ell \in \mathcal{L} \,:\, h \in \mathcal{H}(c,\ell)\}} \sum_{h' \in \mathcal{H}} \sum_{\{t' \in \{1,\ldots,t-1\} \,:\, t'+\tau(t',c,a,h',\ell,h)=t\}} x_{t'}(c,a,h',\ell,h) \\
& + \sum_{c \in \mathcal{C}(a)} \sum_{\{\ell \in \mathcal{L} \,:\, h \in \mathcal{H}(c,\ell)\}} \sum_{b \in \mathcal{B}} \sum_{\ell_1 \in \mathcal{L}} \sum_{\{t' \in \{0,\ldots,t-1\} \,:\, \ell_1 \in \mathcal{L}(t',a,b),\, t'+\tau(t',c,a,\ell_1,\ell,h)=t\}} x_{t'}(c,a,\ell_1,b,\ell,h).
\end{aligned} \tag{2.7}
$$

The left side of (2.7) incorporates the requirement that every ambulance that finishes service at a hospital in time period t has to be sent either to a call in queue or to a station (even if the station is at the hospital itself). The right side of (2.7) represents the number of ambulances of type a that finish service at a hospital h in time period t. In particular, the terms $A_0(c,a,\ell_1,\ell,h)$ and $A_0(c,a,\ell_1,h)$ represent the number of ambulances of type a in service at $t = 0$ (either going to an emergency and then to hospital h or traveling to hospital h after on-site emergency care has been provided) that finish that service in time period t.

(At3) Flow balance equations at the locations between hospitals and stations: For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $b \in \mathcal{B}$, $\ell_1 \in \mathcal{L}(t,a,b)$,

$$
\begin{aligned}
A_{t+1}(a,\ell_1,b) &= \sum_{\{h \in \mathcal{H} \,:\, L(t,a,h,b)=\ell_1\}} y_t(a,h,b) \\
&\quad + \sum_{\{\ell'_1 \in \mathcal{L}(t,a,b) \,:\, L(t,a,\ell'_1,b)=\ell_1\}} \left[A_t(a,\ell'_1,b) - \sum_{c \in \mathcal{C}(a)} \sum_{\ell \in \mathcal{L}} \sum_{h \in H(c,\ell)} x_t(c,a,\ell'_1,b,\ell,h) \right].
\end{aligned} \tag{2.8}
$$

(At4) Flow balance equations for the queues: For each $t = 1, \ldots, T$, $c \in \mathcal{C}$, $\ell \in \mathcal{L}$,

$$
\begin{aligned}
C_{t+1}(c,\ell) &= C_t(c,\ell) + \lambda(t,c,\ell) - \sum_{a \in \mathcal{A}(c)} \sum_{b \in \mathcal{B}} \sum_{h \in \mathcal{H}(c,\ell)} x_t(c,a,b,\ell,h) \\
&\quad - \sum_{a \in \mathcal{A}(c)} \sum_{h' \in \mathcal{H}} \sum_{h \in \mathcal{H}(c,\ell)} x_t(c,a,h',\ell,h) - \sum_{a \in \mathcal{A}(c)} \sum_{b \in \mathcal{B}} \sum_{\ell_1 \in \mathcal{L}(t,a,b)} \sum_{h \in H(c,\ell)} x_t(c,a,\ell_1,b,\ell,h).
\end{aligned} \tag{2.9}
$$

(At5) Station capacity constraints: For each $t = 1, \ldots, T+1$, $b \in \mathcal{B}$,

$$
\sum_{a \in \mathcal{A}} A_t(a,b) \leq A_{\max}(b), \tag{2.10}
$$

where $A_{\max}(b)$ denotes the maximum number of ambulances that can park at station b.

(At6) Ambulance supply at the locations between hospitals and stations constraints: For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $b \in \mathcal{B}$, $\ell_1 \in \mathcal{L}(t, a, b)$,

$$\sum_{c \in \mathcal{C}(a)} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c,\ell)} x_t(c, a, \ell_1, b, \ell, h) \quad \leq \quad A_t(a, \ell_1, b). \tag{2.11}$$

(At7) Ambulance supply at stations constraints: For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $b \in \mathcal{B}$,

$$\sum_{c \in \mathcal{C}(a)} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c,\ell)} x_t(c, a, b, \ell, h) \quad \leq \quad A_t(a, b). \tag{2.12}$$

2.1.5. *Constraints for the ambulance reassignment problem.* The following four sets of constraints apply to the first-stage variables for the ambulance reassignment problem only:

(R1) Flow balance equations at the stations: For each $a \in \mathcal{A}$, $b \in \mathcal{B}$,

$$A_1(a, b) \quad = \quad A_0(a, b) + \sum_{\{h \in \mathcal{H} \,:\, L(0,a,h,b) = b\}} y_0(a, h, b) + \sum_{\{\ell_1 \in \mathcal{L}(0,a,b) \,:\, L(0,a,\ell_1,b) = b\}} A_0(a, \ell_1, b). \tag{2.13}$$

(R2) Flow balance equations at the hospitals: For each $a \in \mathcal{A}$, $h \in \mathcal{H}$,

$$\sum_{b \in \mathcal{B}} y_0(a, h, b) + \sum_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{L}} \sum_{h' \in \mathcal{H}(c,\ell)} x_0(c, a, h, \ell, h') \quad = \quad A_0(a, h). \tag{2.14}$$

(R3) Flow balance equations at the locations between hospitals and stations: For each $a \in \mathcal{A}$, $b \in \mathcal{B}$, $\ell_1 \in \mathcal{L}(0, a, b)$,

$$A_1(a, \ell_1, b) \quad = \sum_{\{h \in \mathcal{H} \,:\, L(0,a,h,b) = \ell_1\}} y_0(a, h, b) + \sum_{\{\ell_1' \in \mathcal{L}(0,a,b) \,:\, L(0,a,\ell_1',b) = \ell_1\}} A_0(a, \ell_1', b). \tag{2.15}$$

(R4) Flow balance equations for the queues: For each $c \in \mathcal{C}$, $\ell \in \mathcal{L}$,

$$C_1(c, \ell) \quad = \quad C_0(c, \ell) - \sum_{h \in \mathcal{H}(c,\ell)} \sum_{a \in \mathcal{A}(c)} \sum_{h' \in \mathcal{H}} x_0(c, a, h', \ell, h). \tag{2.16}$$

For $t = 1, \ldots, T$, constraints (At1), (At3), (At4), (At5), (At6), and (At7) apply, and flow balance constraints (At2) at the hospitals are replaced by constraints (At8).

(At8) Flow balance equations at the hospitals: For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $h \in \mathcal{H}$,

$$\begin{aligned}
& \sum_{c \in \mathcal{C}(a)} \sum_{\ell \in \mathcal{L}} \sum_{h' \in \mathcal{H}(c,\ell)} x_t(c, a, h, \ell, h') + \sum_{b \in \mathcal{B}} y_t(a, h, b) \\
= \; & \sum_{c \in \mathcal{C}(a)} \sum_{\ell_1 \in \mathcal{L}} \sum_{\{\ell \in \mathcal{L} \,:\, \tau(0,c,a,\ell_1,\ell,h) = t\}} A_0(c, a, \ell_1, \ell, h) \\
& + \sum_{c \in \mathcal{C}(a)} \sum_{\{\ell_1 \in \mathcal{L} \,:\, \tau_0(c,a,\ell_1,h) = t\}} A_0(c, a, \ell_1, h) \\
& + \sum_{c \in \mathcal{C}(a)} \sum_{\{\ell \in \mathcal{L} \,:\, h \in \mathcal{H}(c,\ell)\}} \sum_{b \in \mathcal{B}} \sum_{\{t' \in \{1,\ldots,t-1\} \,:\, t' + \tau(t',c,a,b,\ell,h) = t\}} x_{t'}(c, a, b, \ell, h) \\
& + \sum_{c \in \mathcal{C}(a)} \sum_{\{\ell \in \mathcal{L} \,:\, h \in \mathcal{H}(c,\ell)\}} \sum_{h' \in \mathcal{H}} \sum_{\{t' \in \{0,\ldots,t-1\} \,:\, t' + \tau(t',c,a,h',\ell,h) = t\}} x_{t'}(c, a, h', \ell, h) \\
& + \sum_{c \in \mathcal{C}(a)} \sum_{\{\ell \in \mathcal{L} \,:\, h \in \mathcal{H}(c,\ell)\}} \sum_{b \in \mathcal{B}} \sum_{\ell_1 \in \mathcal{L}} \sum_{\{t' \in \{1,\ldots,t-1\} \,:\, \ell_1 \in \mathcal{L}(t',a,b),\, t' + \tau(t',c,a,\ell_1,\ell,h) = t\}} x_{t'}(c, a, \ell_1, b, \ell, h).
\end{aligned} \tag{2.17}$$

2.1.6. *Objective function for the ambulance selection problem.* Let $f_t(c, a, b, \ell, h)$ denote the cost per call if ambulance type $a \in \mathcal{A}(c)$ is dispatched at time t from station $b \in \mathcal{B}$ to serve a call of type $c \in \mathcal{C}$ at location $\ell \in \mathcal{L}$, and transport them to hospital $h \in \mathcal{H}(c, \ell)$; this includes a penalty for the waiting time and other costs. Similarly, let $f_t(c, a, h', \ell, h)$ denote the cost per call if ambulance type $a \in \mathcal{A}(c)$ is dispatched at time t from hospital $h' \in \mathcal{H}$ to serve a call of type $c \in \mathcal{C}$ at location $\ell \in \mathcal{L}$, and transport them to hospital $h \in \mathcal{H}(c, \ell)$; let $f_t(c, a, \ell_1, b, \ell, h)$ denote the cost per call if ambulance type $a \in \mathcal{A}(c)$ at location $\ell_1 \in \mathcal{L}(t, a, b)$

at time t traveling toward station $b \in \mathcal{B}$ is dispatched to serve a call of type $c \in \mathcal{C}$ that arrived at location $\ell \in \mathcal{L}$ at time t, and transport them to hospital $h \in \mathcal{H}(c, \ell)$; let $f_t(a, h, b)$ denote the cost if ambulance type $a \in \mathcal{A}$ is dispatched at time t from hospital $h \in \mathcal{H}$ to station $b \in \mathcal{B}$; let $g_t(c, \ell)$ denote the penalty per call of type $c \in \mathcal{C}$ waiting in queue at location $\ell \in \mathcal{L}$ at the beginning of time t; let $g_t(a, b)$ denote the cost per ambulance of type $a \in \mathcal{A}$ at station $b \in \mathcal{B}$ at the beginning of time t; and let $g_t(a, \ell_1, b)$ denote the cost per ambulance of type $a \in \mathcal{A}$ that moves from location $\ell_1 \in \mathcal{L}(t, a, b)$ to location $L(t, a, \ell_1, b)$ during time t.

