LOCATION AND OPERATION OF AN AMBULANCE FLEET UNDER UNCERTAINTY

1. Introduction

Operations research models and algorithms have been used since several years for the management of fleets of ambulances. More precisely, the ambulance strategic planning problem consists in deciding how to allocate ambulances to bases for a given period or how to choose base locations while the ambulance operational planning problem consists in operating in real time a fleet of ambulances to control the trajectories of ambulances, choosing the ambulances to send to the emergency calls and their destinations when they finish service. For this application, most operations research tools use static optimization models for the strategic planning problem and simulation tools for the operational planning problem, see [1, 3]. The operational planning problem can involve the scheduled times when ambulance crews go on duty and off duty, [3]. In Brazil, a simulation model for Belo Horizonte using Arena was proposed in [2]. Some optimization models based on Approximate Dynamic Programming have however been proposed, [4, 5, 6, 7]. In particular, [7], develops an approximate dynamic programming scheme based on state space aggregation and reinforcement learning.

This paper proposes an optimization-based method for controlling ambulance dispatches. Before we describe the optimization model, we briefly review basic ambulance emergency response operations. We consider requests for emergency medical service that arrive at a call center, for example, a 911-call center in the USA. (Thus we do not consider potential patients who arrive with their own transportation at hospital emergency departments.) A call center operator receives the call and obtains some 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 estimate whether or not a call is related to an emergency that has already been reported. We do not address that question in this paper.) The call center operator decides whether an ambulance should be dispatched to the emergency. If so, a decision is made which ambulance should be dispatched to the emergency, or whether the request should enter 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.) In some systems the call center operator 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 arrives at the location of the emergency is called the response time. It has been found that the response time is highly correlated with the survival probability of the patient, and therefore many emergency medical services and academic papers put great emphasis on the response time performance metric . However, it has been pointed out that the response time is not the only variable under the control of emergency medical services that affects the survival probability of the patient. For example, although many academic papers consider all ambulances to be the same, typically ambulances are not the same.

The outline of the paper is the following. In Section 2, we detail our optimization models. In Section 3, we explain how those problems are solved. Numerical results for the Rio de Janeiro SAMU are presented in Section 4 while the final section 5 provides recommendations and future directions of research.

2. Optimization models

We describe an operational model to be solved

Date: January 10, 2020.

(1) after each emergency call arrives, to determine either which ambulance(s) to dispatch for that call, or whether to put the call in a queue of calls waiting for ambulance(s) to be dispatched, as well as
(2) after each dispatched ambulance finishes its task, for example, by treating a patient on-site or by delivering a patient to a hospital, to determine either for which call in queue to dispatch the ambulance next, or towards which ambulance base to send the ambulance.

The formulation "looks ahead" until the end of the time horizon (for example, until the end of the day), using forecasts of emergency calls in the region under consideration, that is, it approximates the impact of current decisions on the objective function in the future.

Ambulances can be dispatched from ambulance bases, and from hospitals, and from other locations while traveling towards an ambulance base (but not while in service — traveling towards an emergency location, or providing on-site emergency care, or traveling with patient(s) towards a hospital — that is, we do not model "forward" dispatching of ambulances). An ambulance that is not in service must either be at an ambulance base, or traveling toward an ambulance base. If ambulances are allowed to wait at a hospital for a dispatch, then the hospital is regarded as an ambulance base in the model.

Both time and space are discretized for the model. Let 0 denote the current time and let $t = 1, \ldots, T$ denote the time steps until the end 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 time, its location, and its type (which determines its priority). Let $\mathcal{C}$ denote the set of call types, let $\mathcal{A}$ denote the set of ambulances types, let $\mathcal{B}$ denote the set of ambulance bases, 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 ambulances, 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.

Other input includes the initial conditions for the problem. If a call arrives at time 0, then we denote by ℓ_0 and c_0 respectively the location and the type of the call that has just arrived. If an ambulance completes service at a hospital at time 0, then we denote by a_0 and h_0 respectively the ambulance type and the hospital. 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. For each ambulance type $a \in \mathcal{A}$ and ambulance base $b \in \mathcal{B}$, 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.

As mentioned above, ambulances can be dispatched while traveling towards an ambulance base. To facilitate 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 ambulance base $b \in \mathcal{B}$, let $L(t, a, h, b) \in \mathcal{L}$ denote the forecasted location of the ambulance at time $t+1$ if the ambulance travels towards b. In addition, for each current ambulance location $\ell_1 \in \mathcal{L}$ at time t, and destination ambulance base $b \in \mathcal{B}$, 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 $\mathcal{L}_1(t, a, b) := \cup_{h \in \mathcal{H}} \{L(t-1, a, h, b)\} \setminus \{b\}$ 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 $\mathcal{L}_\tau(t, a, b) := \cup_{\ell_1 \in \mathcal{L}_{\tau-1}(t-1, a, b)} \{L(t-1, a, \ell_1, b)\} \setminus \{b\}$ 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 ambulance base $b \in \mathcal{B}$, let $\mathcal{L}(t, a, b) := \cup_{\tau \geq 1} \mathcal{L}_\tau(t, a, b)$ 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 $a \in \mathcal{A}$, ambulance base $b \in \mathcal{B}$, 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 $h \in \cup_{\ell \in \mathcal{L}} \mathcal{H}(c, \ell)$.

