Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

A SIMULATION MODEL FOR EMERGENCY MEDICAL SERVICES CALL CENTERS

Martin van Buuren

Department of Stochastics
Centrum Wiskunde & Informatica
123 Science Park
Amsterdam, 1098XG, The Netherlands

Geert Jan Kommer

Department for Quality of Care and Health Economics
National Institute for Public Health and
the Environment
9 Antonie van Leeuwenhoeklaan
Bilthoven, 3721MA, The Netherlands

Rob van der Mei

Department of Stochastics
Centrum Wiskunde & Informatica
123 Science Park
Amsterdam, 1098XG, The Netherlands

Sandjai Bhulai

Department of Mathematics
VU University Amsterdam
1081A De Boelelaan
Amsterdam, 1081HV, The Netherlands

ABSTRACT

In pre-hospital health care the call center plays an important role in the coordination of emergency medical services (EMS). An EMS call center handles inbound requests for EMS and dispatches an ambulance if necessary. The time needed for triage and dispatch is part of the total response time to the request, which, in turn, is an indicator for the quality of EMS. Calls entering an efficient EMS call center must have short waiting times, centralists should perform the triage efficiently and the dispatch of ambulances must be adequate and swift. This paper presents a detailed discrete event simulation model for EMS call centers. The model provides insight into the EMS call center processes and can be used to address strategic issues, such as capacity and workforce planning. We analyse results of the model that are based on real EMS call center data to illustrate the usefulness of the model.

1 INTRODUCTION

In most countries a request for emergency medical services (EMS) is done by calling a nation-wide valid emergency number. The request is answered by a centralist. Depending on the country's system, the call center handles the request, or the request is passed through to a regional call center; this may be a general call center for emergency services or a specific medical or ambulance call center. Either way, a triage is performed to determine if medical service is needed and if an urgent response is necessary. In the latter case, an ambulance is sent to the incident and care is provided. If the patient needs hospital care, transportation to a hospital is provided.

The time needed for taking the request, performing the triage, and dispatching an ambulance is part of the total response time. In case of life-threatening situations short response times are important. Hence, it is essential to have short triage and dispatch times. The triage and dispatch also need to be performed adequately, that is, the need for EMS is to be determined properly so that an ambulance is sent to incidents only in the case that the patient really needs this service. The number of ambulances is limited and it is important to have an ambulance available for dispatch when needed.

Most EMS call centers have *two classes of centralists*, working in cooperation. The first class contains the *call takers*. They handle the inbound requests and perform communication with the caller, also referred

to as the applicant. *Dispatchers* are contained within the second centralist class, they take care of the outbound calls: the dispatch process and the communication with the ambulance team and hospital. The dispatcher has logistic skills and can be supported by decision support software (DSS) to determine the most appropriate ambulance to send to the incident. Requests can be made by several disjunct applicant classes with their own class-based priority, like civilians, general practitioners, or police officers. For example, civilians have a higher priority than hospitals. Requests applied for by a general practitioner or police generally do not require extensive triage and therefore may have a shorter service time than requests from civilians. Requests from civilians need to be triaged in order to determine the need and urgency of the service. Requests for EMS are often categorized into three urgency levels: high urgent in case of a potential life-threatening situation, medium urgent when a fast response is welcome but not strictly necessary and low urgent in case the service can wait for some time. The third urgency class captures for instance planned transports from and to a hospital. A request is called honored if the call taker decides to send an ambulance. Each honored request is assigned urgency, the urgency is not subject to change during the remainder of handling the service. Requests that are not honored do not get an urgency assigned. Some patients receive emergency service without being transported to a hospital needs. In these cases only ambulance care is provided at the incident location and the service ends there. If transport to a hospital is necessary, the service ends after the patient is transferred to the hospital.

In this paper we develop a discrete event simulation (DES) model for an EMS call center. Our model simulates the communication processes at the EMS call center, and can be used to evaluate the call centers performance. It is a tool for decision makers, managers of EMS call centers, and other policy makers in managing the operational aspects of the EMS call center. All processes are assumed to be stochastic, the simulations makes use of the uncertainties in call volumes and lengths of the communications. The model gives insight into the variance of the workload in different situations, depending on the scenarios. It is possible to study economies of scale in case of increasing call volumes. Outcomes of simulation runs include the workload of the system and the waiting times for different applicant classes. These performance indicators can be examined to explore staffing levels and service level requirements.

1.1 Literature Review

The literature on call centers is broad and vast, see, e.g., the survey papers of Koole and Mandelbaum (2002) and Gans, Koole, and Mandelbaum (2003). The survey shows that call centers are mainly analysed within the framework of queueing theory. In order to keep the queueing models tractable, most papers deal with a single call type with a homogeneous pool of agents. The extension to multiple call types and a heterogeneous pool of servers leads to complex and intractable models, see Bhulai (2009) and Roubos and Bhulai (2012) for an overview of literature. Therefore, simulation is an appropriate tool to analyse more complex call centers, such as EMS call centers.

Simulation models are powerful techniques to understand tactical and operational problems, and they can also be used in EMS. Simulation in EMS originates with Savas (1969). Specific simulation models for ambulance services evolved later; see Henderson and Mason (2005), or Aboueljinane, Sahin, Jemaï, and Marty (2014) for a recent review article. These models can be used to explore facility locations, deployment of ambulances, and resources needed for optimal services. The TIFAR framework van Buuren, van der Mei, Aardal, and Post (2012) has recently been developed for simulation of ambulance services and can be used for studying various dispatch regimes. Most models are designed as a tool for the road domain of the ambulance services. The TIFAR model has been extended and includes the call center domain as well.