For the ambulance selection problem, the objective is to minimize

(2.18)

$$
\begin{aligned}
&\sum_{t=0}^{T} \sum_{c \in \mathcal{C}} \sum_{a \in \mathcal{A}(c)} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c,\ell)} \left[\sum_{b \in \mathcal{B}} f_t(c, a, b, \ell, h) x_t(c, a, b, \ell, h) + \sum_{b \in \mathcal{B}} \sum_{\ell_1 \in \mathcal{L}(t,a,b)} f_t(c, a, \ell_1, b, \ell, h) x_t(c, a, \ell_1, b, \ell, h) \right] \\
&+ \sum_{t=1}^{T} \sum_{c \in \mathcal{C}} \sum_{a \in \mathcal{A}(c)} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c,\ell)} \sum_{h' \in \mathcal{H}} f_t(c, a, h', \ell, h) x_t(c, a, h', \ell, h) + \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} \sum_{h \in \mathcal{H}} f_t(a, h, b) y_t(a, h, b) \\
&+ \sum_{t=0}^{T} \left[\sum_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{L}} g_{t+1}(c, \ell) C_{t+1}(c, \ell) + \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} \left\{ g_{t+1}(a, b) A_{t+1}(a, b) + \sum_{\ell_1 \in \mathcal{L}(t+1,a,b)} g_{t+1}(a, \ell_1, b) A_{t+1}(a, \ell_1, b) \right\} \right].
\end{aligned}
$$

2.1.7. *Objective function for the ambulance reassignment problem.* Using the notation of the previous sections, for the ambulance reassignment problem, the objective is to minimize

(2.19)

$$
\begin{aligned}
&\sum_{t=1}^{T} \sum_{c \in \mathcal{C}} \sum_{a \in \mathcal{A}(c)} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c,\ell)} \left[\sum_{b \in \mathcal{B}} f_t(c, a, b, \ell, h) x_t(c, a, b, \ell, h) + \sum_{b \in \mathcal{B}} \sum_{\ell_1 \in \mathcal{L}(t,a,b)} f_t(c, a, \ell_1, b, \ell, h) x_t(c, a, \ell_1, b, \ell, h) \right] \\
&+ \sum_{t=0}^{T} \sum_{c \in \mathcal{C}} \sum_{a \in \mathcal{A}(c)} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c,\ell)} \sum_{h' \in \mathcal{H}} f_t(c, a, h', \ell, h) x_t(c, a, h', \ell, h) + \sum_{t=0}^{T} \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} \sum_{h \in \mathcal{H}} f_t(a, h, b) y_t(a, h, b) \\
&+ \sum_{t=0}^{T} \left[\sum_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{L}} g_{t+1}(c, \ell) C_{t+1}(c, \ell) + \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} \left\{ g_{t+1}(a, b) A_{t+1}(a, b) + \sum_{\ell_1 \in \mathcal{L}(t+1,a,b)} g_{t+1}(a, \ell_1, b) A_{t+1}(a, \ell_1, b) \right\} \right].
\end{aligned}
$$

2.2. **Stochastic Model.** In this section we describe a stochastic model of ambulance dispatch operations. The stochastic model is used to generate scenarios for the two-stage optimization problem, and is used in a simulation to test the performance of different ambulance dispatch policies.

The same space discretization is used for the optimization problem and the simulation, but the time discretization applies to the optimization problem only. The main random variable is the sequence of emergency calls of each type at each location. Other quantities, such as travel times, are modeled as deterministic. Emergency calls arrive in continuous time according to an exogenous stochastic process. We assume that the stochastic process has the property that with probability 1, at most one emergency call arrives at a continuous time point. (In the numerical tests, emergency calls of type c for location ℓ arrive according to a nonhomogeneous Poisson process with rate $\lambda(\tau, c, \ell)$ at time τ.) Let $\omega(\tau)$ denote a random sequence of emergency calls during time interval (τ, T'), where $T' \gg T$ denotes a specified simulation time horizon, and let $\xi(\tau)$ denote a random sequence of emergency calls during time interval $(\tau, \tau + T)$. For any time τ, let $s(\tau)$ denote the state of the process at time τ. That is, $s(\tau)$ contains information about the location and current assignment of each ambulance, and the calls in queue at time τ, including the data of a call, if any, that has just arrived. Let π denote any deterministic policy that takes the state $s(\tau)$ at any time τ as input, and specifies either an ambulance selection decision or an ambulance reassignment decision $\pi(s(\tau))$, as appropriate. The simulation generates a sample path $\omega(0)$ according to the specified stochastic process, and keeps track of the state and performance metrics of the process. When an ambulance selection decision or an ambulance reassignment decision has to be made at time τ, the simulation provides $s(\tau)$ as input to π, receives the decision $\pi(s(\tau))$ as output from π, and updates the state of the process accordingly.

Next we describe how the optimization problems specified in Section 2.1 are used in a corresponding policy π.

First, suppose that π is called at time τ with a request for an ambulance selection decision, and with state $s(\tau)$ provided as input. Then time τ in the simulation corresponds to time $t = 0$ in the current optimization problem, and the time interval $(\tau, \tau + T)$ is partitioned into discrete time periods indexed by $t = 1, \ldots, T$ for the purpose of the optimization problem. Also, N independent and identically distributed "second-stage scenarios", denoted by $\xi_1(\tau), \ldots, \xi_N(\tau)$, each being a sequence of emergency calls during time interval $(\tau, \tau + T)$, is generated according to the specified stochastic process. Note that scenarios $\xi_1(\tau), \ldots, \xi_N(\tau)$ are independent of the sample path $\omega(\tau)$ used to evaluate the policies. Also note that all the input of the optimization problem either does not change, or can be derived from the state $s(\tau)$ provided as input by the simulation and the generated scenarios $\xi_1(\tau), \ldots, \xi_N(\tau)$. For example, the set $\mathcal{C}$ of emergency call types does not change, the type c_0 of the emergency call at $t = 0$ is part of the state $s(\tau)$, and the number $\lambda_n(t, c, \ell)$ of emergency calls of type c for period t and location ℓ under scenario n can be derived from $\xi_n(\tau)$. Let $x_0 := \big(x_0(c_0, a, b, \ell_0, h), x_0(c_0, a, \ell_1, b, \ell_0, h), a \in \mathcal{A}(c_0), b \in \mathcal{B}, \ell_1 \in \mathcal{L}(0, a, b), h \in \mathcal{H}(c_0, \ell_0)\big)$ denote the first-stage decision variables of the ambulance selection problem, and for each scenario n, let

$$\begin{aligned}(x_n, y_n) \quad := \quad & \big(x_{n,t}(c, a, b, \ell, h), x_{n,t}(c, a, \ell_1, b, \ell, h), x_{n,t}(c, a, h', \ell, h), y_{n,t}(a, h', b), C_{n,t}(c, \ell), A_{n,t}(a, b), \\ & A_{n,t}(a, \ell_1, b), t = 1, \ldots, T, c \in \mathcal{C}, a \in \mathcal{A}(c), b \in \mathcal{B}, \ell \in \mathcal{L}, \ell_1 \in \mathcal{L}(0, a, b), h \in \mathcal{H}(c, \ell), h' \in \mathcal{H}\big)\end{aligned}$$

denote the second-stage decision variables for scenario n. Let

$$\begin{aligned}F(x_0) \quad := \quad & \sum_{c \in \mathcal{C}} \sum_{a \in \mathcal{A}(c)} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c, \ell)} \Bigg[\sum_{b \in \mathcal{B}} f_0(c, a, b, \ell, h) x_0(c, a, b, \ell, h) \\ & \qquad + \sum_{b \in \mathcal{B}} \sum_{\ell_1 \in \mathcal{L}(0, a, b)} f_0(c, a, \ell_1, b, \ell, h) x_0(c, a, \ell_1, b, \ell, h) \Bigg]\end{aligned}$$

denote the first-stage part of the ambulance selection problem objective, and for any scenario n, let the remaining (second-stage) part of the ambulance selection problem objective be denoted by $G(s(\tau), \xi_n(\tau), x_n, y_n)$. Then, for each first-stage decision x_0 and each scenario n, the continuous relaxation of the second-stage of the ambulance selection problem is

$$\begin{aligned} Q(s(\tau), \xi_n(\tau), x_0) \quad := \quad & \min_{x_n, y_n} \quad G(s(\tau), \xi_n(\tau), x_n, y_n) \\ & \text{s.t.} \quad (At1), (At2), (At3), (At4), (At5), (At6), (At7). \end{aligned} \tag{2.20}$$