Additional input includes demand forecasts, travel time and service time forecasts, and path forecasts. Specifically, 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.

If a call arrives at time 0, then the decision variables for the current emergency call are:

- $x_0(a, b, h)$ = the number of ambulances of type $a \in \mathcal{A}(c_0)$ dispatched at time 0 from base $b \in \mathcal{B}$ to serve patient(s) of the current call at location ℓ_0 and transport them to hospital $h \in \mathcal{H}(c_0, \ell_0)$;
- $x_0(a, h', h)$ = the number of ambulances of type $a \in \mathcal{A}(c_0)$ dispatched at time 0 from hospital $h' \in \mathcal{H}$ to serve patient(s) of the current call at location ℓ_0 and transport them to hospital $h \in \mathcal{H}(c_0, \ell_0)$;
- $x_0(a, \ell_1, b, h)$ = the number of ambulances of type $a \in \mathcal{A}(c_0)$ at location $\ell_1 \in \mathcal{L}(0, a, b)$ at time 0 traveling toward base $b \in \mathcal{B}$ dispatched to serve patient(s) of the current call at location ℓ_0 and transport them to hospital $h \in \mathcal{H}(c_0, \ell_0)$.
- $y_0(a, h, b)$ = the number of ambulances of type $a \in \mathcal{A}$ instructed at time 0 to move from hospital $h \in \mathcal{H}$ towards base $b \in \mathcal{B}$;

The decision variables above are binary, or could be larger integers if more than one ambulance is needed to serve the current emergency.

If an ambulance completes service at a hospital at time 0, then the decision variables for the ambulance are:

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

For time steps $t = 1, \ldots, T$, the decision variables are

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

All decision variables are nonnegative.

If a call arrives at time 0, then we have the following four sets of constraints for time 0:

Flow balance equations at the bases: For each $a \in \mathcal{A}$, $b \in \mathcal{B}$,

$$\begin{aligned} A_1(a,b) &= A_0(a,b) + \sum_{\{h\in\mathcal{H}\,:\,L(0,a,h,b)=b\}} y_0(a,h,b) \\ &\quad + \sum_{\{\ell_1\in\mathcal{L}(0,a,b)\,:\,L(0,a,\ell_1,b)=b\}} \left[A_0(a,\ell_1,b) - \sum_{h\in H(c_0,\ell_0)} x_0(a,\ell_1,b,h) \right] \\ &\quad - \sum_{h\in\mathcal{H}(c_0,\ell_0)} x_0(a,b,h). \end{aligned} \tag{2.1}$$

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

$$\sum_{h'\in\mathcal{H}(c_0,\ell_0)} x_0(a,h,h') + \sum_{b\in\mathcal{B}} y_0(a,h,b) = A_0(a,h). \tag{2.2}$$

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

(2.3)

$$A_1(a,\ell_1,b) = \sum_{\{h\in\mathcal{H}\,:\,L(0,a,h,b)=\ell_1\}} y_0(a,h,b) + \sum_{\{\ell_1'\in\mathcal{L}(0,a,b)\,:\,L(0,a,\ell_1',b)=\ell_1\}} \left[A_0(a,\ell_1',b) - \sum_{h\in H(c_0,\ell_0)} x_0(a,\ell_1',b,h) \right]$$

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

$$\begin{aligned} C_1(c,\ell) &= C_0(c,\ell) + \mathbb{I}_{\{c=c_0,\ell=\ell_0\}} \left[1 - \sum_{a\in\mathcal{A}(c)} \sum_{b\in\mathcal{B}} \sum_{h\in\mathcal{H}(c,\ell)} x_0(a,b,h) \right. \\ &\quad \left. - \sum_{a\in\mathcal{A}(c)} \sum_{h'\in\mathcal{H}} \sum_{h\in\mathcal{H}(c,\ell)} x_0(a,h',h) - \sum_{a\in\mathcal{A}(c)} \sum_{b\in\mathcal{B}} \sum_{\ell_1\in\mathcal{L}(0,a,b)} \sum_{h\in H(c,\ell)} x_0(a,\ell_1,b,h) \right] \end{aligned} \tag{2.4}$$

If an ambulance completes service at a hospital at time 0, then we have the following four sets of constraints for time 0:

Flow balance equations at the bases: For each $a \in \mathcal{A}$, $b \in \mathcal{B}$,

$$\begin{aligned} A_1(a,b) &= A_0(a,b) + \sum_{\{h\in\mathcal{H}\,:\,L(0,a,h,b)=b\}} y_0(a,h,b) \\ &\quad + \sum_{\{\ell_1\in\mathcal{L}(0,a,b)\,:\,L(0,a,\ell_1,b)=b\}} A_0(a,\ell_1,b) \end{aligned} \tag{2.5}$$

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

$$\sum_{b\in\mathcal{B}} y_0(a,h,b) + \sum_{c\in\mathcal{C}} \sum_{\ell\in\mathcal{L}} \sum_{h'\in\mathcal{H}(c,\ell)} y_0(a,h,c,\ell,h') = A_0(a,h). \tag{2.6}$$