Policy π then chooses an ambulance selection decision

$$\begin{aligned} \pi(s(\tau)) \quad \in \quad & \operatorname*{argmin}_{x_0} \quad F(x_0) + \frac{1}{N} \sum_{n=1}^{N} Q(s(\tau), \xi_n(\tau), x_0) \\ & \text{s.t.} \quad (S1), (S2), (S3), (S4), (S5). \end{aligned}$$

Similarly, suppose that π is called at time τ with a request for an ambulance reassignment decision, and with state $s(\tau)$ provided as input. Let $\tilde{x}_0 := \big(x_0(c, a_0, h_0, \ell, h), y_t(a_0, h_0, b), c \in \mathcal{C}, b \in \mathcal{B}, \ell \in \mathcal{L}, h \in \mathcal{H}(c, \ell)\big)$ denote the first-stage decision variables, and for each scenario n, let

$$\begin{aligned}(x_n, y_n) \quad := \quad & \big(x_{n,t}(c, a, b, \ell, h), x_{n,t}(c, a, \ell_1, b, \ell, h), x_{n,t}(c, a, h', \ell, h), y_{n,t}(a, h', b), C_{n,t}(c, \ell), A_{n,t}(a, b), \\ & A_{n,t}(a, \ell_1, b), t = 1, \ldots, T, c \in \mathcal{C}, a \in \mathcal{A}(c), b \in \mathcal{B}, \ell \in \mathcal{L}, \ell_1 \in \mathcal{L}(0, a, b), h \in \mathcal{H}(c, \ell), h' \in \mathcal{H}\big)\end{aligned}$$

denote the second-stage decision variables for scenario n. Let

$$\tilde{F}(\tilde{x}_0) \quad := \quad \sum_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c, \ell)} f_0(c, a_0, h_0, \ell, h) x_0(c, a_0, h_0, \ell, h) + \sum_{b \in \mathcal{B}} f_0(a_0, h_0, b) y_0(a_0, h_0, b)$$

denote the first-stage part of the ambulance reassignment problem objective, and for any scenario n, let $G(s(\tau), \xi_n(\tau), x_n, y_n)$ denote the remaining (second-stage) part of the ambulance reassignment problem objective. Then, for each first-stage decision $\tilde{x}_0$ and each scenario n, the continuous relaxation of the second-stage of the ambulance reassignment problem is

$$\begin{aligned} \tilde{Q}(s(\tau), \xi_n(\tau), \tilde{x}_0) \quad := \quad & \min_{x_n, y_n} \quad G(s(\tau), \xi_n(\tau), x_n, y_n) \\ & \text{s.t.} \quad (At1), (At3), (At4), (At5), (At6), (At7), (At8). \end{aligned} \tag{2.21}$$

Policy π then chooses an ambulance reassignment decision

$$\begin{aligned}\pi(s(\tau)) \quad \in \quad \underset{\tilde{x}_0}{\operatorname{argmin}} \qquad & \tilde{F}(\tilde{x}_0) + \frac{1}{N}\sum_{n=1}^{N} \tilde{Q}(s(\tau), \xi_n(\tau), \tilde{x}_0) \\ \text{s.t.} \qquad & (R1), (R2), (R3), (R4).\end{aligned}$$

3. Column Generation

We propose a column generation algorithm to solve the optimization problems (2.20) and (2.21). Problems (2.20) and (2.21) are large linear programs that have to be solved fast in practice. In addition, in both problems the number of decision variables is large relative to the number of constraints, so that in a basic feasible solution most decision variables have value zero. This motivates the column generation algorithm proposed in this section to solve these problems. We describe the algorithm for problem (2.20). A similar algorithm is used for problem (2.21).

The dual variables for problem (2.20) are as follows:

- For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $b \in \mathcal{B}$, let $\beta_t(a, b)$ denote the dual variable associated with the flow balance constraint (At1) at station b.
- For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $h \in \mathcal{H}$, let $\alpha_t(a, h)$ denote the dual variable associated with the flow balance constraint (At2) at hospital h.
- For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $b \in \mathcal{B}$, $\ell_1 \in \mathcal{L}(t, a, b)$, let $\psi_t(a, b, \ell_1)$ denote the dual variable associated with the flow balance constraint (At3) at location ℓ_1.
- For each $t = 1, \ldots, T$, $c \in \mathcal{C}$, $\ell \in \mathcal{L}$, let $\phi_t(c, \ell)$ denote the dual variable associated with the flow balance constraint (At4) for the queue of call type c at location ℓ.
- For each $t = 1, \ldots, T + 1$, $b \in \mathcal{B}$, let $\nu_t(b) \geq 0$ denote the dual variable associated with the capacity constraint (At5) for station b.
- For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $b \in \mathcal{B}$, $\ell_1 \in \mathcal{L}(t, a, b)$, let $\theta_t(a, b, \ell_1) \geq 0$ denote the dual variable associated with the ambulance supply constraint (At6) for location ℓ_1.
- For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $b \in \mathcal{B}$, let $\gamma_t(a, b) \geq 0$ denote the dual variable associated with the ambulance supply constraint (At7) for station b.

For problem (2.20), the Lagrangian relaxation is to minimize

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}(c)}\sum_{b\in\mathcal{B}}\sum_{\ell\in\mathcal{L}}\sum_{h\in\mathcal{H}(c,\ell)} \left[f_t(c,a,b,\ell,h) + \beta_t(a,b) - \alpha_{t+\tau(t,c,a,b,\ell,h)}(a,h) + \phi_t(c,\ell) + \gamma_t(a,b)\right] x_t(c,a,b,\ell,h) \\
&+ \sum_{t=1}^{T}\sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}(c)}\sum_{\ell\in\mathcal{L}}\sum_{h\in\mathcal{H}(c,\ell)}\sum_{h'\in\mathcal{H}} \left[f_t(c,a,h',\ell,h) + \alpha_t(a,h') - \alpha_{t+\tau(t,c,a,h',\ell,h)}(a,h) + \phi_t(c,\ell)\right] x_t(c,a,h',\ell,h) \\
&+ \sum_{t=1}^{T}\sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}(c)}\sum_{b\in\mathcal{B}}\sum_{\ell_1\in\mathcal{L}(t,a,b)}\sum_{\ell\in\mathcal{L}}\sum_{h\in\mathcal{H}(c,\ell)} \Big[f_t(c,a,\ell_1,b,\ell,h) + \beta_t(a,b)\mathbb{1}_{\{L(t,a,\ell_1,b)=b\}} - \alpha_{t+\tau(t,c,a,\ell_1,\ell,h)}(a,h) \\
&\qquad\qquad + \psi_t(a,b,L(t,a,\ell_1,b))\mathbb{1}_{\{L(t,a,\ell_1,b)\in\mathcal{L}(t,a,b)\}} + \phi_t(c,\ell) + \theta_t(a,b,\ell_1)\Big] x_t(c,a,\ell_1,b,\ell,h) \\
&+ \sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}\sum_{h\in\mathcal{H}} \left[f_t(a,h,b) + \alpha_t(a,h) - \beta_t(a,b)\mathbb{1}_{\{L(t,a,h,b)=b\}} - \psi_t(a,b,L(t,a,h,b))\mathbb{1}_{\{L(t,a,h,b)\in\mathcal{L}(t,a,b)\}}\right] y_t(a,h,b) \\
&+ \sum_{t=1}^{T}\sum_{c\in\mathcal{C}}\sum_{\ell\in\mathcal{L}} \left[g_t(c,\ell) - \phi_t(c,\ell) + \phi_{t-1}(c,\ell)\right] C_t(c,\ell) + \sum_{c\in\mathcal{C}}\sum_{\ell\in\mathcal{L}} \left[g_{T+1}(c,\ell) + \phi_T(c,\ell)\right] C_{T+1}(c,\ell) \\
&+ \sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \left[g_t(a,b) - \beta_t(a,b) + \beta_{t-1}(a,b) + \nu_t(b) - \gamma_t(a,b)\right] A_t(a,b) \\
&+ \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \left[g_{T+1}(a,b) + \beta_T(a,b) + \nu_{T+1}(b)\right] A_{T+1}(a,b)
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{t=1}^{T} \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} \sum_{\ell_1\in\mathcal{L}(t,a,b)} \big[g_t(a,\ell_1,b) - \beta_t(a,b)\mathbb{1}_{\{L(t,a,\ell_1,b)=b\}} + \psi_{t-1}(a,b,\ell_1) \\
&\qquad\qquad\qquad - \psi_t(a,b,L(t,a,\ell_1,b))\mathbb{1}_{\{L(t,a,\ell_1,b)\in\mathcal{L}(t,a,b)\}} - \theta_t(a,b,\ell_1)\big]\, A_t(a,\ell_1,b) \\
&+ \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} \sum_{\ell_1\in\mathcal{L}(T+1,a,b)} \left[g_{T+1}(a,\ell_1,b) + \psi_T(a,b,\ell_1)\right] A_{T+1}(a,\ell_1,b) \\
&+ \sum_{t=1}^{T} \sum_{a\in\mathcal{A}} \sum_{h\in\mathcal{H}} \alpha_t(a,h) \Bigg\{ - \sum_{c\in\mathcal{C}(a)} \sum_{\ell_1\in\mathcal{L}} \sum_{\{\ell\in\mathcal{L}\,:\,\tau(0,c,a,\ell_1,\ell,h)=t\}} A_0(c,a,\ell_1,\ell,h) \\
&\qquad\qquad\qquad - \sum_{c\in\mathcal{C}(a)} \sum_{\{\ell_1\in\mathcal{L}\,:\,\tau_0(c,a,\ell_1,h)=t\}} A_0(c,a,\ell_1,h) \Bigg\} \\
&- \sum_{t=1}^{T} \sum_{c\in\mathcal{C}} \sum_{\ell\in\mathcal{L}} \phi_t(c,\ell)\lambda(t,c,\ell) - \sum_{t=1}^{T+1} \sum_{b\in\mathcal{B}} \nu_t(b) A_{\max}(b).
\end{aligned}
$$

The column generation algorithm starts with an initial feasible solution for problem (2.20), for instance obtained with the closest-available-ambulance rule to allocate ambulances to calls and the closest-station rule to choose the station to which to send an ambulance when it finishes its task and there are no calls waiting in queue. Then an optimal primal-dual pair is computed for the restriction of problem (2.20) with only the decision variables that are nonzero in the initial feasible solution. If the optimal dual solution of the restricted problem is feasible for the dual of problem (2.20), then the primal-dual pair is optimal for problem (2.20). Otherwise, there exists a primal decision variable with negative reduced cost, and variables with negative reduced cost are added to the restricted problem to obtain the next restricted problem. Every iteration adds a number of primal decision variables with negative reduced cost until the optimal dual solution of the restriction of problem (2.20) is feasible for the dual of problem (2.20). Next we explain how to find primal decision variables with negative reduced cost. We will consider column generation for three types of primal decision variables: $x_t(c,a,\ell_1,b,\ell,h)$, $x_t(c,a,b,\ell,h)$, and $x_t(c,a,h',\ell,h)$.

3.1. **Column Generation for Variables $x_t(c,a,\ell_1,b,\ell,h)$.** The column generation subproblem to determine the variable $x_t(c,a,\ell_1,b,\ell,h)$ with the smallest objective coefficient in the Lagrangian relaxation is

$$
\begin{aligned}
\min \quad & \big\{ f_t(c,a,\ell_1,b,\ell,h) + \beta_t(a,b)\mathbb{1}_{\{L(t,a,\ell_1,b)=b\}} - \alpha_{t+\tau(t,c,a,\ell_1,\ell,h)}(a,h) \\
& \quad + \psi_t(a,b,L(t,a,\ell_1,b))\mathbb{1}_{\{L(t,a,\ell_1,b)\in\mathcal{L}(t,a,b)\}} + \phi_t(c,\ell) + \theta_t(a,b,\ell_1) \;: \\
& \quad t\in\{1,\ldots,T\}, c\in\mathcal{C}, a\in\mathcal{A}(c), b\in\mathcal{B}, \ell_1\in\mathcal{L}(t,a,b), \ell\in\mathcal{L}, h\in\mathcal{H}(c,\ell)\big\}\,,
\end{aligned}
\tag{3.22}
$$

where the dual variable values are optimal dual values for the previous restricted problem. Solving this column generation subproblem exactly can be time consuming, and is unnecessary for most iterations. Next we show how to compute variables $x_t(c,a,\ell_1,b,\ell,h)$ with negative reduced cost under the following simplifying assumptions:

- (A1) The time $\tau(t,c,a,\ell_1,\ell,h)$ for ambulance type a to travel from ℓ_1 at time t to ℓ, provide on-site emergency care for call type c at ℓ, travel with patient(s) from ℓ to hospital h, and deliver the patient(s) at h, does not depend on the call type c. Thus this time is denoted with $\tau(t,a,\ell_1,\ell,h)$.
- (A2) Similarly, the cost $f_t(c,a,\ell_1,b,\ell,h)$ for ambulance type a to travel from ℓ_1 at time t to ℓ, provide on-site emergency care for call type c at ℓ, travel with patient(s) from ℓ to hospital h, and deliver the patient(s) at h, does not depend on the ambulance station b or the call type c. Thus this cost is denoted with $f_t(a,\ell_1,\ell,h)$.
- (A3) The set $\mathcal{A}(c)$ of ambulances to choose from does not depend on the call type c. Thus $\mathcal{A}(c)=\mathcal{A}$ for all c.
- (A4) The set $\mathcal{H}(c,\ell)$ of hospitals to choose from does not depend on the call type c. Thus this set is denoted with $\mathcal{H}(\ell)$.

Then the calculations can be streamlined as follows:

(1) For each time t and emergency location $\ell \in \mathcal{L}$, let

$$c_t^*(\ell) \quad \in \quad \operatorname{argmin}\left\{\phi_t(c,\ell) \ : \ c \in \mathcal{C}\right\}$$

denote the critical call type at time t and location ℓ.

(2) For each time t, ambulance type $a \in \mathcal{A}$, and intermediate location $\ell_1 \in \mathcal{L}$, let

$$b_t^*(a,\ell_1) \quad \in \quad \operatorname{argmin}\left\{\beta_t(a,b)\mathbb{1}_{\{L(t,a,\ell_1,b)=b\}} + \psi_t(a,b,L(t,a,\ell_1,b))\mathbb{1}_{\{L(t,a,\ell_1,b)\in\mathcal{L}(t,a,b)\}} + \theta_t(a,b,\ell_1) \ : \ b \in \mathcal{B}\right\}$$

denote the critical ambulance station at time t for ambulance type a and intermediate location ℓ_1. Observe that if $\{b \in \mathcal{B} \ : \ \ell_1 \in \mathcal{L}(t,a,b)\} = \varnothing$, then there is no decision variable $x_t(c,a,\ell_1,b,\ell,h)$ for such ℓ_1.

(3) For each time t, ambulance type $a \in \mathcal{A}$, emergency location $\ell \in \mathcal{L}$, and intermediate location $\ell_1 \in \mathcal{L}$, let

$$h_t^*(a,\ell_1,\ell) \quad \in \quad \operatorname{argmin}\left\{f_t(a,\ell_1,\ell,h) - \alpha_{t+\tau(t,a,\ell_1,\ell,h)}(a,h) \ : \ h \in \mathcal{H}(\ell)\right\}$$

denote the critical hospital at time t for ambulance type a, emergency location ℓ, and intermediate location ℓ_1. In practice, the set $\mathcal{H}(\ell)$ is small, often a singleton, so the computation of $h_t^*(a,\ell_1,\ell)$ is quick.

Then the column generation subproblem (3.22) reduces to

$$\begin{aligned}
&\min\Big\{f_t(a,\ell_1,\ell,h_t^*(a,\ell_1,\ell)) + \beta_t(a,b_t^*(a,\ell_1))\mathbb{1}_{\{L(t,a,\ell_1,b_t^*(a,\ell_1))=b_t^*(a,\ell_1)\}} - \alpha_{t+\tau(t,a,\ell_1,\ell,h_t^*(a,\ell_1,\ell))}(a,h_t^*(a,\ell_1,\ell)) \\
&\quad +\psi_t(a,b_t^*(a,\ell_1),L(t,a,\ell_1,b_t^*(a,\ell_1)))\mathbb{1}_{\{L(t,a,\ell_1,b_t^*(a,\ell_1))\in\mathcal{L}(t,a,b_t^*(a,\ell_1))\}} + \phi_t(c_t^*(\ell),\ell) + \theta_t(a,b_t^*(a,\ell_1),\ell_1) \ : \\
&\quad t \in \{0,\ldots,T\}, a \in \mathcal{A}, \ell_1 \in \mathcal{L}, \ell \in \mathcal{L}\Big\}\,.
\end{aligned} \tag{3.23}$$

In most iterations, the column generation subproblem does not have to be solved to optimality. Recall that in each column generation iteration, it is sufficient to find a variable $x_t(c,a,\ell_1,b,\ell,h)$ with negative objective value in the column generation subproblem, or to verify that no variable $x_t(c,a,\ell_1,b,\ell,h)$ has negative objective value. Also, for any emergency location ℓ, an (available) ambulance at an intermediate location ℓ_1 that is close to ℓ is more attractive than an ambulance at an intermediate location ℓ_1 that is far from ℓ (this is also the intuition underlying the popular closest-available-ambulance dispatch heuristic). These observations motivate Algorithm 1 below to solve problem (3.23).

3.2. **Column Generation for Variables $x_t(c,a,b,\ell,h)$.** Next, we consider the column generation subproblem to find a variable $x_t(c,a,b,\ell,h)$ with the smallest objective coefficient in the Lagrangian relaxation, that is,

$$\begin{aligned}
\min \quad &\left\{f_t(c,a,b,\ell,h) + \beta_t(a,b) - \alpha_{t+\tau(t,c,a,b,\ell,h)}(a,h) + \phi_t(c,\ell) + \gamma_t(a,b) \ :\right. \\
&\quad \left. t \in \{0,\ldots,T\}, c \in \mathcal{C}, a \in \mathcal{A}(c), b \in \mathcal{B}, \ell \in \mathcal{L}, h \in \mathcal{H}(c,\ell)\right\}.
\end{aligned} \tag{3.24}$$

Consider the following simplifying assumptions:

(B1) The time $\tau(t,c,a,b,\ell,h)$ for ambulance type a to travel from station b at time t to ℓ, provide on-site emergency care for call type c at ℓ, travel with patient(s) from ℓ to hospital h, and deliver the patient(s) at h, does not depend on the call type c. Thus this time is denoted with $\tau(t,a,b,\ell,h)$.

(B2) The cost $f_t(c,a,b,\ell,h)$ for ambulance type a to travel from station b at time t to ℓ, provide on-site emergency care for call type c at ℓ, travel with patient(s) from ℓ to hospital h, and deliver the patient(s) at h, does not depend on the call type c. Thus this cost is denoted with $f_t(a,b,\ell,h)$.

(A3) The set $\mathcal{A}(c)$ of ambulances to choose from does not depend on the call type c. Thus $\mathcal{A}(c) = \mathcal{A}$ for all c.

(A4) The set $\mathcal{H}(c,\ell)$ of hospitals to choose from does not depend on the call type c. Thus this set is denoted with $\mathcal{H}(\ell)$.