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

$$\begin{aligned} A_1(a,\ell_1,b) &= \sum_{\{h\in\mathcal{H}\,:\,L(0,a,h,b)=\ell_1\}} y_0(a,h,b) \\ &\quad + \sum_{\{\ell_1'\in\mathcal{L}(0,a,b)\,:\,L(0,a,\ell_1',b)=\ell_1\}} A_0(a,\ell_1',b) \end{aligned} \tag{2.7}$$

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

$$C_1(c,\ell) = C_0(c,\ell) - \sum_{h\in\mathcal{H}(c,\ell)} \sum_{a\in\mathcal{A}(c)} \sum_{h'\in\mathcal{H}} y_0(a,h',c,\ell,h) \tag{2.8}$$

The following constraints apply for times $t = 1,\ldots,T$:

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

$$
\begin{aligned}
A_{t+1}(a,b) \;=\; & A_t(a,b) + \sum_{\{h' \in \mathcal{H} \,:\, L(t,a,h',b)=b\}} y_t(a,h',b) \\
& + \sum_{\{\ell_1 \in \mathcal{L}(t,a,b) \,:\, L(t,a,\ell_1,b)=b\}} \left[A_t(a,\ell_1,b) - \sum_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{L}} \sum_{h \in H(c,\ell)} x_t(c,a,\ell_1,b,\ell,h) \right] \\
& - \sum_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c,\ell)} x_t(c,a,b,\ell,h).
\end{aligned} \tag{2.9}
$$

Flow balance equations at the hospitals: If a call arrives at time 0, then we have the following flow balance constraints at the hospitals: For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $h \in \mathcal{H}$,

$$
\begin{aligned}
& \sum_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{L}} \sum_{h' \in \mathcal{H}(c,\ell)} x_t(c,a,h,\ell,h') + \sum_{b \in \mathcal{B}} y_t(a,h,b) \\
= \; & \sum_{c \in \mathcal{C}} \sum_{\ell_1 \in \mathcal{L}} \sum_{\{\ell \in \mathcal{L} \,:\, \tau(0,c,a,\ell_1,\ell,h)=t\}} A_0(c,a,\ell_1,\ell,h) \\
& + \sum_{c \in \mathcal{C}} \sum_{\{\ell_1 \in \mathcal{L} \,:\, \tau_0(c,a,\ell_1,h)=t\}} A_0(c,a,\ell_1,h) \\
& + \sum_{\{b \in \mathcal{B} \,:\, \tau(0,c_0,a,b,\ell_0,h)=t\}} x_0(a,b,h) \\
& + \sum_{c \in \mathcal{C}} \sum_{\{\ell \in \mathcal{L} \,:\, h \in \mathcal{H}(c,\ell)\}} \sum_{b \in \mathcal{B}} \sum_{\{t' \in \{1,\ldots,t-1\} \,:\, t'+\tau(t',c,a,b,\ell,h)=t\}} x_{t'}(c,a,b,\ell,h) \\
& + \sum_{\{h' \in \mathcal{H} \,:\, \tau(0,c_0,a,h',\ell_0,h)=t\}} x_0(a,h',h) \\
& + \sum_{c \in \mathcal{C}} \sum_{\{\ell \in \mathcal{L} \,:\, h \in \mathcal{H}(c,\ell)\}} \sum_{h' \in \mathcal{H}} \sum_{\{t' \in \{1,\ldots,t-1\} \,:\, t'+\tau(t',c,a,h',\ell,h)=t\}} x_{t'}(c,a,h',\ell,h) \\
& + \sum_{\{\ell_1 \in \mathcal{L} \,:\, \tau(0,c_0,a,\ell_1,\ell_0,h)=t\}} \sum_{b \in \mathcal{B}} x_0(a,\ell_1,b,h) \\
& + \sum_{c \in \mathcal{C}} \sum_{\{\ell \in \mathcal{L} \,:\, h \in \mathcal{H}(c,\ell)\}} \sum_{b \in \mathcal{B}} \sum_{\ell_1 \in \mathcal{L}} \sum_{\{t' \in \{1,\ldots,t-1\} \,:\, \ell_1 \in \mathcal{L}(t',a,b),\, t'+\tau(t',c,a,\ell_1,\ell,h)=t\}} x_{t'}(c,a,\ell_1,b,\ell,h)
\end{aligned} \tag{2.10}
$$

If an ambulance completes service at a hospital at time 0, then we have the following flow balance constraints at the hospitals: For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $h \in \mathcal{H}$,