Then the calculations can be streamlined as follows:

(1) For each time t and emergency location $\ell \in \mathcal{L}$, let

$$\hat{c}_t(\ell) \quad = \quad c_t^*(\ell) \quad \in \quad \operatorname{argmin}\left\{\phi_t(c,\ell) \ : \ c \in \mathcal{C}\right\}$$

denote the critical call type at time t and location ℓ.

For each emergency location $\ell \in \mathcal{L}$, construct a list $\mathcal{L}(\ell)$ of intermediate locations $\ell_1 \in \mathcal{L}$ sorted from closest to ℓ to furthest from ℓ;
Find an initial feasible solution, and solve the restricted version of the continuous relaxation of problem (2.1)–(2.12) with the decision variables that are nonzero in the initial feasible solution;
optimality_verified $\leftarrow$ false;
while not optimality_verified **do**
  optimality_verified $\leftarrow$ true;
  Use the optimal dual variables for the restricted version of the continuous relaxation of problem (2.1)–(2.12) to compute $c_t^*(\ell)$, $b_t^*(a, \ell_1)$, and $h_t^*(a, \ell_1, \ell)$;
  for $t \in \{0, \ldots, T\}$ **do**
    for $a \in \mathcal{A}$ **do**
      for $\ell \in \mathcal{L}$ **do**
        negative_found $\leftarrow$ false;
        for $\ell_1 \in \mathcal{L}(\ell)$ (from closest to furthest) while not negative_found **do**
          if objective value of column generation subproblem (3.23) is negative **then**
            negative_found $\leftarrow$ true;
            optimality_verified $\leftarrow$ false;
            Add variable $x_t(c_t^*(\ell), a, \ell_1, b_t^*(a, \ell_1), \ell, h_t^*(a, \ell_1, \ell))$ to the restricted version of the continuous relaxation of problem (2.1)–(2.12);
          end if
        end for
      end for
    end for
  end for
  if not optimality_verified **then**
    Solve the current restricted version of the continuous relaxation of problem (2.1)–(2.12);
  end if
end while

Algorithm 1: Column generation algorithm for variables $x_t(c, a, \ell_1, b, \ell, h)$

(2) For each time t, ambulance type $a \in \mathcal{A}$, station $b \in \mathcal{B}$, and emergency location $\ell \in \mathcal{L}$, let

$$\hat{h}_t(a, b, \ell) \quad \in \quad \operatorname{argmin}\left\{f_t(a, b, \ell, h) - \alpha_{t+\tau(t,a,b,\ell,h)}(a, h) \ : \ h \in \mathcal{H}(\ell)\right\}$$

denote the critical hospital at time t for ambulance type a, station b, and emergency location ℓ. As mentioned before, typically the set $\mathcal{H}(\ell)$ is small, and thus the computation of $\hat{h}_t(a, b, \ell)$ is quick.

Then the column generation subproblem (3.24) reduces to

$$\begin{aligned} \min \quad & \Big\{f_t(a, b, \ell, \hat{h}_t(a, b, \ell)) + \beta_t(a, b) - \alpha_{t+\tau(t,a,b,\ell,\hat{h}_t(a,b,\ell))}(a, \hat{h}_t(a, b, \ell)) + \phi_t(\hat{c}_t(\ell), \ell) + \gamma_t(a, b) \ : \\ & \quad t \in \{0, \ldots, T\}, a \in \mathcal{A}, b \in \mathcal{B}, \ell \in \mathcal{L}\Big\}. \end{aligned} \tag{3.25}$$

For any emergency location ℓ, an ambulance at a station b that is close to ℓ is more attractive than an ambulance at a station b that is far from ℓ. These observations motivate Algorithm 2 to solve problem (3.25).

3.3. **Column Generation for Variables $x_t(c, a, h', \ell, h)$.** Next, we consider the column generation subproblem to find a variable $x_t(c, a, h', \ell, h)$ with the smallest objective coefficient in the Lagrangian relaxation, that is,

$$\begin{aligned} \min \quad & \big\{f_t(c, a, h', \ell, h) + \alpha_t(a, h') - \alpha_{t+\tau(t,c,a,h',\ell,h)}(a, h) + \phi_t(c, \ell) \ : \\ & \quad t \in \{1, \ldots, T\}, c \in \mathcal{C}, a \in \mathcal{A}(c), h' \in \mathcal{H}, \ell \in \mathcal{L}, h \in \mathcal{H}(c, \ell)\big\}. \end{aligned} \tag{3.26}$$

Consider the following simplifying assumptions:

(C1) The time $\tau(t, c, a, h', \ell, h)$ for ambulance type a to travel from hospital h' at time t to ℓ, provide on-site emergency care for call type c at ℓ, travel with patient(s) from ℓ to hospital h, and deliver the patient(s) at h, does not depend on the call type c. Thus this time is denoted with $\tau(t, a, h', \ell, h)$.

For each emergency location $\ell \in \mathcal{L}$, construct a list $\mathcal{B}(\ell)$ of bases $b \in \mathcal{B}$ sorted from closest to ℓ to furthest from ℓ;
Find an initial feasible solution, and solve the restricted version of the continuous relaxation of problem (2.1)–(2.12) with the decision variables that are nonzero in the initial feasible solution;
optimality_verified $\leftarrow$ false;
while not optimality_verified **do**
  optimality_verified $\leftarrow$ true;
  Use the optimal dual variables for the restricted version of the continuous relaxation of problem (2.1)–(2.12) to compute $\hat{c}_t(\ell)$ and $\hat{h}_t(a, b, \ell)$;
  for $t \in \{0, \ldots, T\}$ **do**
    for $a \in \mathcal{A}$ **do**
      for $\ell \in \mathcal{L}$ **do**
        negative_found $\leftarrow$ false;
        for $b \in \mathcal{B}(\ell)$ (from closest to furthest) while not negative_found **do**
          if objective value of column generation subproblem (3.25) is negative **then**
            negative_found $\leftarrow$ true;
            optimality_verified $\leftarrow$ false;
            Add variable $x_t(\hat{c}_t(\ell), a, b, \ell, \hat{h}_t(a, b, \ell))$ to the restricted version of the continuous relaxation of problem (2.1)–(2.12);
          end if
        end for
      end for
    end for
  end for
  if not optimality_verified **then**
    Solve the current restricted version of the continuous relaxation of problem (2.1)–(2.12);
  end if
end while

Algorithm 2: Column generation algorithm for variables $x_t(c, a, b, \ell, h)$

(C2) The cost $f_t(c, a, h', \ell, h)$ for ambulance type a to travel from hospital h' at time t to ℓ, provide on-site emergency care for call type c at ℓ, travel with patient(s) from ℓ to hospital h, and deliver the patient(s) at h, does not depend on the call type c. Thus this cost is denoted with $f_t(a, h', \ell, h)$.

(A3) The set $\mathcal{A}(c)$ of ambulances to choose from does not depend on the call type c. Thus $\mathcal{A}(c) = \mathcal{A}$ for all c.

(A4) The set $\mathcal{H}(c, \ell)$ of hospitals to choose from does not depend on the call type c. Thus this set is denoted with $\mathcal{H}(\ell)$.

Then the calculations can be streamlined as follows:

(1) For each time t and emergency location $\ell \in \mathcal{L}$, let

$$\check{c}_t(\ell) \quad = \quad c_t^*(\ell) \quad \in \quad \operatorname{argmin}\{\phi_t(c, \ell) \ : \ c \in \mathcal{C}\}$$

denote the critical call type at time t and location ℓ.

(2) For each time t, ambulance $a \in \mathcal{A}$, hospital $h' \in \mathcal{H}$, and emergency location $\ell \in \mathcal{L}$, let

$$\check{h}_t(a, h', \ell) \quad \in \quad \operatorname{argmin}\left\{f_t(a, h', \ell, h) - \alpha_{t+\tau(t,a,h',\ell,h)}(a, h) \ : \ h \in \mathcal{H}(\ell)\right\}$$

denote the critical hospital at time t for ambulance a, hospital h', and emergency location ℓ.

Then the column generation subproblem (3.26) reduces to

$$\begin{aligned} \min \quad & \Big\{f_t(a, h', \ell, \check{h}_t(a, h', \ell)) + \alpha_t(a, h') - \alpha_{t+\tau(t,a,h',\ell,\check{h}_t(a,h',\ell))}(a, \check{h}_t(a, h', \ell)) + \phi_t(\check{c}_t(\ell), \ell) \ : \\ & \quad t \in \{1, \ldots, T\}, a \in \mathcal{A}, h' \in \mathcal{H}, \ell \in \mathcal{L}\Big\}. \end{aligned} \tag{3.27}$$

For any emergency location ℓ, an ambulance at a hospital h' that is close to ℓ is more attractive than an ambulance at a hospital h' that is far from ℓ. These observations motivate Algorithm 3 to solve problem (3.27).

For each emergency location $\ell \in \mathcal{L}$, construct a list $\mathcal{H}(\ell)$ of hospitals $h' \in \mathcal{H}$ sorted from closest to ℓ to furthest from ℓ;
Find an initial feasible solution, and solve the restricted version of the continuous relaxation of problem (2.1)–(2.12) with the decision variables that are nonzero in the initial feasible solution;
optimality_verified $\leftarrow$ false;
while not optimality_verified **do**
  optimality_verified $\leftarrow$ true;
  Use the optimal dual variables for the restricted version of the continuous relaxation of problem (2.1)–(2.12) to compute $\check{c}_t(\ell)$ and $\check{h}_t(a, h', \ell)$;
  for $t \in \{1, .., T\}$ **do**
    for $a \in \mathcal{A}$ **do**
      for $\ell \in \mathcal{L}$ **do**
        negative_found $\leftarrow$ false;
        for $h' \in \mathcal{H}(\ell)$ (from closest to furthest) while not negative_found **do**
          if objective value of column generation subproblem (3.27) is negative **then**
            negative_found $\leftarrow$ true;
            optimality_verified $\leftarrow$ false;
            Add variable $x_t(\check{c}_t(\ell), a, h', \ell, \check{h}_t(a, h', \ell))$ to the restricted version of the continuous relaxation of problem (2.1)–(2.12);
          end if
        end for
      end for
    end for
  end for
  if not optimality_verified **then**
    Solve the current restricted version of the continuous relaxation of problem (2.1)–(2.12);
  end if
end while

Algorithm 3: Column generation algorithm for variables $x_t(c, a, h', \ell, h)$

We recall that the stopping criterion (controlled by flag optimality_verified) of Algorithms 1, 2, 3, and 4 is that an optimal dual solution of a restricted version of problem (2.20) is feasible for the dual of problem (2.20). In this case, we have indeed found an optimal solution to the primal and dual of the relaxation of the ambulance selection problem.

Gathering our previous developments, we also propose Algorithm 4 below which is a single column generation algorithm for variables $x_t(c, a, b, \ell, h)$, $x_t(c, a, h', \ell, h)$, and $x_t(c, a, \ell_1, b, \ell, h)$ with negative reduced cost for problem (2.20).

We recall that the stopping criterion (controlled by flag optimality_verified) of Algorithms 1, 2, 3, and 4 is that the dual solution of the restricted version of the continuous relaxation of problem (2.1)–(2.12) is feasible for the dual of the continuous relaxation of the original (not restricted using a subset of columns) ambulance selection problem. In this case, we have indeed found an optimal solution to the primal and dual of the relaxation of the ambulance selection problem.

4. Case Study

4.1. **Comparison Between the Proposed Policy and the Closest-Available-Ambulance Heuristic on a real emergency medical service.** The performance of the optimization-based policy for ambulance dispatch decisions described in Section 2 was evaluated using simulation. The problem data were obtained from the Rio de Janeiro emergency medical service (SAMU). This SAMU has 3 types of ambulances (basic, intermediate, and advanced). The call types were grouped into 3 types of calls (also called basic, intermediate, and advanced) according to the type of ambulance needed to serve each type of call. Specifically, basic calls can be served by any type of ambulance, intermediate calls can be served by intermediate and advanced ambulances, and advanced calls can be served by advanced ambulances only. The SAMU has 48 ambulance

For each emergency location $\ell \in \mathcal{L}$, construct a list $\mathcal{B}(\ell)$ of bases $b \in \mathcal{B}$ sorted from closest to ℓ to furthest from ℓ;
For each emergency location $\ell \in \mathcal{L}$, construct a list $\mathcal{H}(\ell)$ of hospitals $h' \in \mathcal{H}$ sorted from closest to ℓ to furthest from ℓ;
For each emergency location $\ell \in \mathcal{L}$, construct a list $\mathcal{L}(\ell)$ of intermediate locations $\ell_1 \in \mathcal{L}$ sorted from closest to ℓ to furthest from ℓ;
Find an initial feasible solution, and solve the restricted version of the continuous relaxation of problem (2.1)–(2.12) with the decision variables that are nonzero in the initial feasible solution;
optimality_verified $\leftarrow$ false;
while not optimality_verified **do**
optimality_verified $\leftarrow$ true;
Use the optimal dual variables for the restricted version of the continuous relaxation of problem (2.1)–(2.12) to compute $\hat{c}_t(\ell) = \check{c}_t(\ell) = c_t^*(\ell)$, $\hat{h}_t(a,b,\ell)$, $\check{h}_t(a,h',\ell)$, $h_t^*(a,\ell_1,\ell)$, and $b_t^*(a,\ell_1)$;
for $t \in \{0,\dots,T\}$ **do**
for $a \in \mathcal{A}$ **do**
for $\ell \in \mathcal{L}$ **do**
negative_found $\leftarrow$ false;
for $b \in \mathcal{B}(\ell)$ (from closest to furthest) while not negative_found **do**
if objective value of column generation subproblem (3.25) is negative **then**
negative_found $\leftarrow$ true;
optimality_verified $\leftarrow$ false;
Add variable $x_t(\hat{c}_t(\ell),a,b,\ell,\hat{h}_t(a,b,\ell))$ to the restricted version of the continuous relaxation of problem (2.1)–(2.12);
end if
end for
if $t \geq 1$ (variables $x_t(c,a,h',\ell,h)$ are defined only for $t \geq 1$ (not for $t = 0$)) **then**
negative_found $\leftarrow$ false;
for $h' \in \mathcal{H}(\ell)$ (from closest to furthest) while not negative_found **do**
if objective value of column generation subproblem (3.27) is negative **then**
negative_found $\leftarrow$ true;
optimality_verified $\leftarrow$ false;
Add variable $x_t(\check{c}_t(\ell),a,h',\ell,\check{h}_t(a,h',\ell))$ to the restricted version of the continuous relaxation of problem (2.1)–(2.12);
end if
end for
end if
negative_found $\leftarrow$ false;
for $\ell_1 \in \mathcal{L}(\ell)$ (from closest to furthest) while not negative_found **do**
if objective value of column generation subproblem (3.23) is negative **then**
negative_found $\leftarrow$ true;
optimality_verified $\leftarrow$ false;
Add variable $x_t(c_t^*(\ell),a,\ell_1,b_t^*(a,\ell_1),\ell,h_t^*(a,\ell_1,\ell))$ to the restricted version of the continuous relaxation of problem (2.1)–(2.12);
end if
end for
end for
end for
end for
if not optimality_verified **then**
Solve the current restricted version of the continuous relaxation of problem (2.1)–(2.12);
end if
end while

Algorithm 4: Column generation algorithm for variables $x_t(c,a,b,\ell,h)$, $x_t(c,a,h',\ell,h)$, and $x_t(c,a,\ell_1,b,\ell,h)$

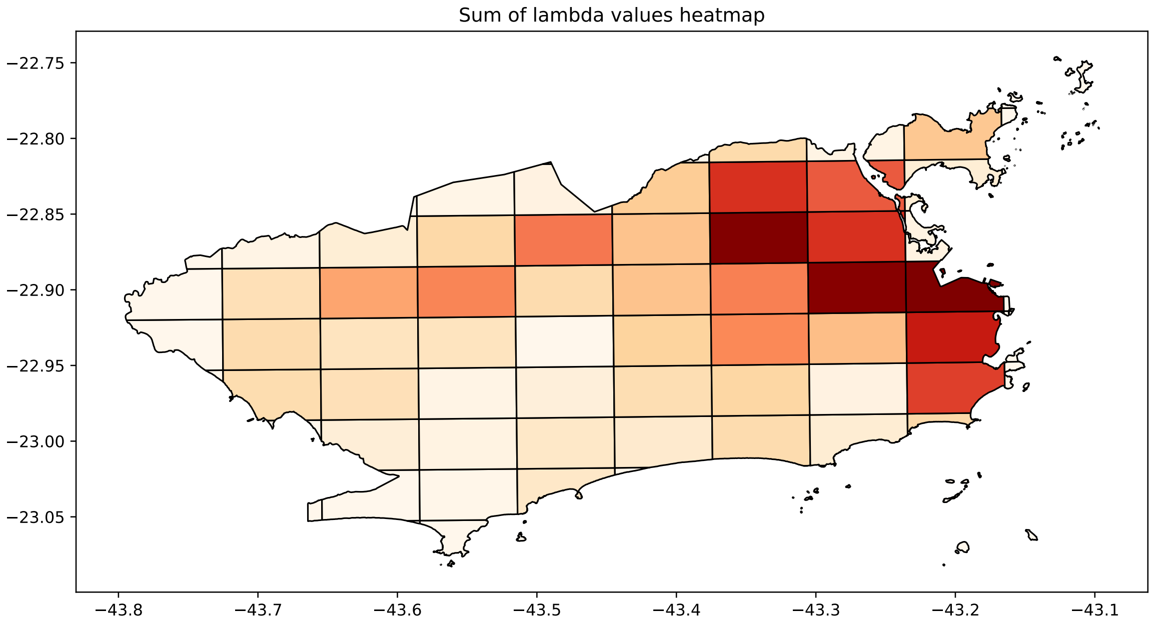

FIGURE 1. Discretization of the service region of the Rio de Janeiro SAMU into 10×10 rectangles, and a heatmap of the mean intensity of emergency calls to the Rio de Janeiro SAMU for the period January 2016–February 2018.

stations and 10 associated hospitals. A rectangle that contains the service region of the Rio de Janeiro SAMU was discretized into 10×10 rectangles. Of these 100 rectangles, 76 have an intersection with the SAMU service region, and are shown in Figure 1. Distances between rectangles were approximated with geodesic distances, and for travel time calculations the ambulance speeds were fixed to 30 km/h. For the optimization problem, time was discretized into 30 minute intervals.

Emergency call data for the SAMU for the period January 2016–February 2018 were used to calibrate the intensities $\lambda(t, c, \ell)$ of calls for each 30 minute period t of each day of the week, for each call type c, and for each discretized location ℓ, of a Poisson process call arrival model via maximizing a regularized likelihood function. Figure 1 also shows a heatmap of the sum of the intensities of the Poisson process (sum over the call types and time periods) for each discretized location ℓ. For the numerical tests, we considered a time window with high intensities of call arrivals (specifically, Fridays between 18h00 and 20h00), and we simulated the call arrivals and dispatch decisions during this time window. The algorithms were implemented using C++14 and run on a computer with a Ryzen 5 2600 processor with 8GB of RAM memory, in a Ubuntu 20.04 OS.