$$
\begin{aligned}
& \sum_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{L}} \sum_{h' \in \mathcal{H}(c,\ell)} x_t(c,a,h,\ell,h') + \sum_{b \in \mathcal{B}} y_t(a,h,b) \\
= \; & \sum_{c \in \mathcal{C}} \sum_{\ell_1 \in \mathcal{L}} \sum_{\{\ell \in \mathcal{L} \,:\, \tau(0,c,a,\ell_1,\ell,h)=t\}} A_0(c,a,\ell_1,\ell,h) \\
& + \sum_{c \in \mathcal{C}} \sum_{\{\ell_1 \in \mathcal{L} \,:\, \tau_0(c,a,\ell_1,h)=t\}} A_0(c,a,\ell_1,h) \\
& + \sum_{c \in \mathcal{C}} \sum_{\{\ell \in \mathcal{L} \,:\, h \in \mathcal{H}(c,\ell)\}} \sum_{\{h' \in \mathcal{H} \,:\, \tau(0,c,a,h',\ell,h)=t\}} y_0(a,c,h',\ell,h) \\
& + \sum_{c \in \mathcal{C}} \sum_{\{\ell \in \mathcal{L} \,:\, h \in \mathcal{H}(c,\ell)\}} \sum_{b \in \mathcal{B}} \sum_{\{t' \in \{1,\ldots,t-1\} \,:\, t'+\tau(t',c,a,b,\ell,h)=t\}} x_{t'}(c,a,b,\ell,h) \\
& + \sum_{c \in \mathcal{C}} \sum_{\{\ell \in \mathcal{L} \,:\, h \in \mathcal{H}(c,\ell)\}} \sum_{h' \in \mathcal{H}} \sum_{\{t' \in \{1,\ldots,t-1\} \,:\, t'+\tau(t',c,a,h',\ell,h)=t\}} x_{t'}(c,a,h',\ell,h) \\
& + \sum_{c \in \mathcal{C}} \sum_{\{\ell \in \mathcal{L} \,:\, h \in \mathcal{H}(c,\ell)\}} \sum_{b \in \mathcal{B}} \sum_{\ell_1 \in \mathcal{L}} \sum_{\{t' \in \{1,\ldots,t-1\} \,:\, \ell_1 \in \mathcal{L}(t',a,b),\, t'+\tau(t',c,a,\ell_1,\ell,h)=t\}} x_{t'}(c,a,\ell_1,b,\ell,h)
\end{aligned} \tag{2.11}
$$

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

$$
\begin{aligned}
A_{t+1}(a, \ell_1, b) \quad = \quad & \sum_{\{h' \in \mathcal{H} \,:\, L(t,a,h',b) = \ell_1\}} y_t(a, h', b) \\
& + \sum_{\{\ell_1' \in \mathcal{L}(t,a,b) \,:\, L(t,a,\ell_1',b) = \ell_1\}} \left[A_t(a, \ell_1', b) - \sum_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{L}} \sum_{h \in H(c,\ell)} x_t(c, a, \ell_1', b, \ell, h) \right].
\end{aligned} \tag{2.12}
$$

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

$$
\begin{aligned}
C_{t+1}(c, \ell) \quad = \quad & C_t(c, \ell) + \lambda(t, c, \ell) - \sum_{a \in \mathcal{A}(c)} \sum_{b \in \mathcal{B}} \sum_{h \in \mathcal{H}(c,\ell)} x_t(c, a, b, \ell, h) \\
& - \sum_{a \in \mathcal{A}(c)} \sum_{h' \in \mathcal{H}} \sum_{h \in \mathcal{H}(c,\ell)} x_t(c, a, h', \ell, h) - \sum_{a \in \mathcal{A}(c)} \sum_{b \in \mathcal{B}} \sum_{\ell_1 \in \mathcal{L}(t,a,b)} \sum_{h \in H(c,\ell)} x_t(c, a, \ell_1, b, \ell, h).
\end{aligned} \tag{2.13}
$$

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

$$
\sum_{a \in \mathcal{A}} A_t(a, b) \quad \leq \quad \bar{A}(b). \tag{2.14}
$$

Ambulance dispatch at locals constraints: For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $b \in \mathcal{B}$, $\ell_1 \in \mathcal{L} \setminus \{b\}$,

$$
\sum_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c,\ell)} x_t(c, a, \ell_1, b, \ell, h) \leq A_t(a, \ell_1, b) \tag{2.15}
$$

Ambulance dispatch at bases constraints: For each $t = 1, \ldots, T$, $a \in \mathcal{A}$, $b \in \mathcal{B}$,

$$
\sum_{c \in \mathcal{C}} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}(c,\ell)} x_t(c, a, b, \ell, h) \leq A_t(a, b) \tag{2.16}
$$

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 base $b \in \mathcal{B}$ 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 base $b \in \mathcal{B}$ 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 base $b \in \mathcal{B}$; 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 base $b \in \mathcal{B}$ at the beginning of time t; and let $g_t(a, \ell_1, b)$ 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.

If a call arrives at time 0, then our objective is to minimize

(2.17)