We present two sets of results. First we show that the proposed method results in smaller response times than the response times obtained with the popular closest-available-ambulance rule. Second, we show that if the ambulance selection problem or the ambulance reassignment problem is of large size, then the column generation algorithms result in smaller solution times than a state-of-the-art solver (Gurobi) used to solve the problems without column generation.

In this section, we compare the response times of the method described in Section 2 with the response times of the closest-available-ambulance heuristic, for a time window of 2 hours (Friday 18h00–20h00). The setting was simplified by assuming that the patients of each call must be taken to the closest hospital and that each ambulance must go to the closest station if the call queue is empty. The average response time, denoted by `Mean Heuristic` and `Mean Model`, and the maximum response time, denoted by `Max Heuristic` and `Max Model`, for respectively the closest-available-ambulance heuristic and the policy obtained with our model, are reported in Table 2. The following notation is used in the table headings:

- T: the number of time periods used in the optimization problem;
- $L = |\mathcal{L}|$: the number of discrete locations used in the model;

T	R	B	H	A	S	Mean Heuristic	Mean Model	Max Heuristic	Max Model
4	76	4	4	10	4	2754.7	2213.1	6304.5	5994.5
4	76	4	4	10	6	2686.3	2252.5	6578.4	6494.0
4	76	4	4	10	8	2559.2	2139.4	6275.3	5885.9
4	76	4	4	10	10	2547.1	2111.5	6347.1	6192.1
12	76	8	8	12	4	1985.5	1800.7	5331.5	5294.3
12	76	12	10	12	4	1908.7	1620.2	5059.6	5052.7
12	76	20	10	20	4	1095.2	793.0	4309.0	3206.0
12	76	20	10	8	4	3190.8	2911.7	6454.5	6417.4
12	76	20	10	20	8	1090.3	782.8	3917.6	3553.7

TABLE 2. Comparison between the closest available ambulance heuristic and our optimization model, for different numbers of scenarios, ambulances, hospitals, and bases.

- $B = |\mathcal{B}|$: the number of stations (a subset of the 48 stations of the Rio de Janeiro SAMU);
- $H = |\mathcal{H}|$: the number of hospitals (depending on the instance this is either a subset of the set of 10 public hospitals that the Rio de Janeiro SAMU works with, or all 10 hospitals);
- A: the number of ambulances;
- S: the number of scenarios used in the second-stage problems.

In each setting, both the mean and the maximum response times are smaller for the proposed policy. An increase in the number of hospitals and stations results in a decrease in response times, and an increase in the number of ambulances results in an even greater decrease in response times. An increase in the number of scenarios seems to have relatively little effect on the response times.

4.2. **Runtime Comparisons with and without Column Generation on simulated instances.** We considered several simulated instances of the ambulance selection problem and of the ambulance reassignment problem in the plane, and we solved the continuous relaxations of the second-stage problems either using Gurobi without column generation or using the column generation algorithm from Section 3. For these experiments, we used emergency calls generated according to a homogeneous Poisson process in a square, and considered the following discretizations of the square: (i) into $12 \times 12 = 144$ identical smaller squares, (ii) into $13 \times 13 = 169$ identical smaller squares, (iii) into $14 \times 14 = 196$ identical smaller squares, (iv) and into $15 \times 15 = 225$ identical smaller squares, while distances were calculated using the Euclidean norm.

The CPU times required to solve the ambulance selection problem and the ambulance reassignment problem for these instances are reported in respectively Tables 3 and 4. The following notation is used in the table headings:

- T: the number of time periods used in the optimization problem;
- A: the number of ambulances;
- C: the number of calls;
- B: the number of stations;
- H: the number of hospitals;
- L: the number of discrete locations;
- K: the number of closest hospitals to choose from to deliver the patient.

For larger problem sizes, the column generation algorithm is faster than using Gurobi without column generation.

5. CONCLUSION

The main contribution of this paper is a model of ambulance fleet operations. The model incorporates many important aspects of ambulance fleet operations that are ignored in most of the existing literature, including the following:

(1) The model incorporates different emergency types and the consequences of emergency types, such as ambulance requirements, hospital requirements, and marginal value of response time.

T	A	C	B	H	$\|\mathcal{L}\|$	Time (s) without column generation	Time (s) with column generation	Columns added
24	10	100	4	4	144	26.76	2.62	1029
24	10	100	4	4	144	110.24	12.51	926
24	10	100	4	4	169	21.57	4.49	744
24	10	100	4	4	196	37.07	4.61	914
24	10	100	4	4	225	142.25	4.73	873
24	10	100	4	10	144	78.73	28.41	332
24	10	150	10	10	144	779.35	243.52	232

Table 3. Runtime comparison to solve instances of the ambulance selection problem with and without column generation Algorithm 1 (column generation for variables $x_t(c, a, \ell_1, b, \ell, h)$).

T	A	C	B	H	$\|\mathcal{L}\|$	Time (s) without column generation	Time (s) with column generation	Columns added
24	10	100	4	4	144	1.97	2.71	591
24	10	100	4	4	144	10.75	13.17	542
24	10	100	4	4	169	12.23	14.35	1158
24	10	100	4	4	196	12.40	14.85	1123
24	10	100	4	4	225	13.11	16.03	1054
24	10	100	4	10	144	60.32	45.27	691
24	10	150	10	10	144	725.95	271.69	2650

Table 4. Runtime comparison to solve instances of the request selection problem with and without column generation Algorithm 1 (column generation for variables $x_t(c, a, \ell_1, b, \ell, h)$).

(2) The model facilitates different ambulance and crew types.
(3) The model allows hospital choice, taking into account the emergency type and the location of the emergency relative to hospitals.
(4) The model makes provision for a queue of waiting emergencies.
(5) The model incorporates both ambulance selection decisions as well as ambulance reassignment decisions.
(6) The model allows ambulances on their way to a station to be dispatched to an emergency.
(7) The model takes into account the future consequences of dispatch decisions.

We proposed a policy based on a rolling horizon approach combined with a two-stage stochastic problem solved each time an ambulance dispatch decision is required. An algorithm based on column generation was proposed for solving the two-stage stochastic problem. The model and algorithm were tested using a simulation calibrated with data of the Rio de Janeiro emergency medical service. The proposed policy performs much better than the popular closest-available-ambulance heuristic. Additional work is needed to reduce the time taken to compute the decisions for the proposed policy.

References

[1] R. Alanis, A. Ingolfsson, and B. Kolfal. A Markov chain model for an EMS system with repositioning. *Production and Operations Management*, 22(1):216–231, 2013.

[2] T. Andersson and P. Värbrand. Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society*, 58(2):195–201, 2007.

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

[4] D. Bandara, M. E. Mayorga, and L. A. McLay. Optimal dispatching strategies for emergency vehicles to increase patient survivability. *International Journal of Operational Research*, 15(2):195–214, 2012.

[5] D. Bandara, M. E. Mayorga, and L. A. McLay. Priority dispatching strategies for EMS systems. *Journal of the Operational Research Society*, 65:572–587, 2014.

[6] V. Bélanger, Y. Kergosien, A. Ruiz, and P. Soriano. An empirical comparison of relocation strategies in real-time ambulance fleet management. *Computers and Industrial Engineering*, 94:216–229, 2016.

[7] G. N. Berlin and J. C. Liebman. Mathematical analysis of emergency ambulance location. *Socio-Economic Planning Sciences*, 8(6):323–328, 1974.

[8] T. H. Blackwell and J. S. Kaufman. Response time effectiveness: Comparison of response time and survival in an urban emergency medical service system. *Academic Emergency Medicine*, 9(4):288–295, 1991.

[9] L. Brotcorne, G. Laporte, and F. Semet. Ambulance location and relocation models. *European Journal of Operational Research*, 147(3):451–463, 2003.

[10] T. H. Burwell, J. P. Jarvis, and M. A. McKnew. Modeling co-located servers and dispatch ties in the hypercube model. *Computers and Operations Research*, 20(2):113–119, 1993.

[11] G. M. Carter, J. M. Chaiken, and E. Ignall. Response areas for two emergency units. *Operations Research*, 20(3):571–594, 1972.

[12] J. M. Chaiken and R. C. Larson. Methods for allocating urban emergency units: A survey. *Management Science*, 19(4):P110–P130, 1972.

[13] S. H. Cho, H. Jang, T. Lee, and J. Turner. Simultaneous location of trauma centers and helicopters for emergency medical service planning. *Operations Research*, 62(4):751–771, 2014.

[14] R. L. Church and C. S. ReVelle. The maximal covering location problem. *Papers of the Regional Science Association*, 32:101–118, 1974.

[15] S. Cretin and T. R. Willemain. A model of preshospital death from ventricular fibrillation following myocardial infarction. *Health Services Research*, 14(3):221–234, 1979.

[16] M. S. Daskin. A maximum expected covering location model: Formulation, properties and heuristic solution. *Transportation Science*, 7(1):48–70, 1983.

[17] M. S. Daskin and E. H. Stern. A hierarchical objective set covering model for Emergency Medical Service vehicle deployment. *Transportation Science*, 15(2):137–152, 1981.

[18] D. Degel, L. Wiesche, S. Rachuba, and B. Werners. Time-dependent ambulance allocation considering data-driven empirically required coverage. *Health Care Management Science*, 18:444–458, 2015.

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

[20] E. Erkut, A. Ingolfsson, T. Sim, and G. Erdoğan. Computational comparison of five maximal covering models for locating ambulances. *Geographical Analysis*, 41:43–65, 2009.

[21] J. A. Fitzsimmons. A methodology for emergency ambulance deployment. *Management Science*, 19(6):627–636, 1973.

[22] M. Gendreau, G. Laporte, and F. Semet. Solving an ambulance location model by Tabu Search. *Location Science*, 5(2):75–58, 1997.