$$\sum_{a\in\mathcal{A}(c_0)}\left[\sum_{b\in\mathcal{B}}\sum_{h\in\mathcal{H}(c_0,\ell_0)} f_0(c_0,a,b,\ell_0,h)x_0(a,b,h)+\sum_{h'\in\mathcal{H}}\sum_{h\in\mathcal{H}(c_0,\ell_0)} f_0(c_0,a,h',\ell_0,h)x_0(a,h',h)\right.$$
$$\left.+\sum_{b\in\mathcal{B}}\sum_{\ell_1\in\mathcal{L}(0,a,b)} f_0(c_0,a,\ell_1,b,\ell_0,h)x_0(a,\ell_1,b,h)\right]+\sum_{a\in\mathcal{A}}\sum_{h\in\mathcal{H}}\sum_{b\in\mathcal{B}} f_0(a,h,b)y_0(a,h,b)$$
$$+\sum_{t=1}^{T}\sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}(c)}\sum_{\ell\in\mathcal{L}}\sum_{h\in\mathcal{H}(c,\ell)}\left[\sum_{b\in\mathcal{B}} f_t(c,a,b,\ell,h)x_t(c,a,b,\ell,h)+\sum_{h'\in\mathcal{H}} f_t(c,a,h',\ell,h)x_t(c,a,h',\ell,h)\right.$$
$$\left.+\sum_{b\in\mathcal{B}}\sum_{\ell_1\in\mathcal{L}(t,a,b)} f_t(c,a,\ell_1,b,\ell,h)x_t(c,a,\ell_1,b,\ell,h)\right]+\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}\sum_{h\in\mathcal{H}} f_t(a,h,b)y_t(a,h,b)$$
$$+\sum_{t=0}^{T}\left[\sum_{c\in\mathcal{C}}\sum_{\ell\in\mathcal{L}} g_{t+1}(c,\ell)C_{t+1}(c,\ell)+\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}\left\{g_{t+1}(a,b)A_{t+1}(a,b)+\sum_{\ell_1\in\mathcal{L}(t+1,a,b)} g_{t+1}(a,\ell_1,b)A_{t+1}(a,\ell_1,b)\right\}\right]$$

If an ambulance completes service at a hospital at time 0, then our objective is to minimize

(2.18)

$$\sum_{b\in\mathcal{B}} f_0(a_0,h_0,b)y_0(a_0,h_0,b)+\sum_{\{c\in\mathcal{C}\,:\,a_0\in\mathcal{A}(c)}\sum_{\ell\in\mathcal{L}}\sum_{h\in\mathcal{H}(c,\ell)} f_0(c,a_0,h_0,\ell,h)y_0(a_0,h_0,c,\ell,h)$$
$$+\sum_{t=1}^{T}\sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}(c)}\sum_{\ell\in\mathcal{L}}\sum_{h\in\mathcal{H}(c,\ell)}\left[\sum_{b\in\mathcal{B}} f_t(c,a,b,\ell,h)x_t(c,a,b,\ell,h)+\sum_{h'\in\mathcal{H}} f_t(c,a,h',\ell,h)x_t(c,a,h',\ell,h)\right.$$
$$\left.+\sum_{b\in\mathcal{B}}\sum_{\ell_1\in\mathcal{L}(t,a,b)} f_t(c,a,\ell_1,b,\ell,h)x_t(c,a,\ell_1,b,\ell,h)\right]+\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}\sum_{h\in\mathcal{H}} f_t(a,h,b)y_t(a,h,b)$$
$$+\sum_{t=0}^{T}\left[\sum_{c\in\mathcal{C}}\sum_{\ell\in\mathcal{L}} g_{t+1}(c,\ell)C_{t+1}(c,\ell)+\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}\left\{g_{t+1}(a,b)A_{t+1}(a,b)+\sum_{\ell_1\in\mathcal{L}(t+1,a,b)} g_{t+1}(a,\ell_1,b)A_{t+1}(a,\ell_1,b)\right\}\right]$$

3. Solution methods

3.1. **Solving the optimization problems.** The optimization problems presented in the previous section were solved using Gurobi solver and to prepare the ground for the use of more sophisticated two-stage or multistage stochastic programming models, the following extensions have been implemented too:

(i) Given a set of scenario calls over the optimization period, we solved in parallel the optimization problems given in the previous section for this set of calls. In this situation we solve as many optimization problems in parallel as there are sets of scenario calls.
(ii) We considered two-stage models where the second stage is given by a set of scenarios for future calls. We implemented a simple strategy consisting in choosing the solution with the minimal objective value (evaluated using the code from (i)) among a set of candidate solutions corresponding to ambulances in the neighborhood of the call when ambulances need to be sent to a call and a small set of bases when an ambulance finishes service.
(iii) We implemented a heuristic which is detailed in the next section.
(iv) We implemented a simulator which simulates a set of scenario calls over the day and for each scenario, computes the waiting time using the heuristic and our optimization models.

The corresponding implementations, together with a Readme file explaining the functions implemented, and how to compile and execute them can be found in the folder Code.zip of the project.

3.2. **A heuristic.** In this section, we propose a heuristic consisting in dispatching, at every time t, the nearest ambulance to each call received, if possible. The algorithm simulates ambulance routes and the calls received at each time t. For each ambulance α, let:

- $type(\alpha)$ be its ambulance type $a \in \mathcal{A}$;
- $location(\alpha) \in \mathcal{L}$ be its current location;

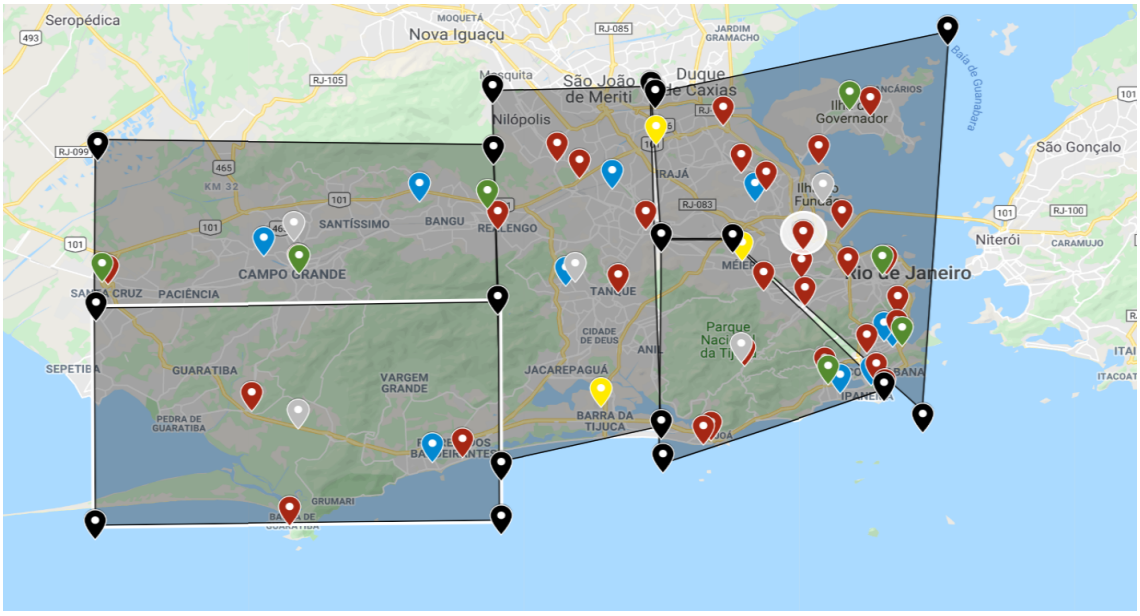

FIGURE 1. Map of hospitals, bases, and locations.

- $busy(\alpha) \in \{0, 1\}$ indicates if α is currently dispatched to a call;
- $call\_destination(\alpha)$ be the location $\ell \in \mathcal{L}$ where α is dispatched to if α is busy, or a null value otherwise;
- $hospital\_destination(\alpha) \in \mathcal{H}$ be the hospital where α must deliver patients of the call it is dispatched to, if it is busy or a null value otherwise;
- $base\_return(\alpha) \in \mathcal{B}$ be the base where α must return after it finishes serving a call if it is not busy, or a null value otherwise.

For each call κ, let:

- $priority(\kappa) \in \mathcal{C}$ be its type;
- $location(\kappa) \in \mathcal{L}$ be its location;
- $received\_time(\kappa) \in \{0, 1, ..., T\}$ be the time κ was received;
- $status(\kappa) \in \{0, 1\}$ indicates if κ was answered or not;
- $answer\_time(\kappa) \in \{0, 1, ..., T\}$ be the time κ was answered, or a null value if *kappa* was not answered yet.

Given a set A of ambulances and the sets $K(t)$ of calls for each time $t \in \{1, ..., T\}$, Algorithm 1 describes the heuristic. Initially, let κ_0 represent the call c_0, ℓ_0, with $priority(\kappa_0) = c_0$, $location(\kappa_0) = \ell_0$, $received\_time(\kappa_0) = 0$, $status(\kappa_0) = 0$ and $answer\_time(\kappa_0)$ with a null value. The best ambulance α^* and hospital h^* obtained in lines 1 and 17 are the ones where the travel time to go from $location(\alpha^*)$ to $location(\kappa)$ (or to κ_0 in line 1) plus the travel time to go from $location(\kappa)$ to $hospital\_destination(\alpha*)$ are minimum. Additionally, $busy(\alpha^*)$ must be zero and $h^* \in \mathcal{H}(c, l)$ where $c = type(\kappa)$ and $l = location(\kappa)$.

4. Numerical experiments

We applied our methodology to the management of the ambulances of Rio de Janeiro SAMU using data provided by this SAMU, in particular the history of emergency calls for the last 2 years to Rio de Janeiro SAMU, and the locations of bases and hospitals.

A map of the studied region with 10 locations, 33 bases, and 9 hospitals is given in Figure 1.

In this figure, the blue points are call locations, the red points are ambulance bases, the green points are hospitals, the yellow points are hospital bases, the black points are the vertices of the regions, the grey points are the centroids of those regions.

Algorithm 1 Heuristic that simulates the ambulance dispatch over the time horizon

Find the best ambulance $\alpha^* \in A$ and the best hospital h^* capable of answering κ_0;
if a^* and h^* exists **then**
  Update α^*: $busy(\alpha^*) \leftarrow 1$, $call\_destination(\alpha^*) \leftarrow \ell_0$, $hospital\_destination(\alpha^*) \leftarrow h^*$;
  Update κ_0: $status(\kappa_0) \leftarrow 1$, $answer\_time(\kappa_0) \leftarrow 0$;
else
  Move κ_0 to $K(1)$;
end if
for $\alpha \in A$ **do**
  Update $location(\alpha)$ according to the forecast location function L
  if α arrives at a hospital **then**
    $busy(\alpha) \leftarrow 0$;
    $base\_return(\alpha) \leftarrow b^*$, where b^* is the nearest base;
  end if
end for
for $t \in \{1, .., T\}$ **do**
  for $\kappa \in K(t)$ **do**
    Find the best a^* and h^* capable of answering κ;
    if α^* and h^* exists **then**
      Update α^*: $busy(\alpha^*) \leftarrow 1$, $call\_destination(\alpha^*) \leftarrow location(\kappa)$, $hospital\_destination(\alpha^*) \leftarrow h^*$;
      Update κ: $status(\kappa) \leftarrow 1$, $answer\_time(\kappa) \leftarrow t$;
    else
      Move κ to $K(t+1)$
    end if
  end for
  for $\alpha \in A$ **do**
    Update $location(\alpha)$ according to the forecast location function L
    if α arrives at a hospital **then**
      $busy(\alpha) \leftarrow 0$;
      $base\_return(\alpha) \leftarrow b^*$, where b^* is the nearest base;
    end if
  end for
end for