[23] M. Gendreau, G. Laporte, and F. Semet. A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*, 27(12):1641–1653, 2001.

[24] M. Gendreau, G. Laporte, and F. Semet. The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operational Research Society*, 57(1):22–28, 2006.

[25] J. Goldberg, R. Dietrich, J. M. Chen, M. G. Mitwasi, T. Valenzuela, and E. Criss. A simulation model for evaluating a set of emergency vehicle base locations: Development, validation, and usage. *Socio-Economic Planning Sciences*, 24(2):125–141, 1990.

[26] J. Goldberg, R. Dietrich, J. M. Chen, M. G. Mitwasi, T. Valenzuela, and E. Criss. Validating and applying a model for locating emergency medical vehicles in Tucson, AZ. *European Journal of Operational Research*, 49(3):308–324, 1990.

[27] J. Goldberg and L. Paz. Locating emergency vehicle bases when service time depends on call location. *Transportation Science*, 25(4):264–280, 1991.

[28] J. Goldberg and F. Szidarovszky. Methods for solving nonlinear equations used in evaluating emergency vehicle busy probabilities. *Operations Research*, 39(6):903–916, 1991.

[29] L. V. Green and P. J. Kolesar. Improving emergency responsiveness with management science. *Management Science*, 50(8):1001–1014, 2004.

[30] V. Guigues and C. Sagastizabal. The value of rolling-horizon policies for risk-averse hydro-thermal planning. *European Journal of Operational Research*, 217:129–140, 2012.

[31] V. Guigues and C. Sagastizabal. Risk-averse feasible policies for large-scale multistage stochastic linear programs. *Mathematical Programming*, 138:167–198, 2013.

[32] A. Haghani, Q. Tian, and H. Hu. Simulation model for real-time emergency vehicle dispatching and routing. *Transportation Research Record*, 1882:176–183, 2004.

[33] S. G. Henderson and A. J. Mason. Estimating ambulance requirements in Auckland, New Zealand. In *Proceedings of the 1999 Winter Simulation Conference*, volume 2, pages 1670–1674, 1999.

[34] S. G. Henderson and A. J. Mason. Ambulance service planning: Simulation and data visualisation. In M. Brandeau, F. Sainfort, and W. Pierskalla, editors, *Operations Research and Health Care: A Handbook of Methods and Applications, International Series in Operations Research and Management Science 70*, chapter 4, pages 77–102. Kluwer, Dordecht, 2004.

[35] E. D. Hill, J. L. Hill, and L. M. Jacobs. Planning for emergency ambulance service systems. *The Journal of Emergency Medicine*, 1:331–338, 1984.

[36] K. Hogan and C. S. Revelle. Concepts and applications of backup coverage. *Management Science*, 32(11):1434–1444, 1986.

[37] A. Ingolfsson, S. Budge, and E. Erkut. Optimal ambulance location with random delays and travel times. *Health Care Management Science*, 11:262–274, 2008.

[38] C. J. Jagtenberg, S. Bhulai, and R. D. Van der Mei. An efficient heuristic for real-time ambulance redeployment. *Operations Research for Health Care*, 12(4):27–35, 2015.

[39] C. J. Jagtenberg, S. Bhulai, and R. D. Van der Mei. Dynamic ambulance dispatching: Is the closest-idle policy always optimal? *Operations Research for Health Care*, 20(4):517–531, 2017.

[40] C. J. Jagtenberg, P. L. Van den Berg, and R. D. Van der Mei. Benchmarking online dispatch algorithms for Emergency Medical Services. *European Journal of Operational Research*, 258(2):715–725, 2017.

[41] J. P. Jarvis. Approximating the equilibrium behavior of multi-server loss systems. *Management Science*, 31(2):235–239, 1985.

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

[43] R. C. Larson. A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers and Operations Research*, 1:67–95, 1974.

[44] R. C. Larson. Approximating the performance of urban emergency service systems. *Operations Research*, 23(5):845–868, 1975.

[45] R. C. Larson and K. A. Stevenson. On insensitivities in urban redistricting and facility location. *Operations Research*, 20(3):595–612, 1972.

[46] S. Lee. The role of preparedness in ambulance dispatching. *Journal of the Operational Research Society*, 62(10):1888–1897, 2011.

[47] S. Lee. The role of centrality in ambulance dispatching. *Decision Support Systems*, 54(1):282–291, 2012.

[48] X. Li and C. Saydam. Balancing ambulance crew workloads via a tiered dispatch policy. *Pesquisa Operacional*, 36(3):399–419, 2016.

[49] X. Li, Z. Zhao, X. Zhu, and T. Wyatt. Covering models and optimization techniques for emergency response facility location and planning: A review. *Transportation Science*, 74(3):281–310, 2011.

[50] C. S. Lim, R. Mamat, and T. Bräunl. Impact of ambulance dispatch policies on performance of Emergency Medical Services. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):624–632, 2011.

[51] A. J. Mason. Simulation and real-time optimised relocation for improving ambulance operations. In B. T. Denton, editor, *Handbook of Healthcare Operations Management: Methods and Applications, International Series in Operations Research and Management Science 184*, chapter 11, pages 289–317. Springer, New York, 2013.

[52] M. S. Maxwell, S. G. Henderson, and H. Topaloglu. Ambulance redeployment: An approximate dynamic programming approach. In M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, editors, *Proceedings of the 2009 Winter Simulation Conference*, pages 1850–1860, 2009.

[53] M. S. Maxwell, S. G. Henderson, and H. Topaloglu. Tuning approximate dynamic programming policies for ambulance redeployment via direct search. *Stochastic Systems*, 3(2):322–361, 2013.

[54] M. S. Maxwell, E. C. Ni, C. Tong, S. G. Henderson, H. Topaloglu, and S. R. Hunter. A bound on the performance of an optimal ambulance redeployment policy. *Operations Research*, 62(5):1014–1027, 2014.

[55] M. S. Maxwell, M. Restrepo, S. G. Henderson, and H. Topaloglu. Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing*, 22(2):266–281, 2010.

[56] M. E. Mayorga, D. Bandara, and L. A. McLay. Districting and dispatching policies for emergency medical service systems to improve patient survival. *IIE Transactions on Healthcare Systems Engineering*, 3(1):39–56, 2013.

[57] R. Nair and E. Miller-Hooks. Evaluation of relocation strategies for emergency medical service vehicles. *Transportation Research Record*, 2137:63–73, 2009.

[58] J. Nicholl, P. Coleman, G. Parry, J. Turner, and S. Dixon. Emergency priority dispatch systems—a new era in the provision of ambulance services in the UK. *Pre-hospital Immediate Care*, 3:71–75, 1999.

[59] D. E. Persse, C. B. Key, R. N. Bradley, C. C. Miller, and A. Dhingra. Cardiac arrest survival as a function of ambulance deployment strategy in a large urban emergency medical services system. *Resuscitation*, 59:97–104, 2003.

[60] J. F. Repede and J. J. Bernardo. Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *European Journal of Operational Research*, 75:567–581, 1994.

[61] M. Restrepo, S. G. Henderson, and H. Topaloglu. Erlang loss models for the static deployment of ambulances. *Health Care Management Science*, 12(67):67–79, 2009.

[62] C. ReVelle, D. Bigman, D. Schilling, J. Cohon, and R. Church. Facility location: A review of context-free and EMS models. *Health Services Research*, 12(2):129–146, 1977.

[63] C. ReVelle and K. Hogan. The maximum availability location problem. *Transportation Science*, 23(3):192–200, 1989.

[64] D. A. Schilling, D. J. Elzinga, J. Cohon, R. L. Church, and C. S. ReVelle. The TEAM/FLEET models for simultaneous facility and equipment siting. *Transportation Science*, 13(2):163–175, 1979.

[65] V. Schmid. Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research*, 219:611–621, 2012.

[66] V. Schmid and K. F. Doerner. Ambulance location and relocation problems with time-dependent travel times. *European Journal of Operational Research*, 207:1293–1303, 2010.

[67] P. Sorensen and R. Church. Integrating expected coverage and local reliability for emergency medical services location problems. *Socio-Economic Planning Sciences*, 44(1):8–18, 2010.

[68] I. G. Stiell, L. P. Nesbitt, W. Pickett, D. Munkley, D. W. Spaite, J. Banek, B. Field, L. Luinstra-Toohey, J. Maloney, J. Dreyer, M. Lyver, T. Campeau, and G. A. Wells. The OPALS major trauma study: Impact of advanced life-support on survival and morbidity. *Canadian Medical Association Journal*, 178(9):1141–1152, 2008.

[69] C. Swoveland, D. Uyeno, I. Vertinsky, and R. Vickson. Ambulance location: A probabilistic enumeration approach. *Management Science*, 20(4):686–698, 1973.

[70] C. Swoveland, D. Uyeno, I. Vertinsky, and R. Vickson. A simulation-based methodology for optimization of ambulance service policies. *Socio-Economic Planning Sciences*, 7(6):697–703, 1973.

[71] C. Toregas, R. Swain, C. ReVelle, and L. Bergman. The location of emergency service facilities. *Operations Research*, 19(6):1363–1373, 1971.

[72] T. C. Van Barneveld, S. Bhulai, and R. D. Van der Mei. The effect of ambulance relocations on the performance of ambulance service providers. *European Journal of Operational Research*, 252:257–269, 2016.

[73] T. C. Van Barneveld, S. Bhulai, and R. D. Van der Mei. A dynamic ambulance management model for rural areas: Computing redeployment actions for relevant performance measures. *Health Care Management Science*, 20:165–186, 2017.

[74] R. A. Volz. Optimum ambulance location in semi-rural areas. *Transportation Science*, 5(2):193–203, 1971.