The heatmaps of the history of calls of priorities 1 (highest priority), 2 (intermediate priority), and 3 (lowest priority) are given in respectively Figures 2, 3, and 4. We calibrated a Hawkes process for the process of calls using this historical data which was used to generate scenarios of calls.

We run our optimization model for a daily planning horizon ($T = 24$ hours) and several instances corresponding to several combinations of the number of locations, bases, and hospitals. The CPU time needed to solve the corresponding problems and the time needed to load the data of these problems is given in Table 1.

Though all variables in the problem are integer and the problems are high dimensional, all problems were solved quite quickly.

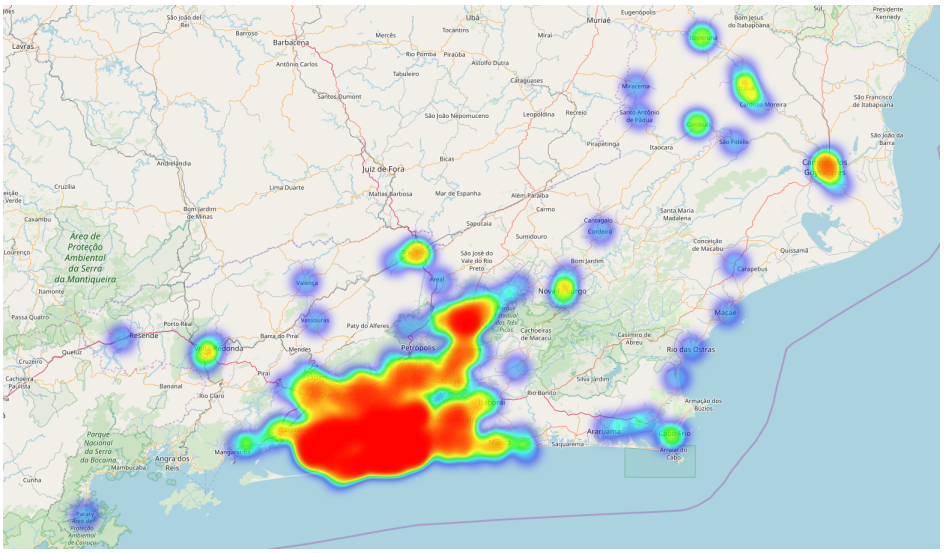

FIGURE 2. Heat map of priority 1 calls.

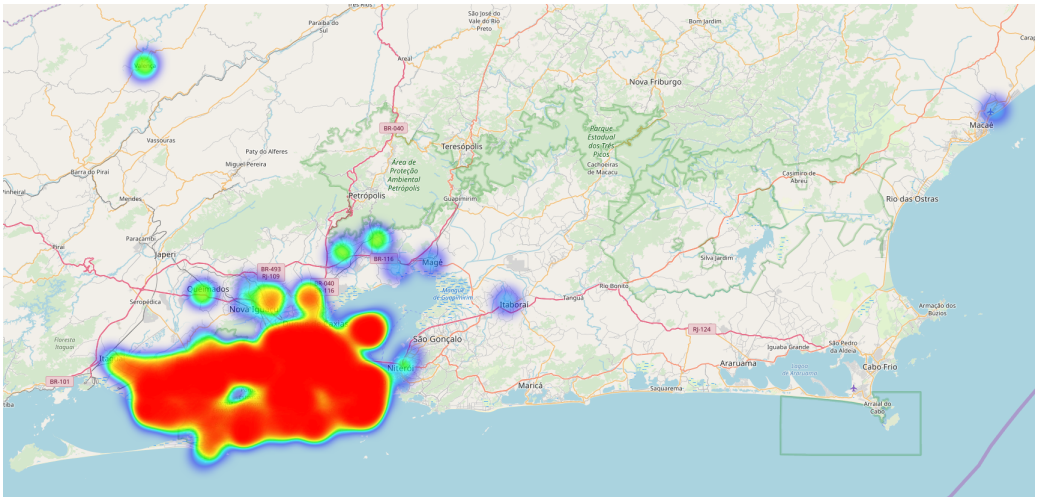

FIGURE 3. Heat map of priority 2 calls.

TABLE 1. Run times and load times (in seconds) for the deterministic model proposed, using Rio de Janeiro case study. The last line (*) was a test on a 4x8 grid map.

Times	Locations	Bases	Hospitals	Ambulances	Calls	Run Time(s)	Load Time(s)
24	5	4	4	48	1036	0.38	0.61
24	10	4	4	48	2071	2.54	0.47
24	10	6	6	48	2071	8.95	0.65
24	10	8	8	48	2071	22.87	0.94
24	10	10	10	48	2071	62.13	1.43
24	10	12	10	48	2071	85.58	1.74
24	10	14	10	48	2071	76.67	2.09
24	10	16	10	48	2071	143.21	2.39
24	10	18	10	48	2071	148.94	2.86
24	10	20	10	48	2071	189.70	3.30
24	10	22	10	48	2071	198.46	3.58
24	10	24	10	48	2071	167.13	3.95
24	32*	10	10	48	6625	410.17	17.33

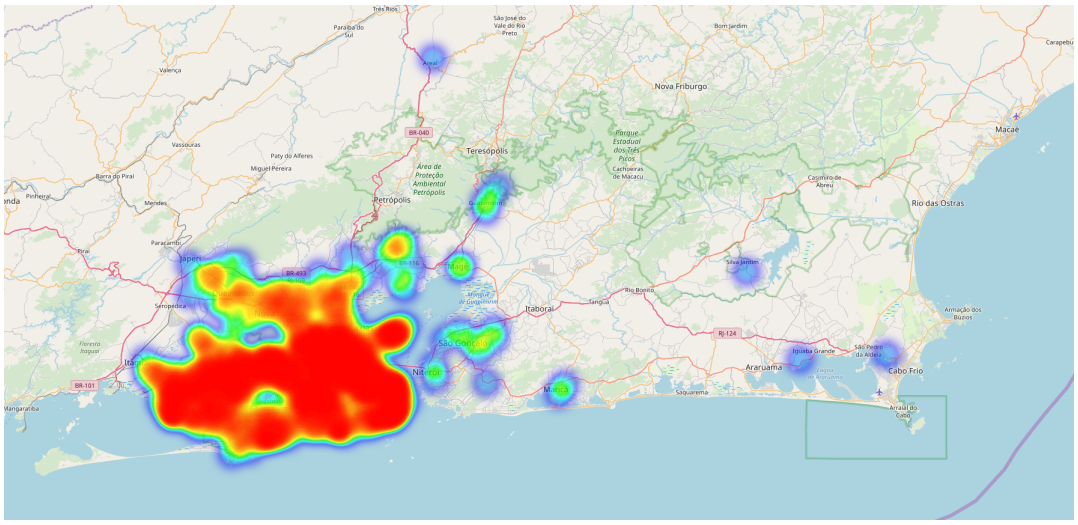

FIGURE 4. Heat map of priority 3 calls.

Finally, we run the simulator computing the average waiting time with our strategy for several instances (combinations of the number of locations, bases, and hospitals) over a set of simulated scenarios. We observe that the waiting time is quite low and much lower than the average waiting time obtained using current Rio de Janeiro SAMU strategy.

TABLE 2. Average waiting time (in seconds) and total number of answered calls of the heuristic in the simulation of calls, using the same instances as in Table 1.

Times	Locations	Bases	Hospitals	Ambulances	Calls	Waiting Time(avg)	Answered Calls
24	5	4	4	48	1036	0	1036
24	10	4	4	48	2071	7.07	715
24	10	6	6	48	2071	5.30	1077
24	10	8	8	48	2071	5.13	1105
24	10	10	10	48	2071	5.15	1103
24	10	12	10	48	2071	5.20	1098
24	10	14	10	48	2071	5.20	1098
24	10	16	10	48	2071	5.22	1096
24	10	18	10	48	2071	5.21	1097
24	10	20	10	48	2071	5.23	1095
24	10	22	10	48	2071	5.23	1095
24	10	24	10	48	2071	5.25	1092
24	32*	10	10	48	6625	10.51	257

5. CONCLUSION

To our knowledge, this paper is the first to propose a dynamic model for the ambulance operational model taking into account nearly all realistic constraints of this problem.

Numerical results are encouraging and our solution method is more efficient than the one currently used for the dispatch of ambulances by Rio de Janeiro SAMU. Therefore, our recommendation is to integrate our optimization tools to the software in charge of the management of the ambulances of Rio de Janeiro SAMU.

We intend to pursue the line of research initiated in this project considering the following extensions:

- consider two-stage models to be solved with decomposition techniques such as the L-shaped method;
- consider multistage stochastic models and write corresponding Dynamic Programming equations to be solved with SDDiP or ADP;

- refine the statistical analysis which generates scenarios of calls to include regressors such as population density in the analysis;
- use decomposition techniques such as column generation to solve more quickly the optimization problems given in Section 2.

References

[1] 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 & Management Science 70*, chapter 4, pages 77–102. Kluwer, Dordecht, 2004.

[2] Pinto L.R., Silva P.M.S., and Young T.P. A generic method to develop simulation models for ambulance systems. *Simulation Modelling Practice and Theory*, 51:170–183, 2015.

[3] 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 & Management Science 184*, chapter 11, pages 289–317. Springer, New York, 2013.

[4] M. S. Maxwell, S. G. Henderson, and H. Topaloglu. Approximate dynamic programming for ambulance redeployment. *Stochastic Systems*, 3(2):322–361, 2013.

[5] 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.

[6] 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.

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