Kozan and Mesken (2005) have modeled the EMS call center within a simulation context. They developed a simulation model to analyze the effects of varying call volumes, personnel resources, and workload distributions on the performance of the call center. Ross (2007) studies the Toronto EMS call center and develops a simulation model to examine the effect of changes of the dispatch processes on the workload of the call center staff. For this research, communication flows at the call center are identified

and different dispatch systems are evaluated. Other preliminary work on EMS call center simulation can be found in Puts (2011) and Dwars (2013).

1.2 Contribution of our model

The main difference between an EMS call center and a regular call center lies in the fact that the EMS call center has communication tasks in the coordination of services and it has to cope with urgent requests. EMS call center centralists have extensive communications with the patient or the person requesting services, and also with the ambulance teams, hospital, and, if necessary, with police, firefighters, and other EMS call centers. Sometimes, the communications have to be performed with some sense of urgency. The importance of an efficient EMS call center together with the small number of existing models raises the need for the development of a DES model for the EMS call center domain.

2 SIMULATION MODEL

2.1 Model Description

Our model simulates the communication processes at an EMS call center in detail. These processes are considered to be queuing systems in which centralists are the servers and communication moments are the tasks. Inbound requests are seen as tasks for the server. The service time of the task depends on the applicant and the urgency. In the call taking a triage is performed and if an ambulance is to be dispatched, incident details are passed through to a dispatcher. The dispatcher determines which ambulance is to be sent to the incident. The dispatcher can have several communication moments with the ambulance team and hospital concerning, for instance, details of the incident, patient and trauma, or traffic services. Every contact between an ambulance team and the dispatcher is modeled by a task that first enters a priority queue at the dispatcher.

The call taker and dispatcher are modeled as two sequential queuing systems with different priorities. In case of multiple call takers, we model them as a server pool. The request arrival process from applicants is assumed to be a Poisson process with exponentially distributed inter arrival times. The arrival rate depends on the applicant class. All service time distributions are model inputs and are estimated using actual data sets.

The model contains a sequence of blocks representing the inter arrival times, contact probabilities, routing, and the service time of follow-up contactmoments with ambulance teams. There are $M_U \in \mathbb{N}$ urgency classes that can be assigned to a request by a call taker. We denote the totally ordered set of urgency classes $U = \{u_1, u_2, \ldots, u_{M_U}\}$. The order of urgency classes is strictly decreasing, meaning that u_1 is the highest urgency class and $u_{i-1} > u_i$, $\forall i \in \{2, \ldots, M_U\}$. Recall that every request is first handled by a call taker and, only if it is honored, is handled in a later stage by a dispatcher. We have n_1 call takers and n_2 dispatchers, who operate in the call taking and dispatcher domains, respectively; see Figure 1. Incoming requests by various classes of applicants generate tasks for the call taker and are modeled by multiple incoming streams, representing the inbound demand from civilians, GPs, police, etc. Similar applicant classes may be bundled into one incoming stream. Streams are modeled by Poisson arrivals with rate λ_i, $i \in \{1, \ldots, n\}$ for a fixed number of incoming streams $n \in \mathbb{N}$.

The call taker has a fixed number of priority queues $M_1 \in \mathbb{N}$, denoted by $pq_1^{CT}, pq_2^{CT}, \ldots, pq_{M_1}^{CT}$. The priorities are strictly decreasing, i.e. pq_1^{CT} is the priority queue with calls of the highest priority. These priority queues are also filled by a feedback loop containing the extra communication moments in case the call taker gives instructions to the applicant until the ambulance arrives. Tasks originating from the same incoming stream are led to the same priority queue.

All tasks are non-impatient, meaning that, once they have entered the system, they wait until they are served. This holds for any priority queue in our model. All priority queues have an infinite capacity and no overflow regulations. Tasks from the call taker queues are handled on a preemptive priority first-in first-served basis by a fixed number of $n_1 \in \mathbb{N}$ servers. This means that when a task enters priority queue

pq_1^{CT} and finds all servers busy, a task of the lowest priority in service is put on hold and the respective server starts serving the task from priority queue pq_1^{CT}. This task resumes service from the point at which it was put on hold when a server becomes available, but only if there are no tasks to serve with a higher priority. This behavior is commonly denoted by $M/G/n_1/\infty/PNPN$. A similar behavior can also be found at the dispatcher.

If a call taker task from an incoming stream ends service, there are two possibilities. Either with probability qh no ambulance is needed and the request exits the system. Otherwise, with probability ph an ambulance is to be sent to the incident. In the latter case the request gets an urgency $u \in U$ assigned and we say that a request is honored. The probability $ph = 1 - qh$ depends on the incoming stream.

With probability $pe(i,u)$ the call taker gives the applicant extra information; see the *Need extra information?'* conditional fork in Figure 1. If there is no extra information required, the task continues through an initial dispatch stream and $block_1$ to one of the priority queues of the dispatcher. The priority depends on the urgency of the request. However, when the applicant receives extra information in the model, the task splits into two separate tasks. One goes to the priority queues at the call takers using a feedback stream, while the other goes to the dispatcher. An extra information task exists for the system at service completion. Both types of routing to a priority queue described in this paragraph may depend on the request's incoming stream and urgency.

Task handling at the dispatcher server pool is done by a $M/G/n_2/\infty/PNPN$ policy, just like the call taker. The $M_2 \in \mathbb{N}$ priority queues are filled by honored requests of the call taker and communication moments with ambulance and hospital, represented by blocks. The service time distribution depends on the call's urgency, incoming stream, and the block where it originates from.

A block chain describes the behavior of the dispatch, driving to an incident, taking care of the patient at the incident, transportation to the hospital and delivering the patient. A block starts with a period of

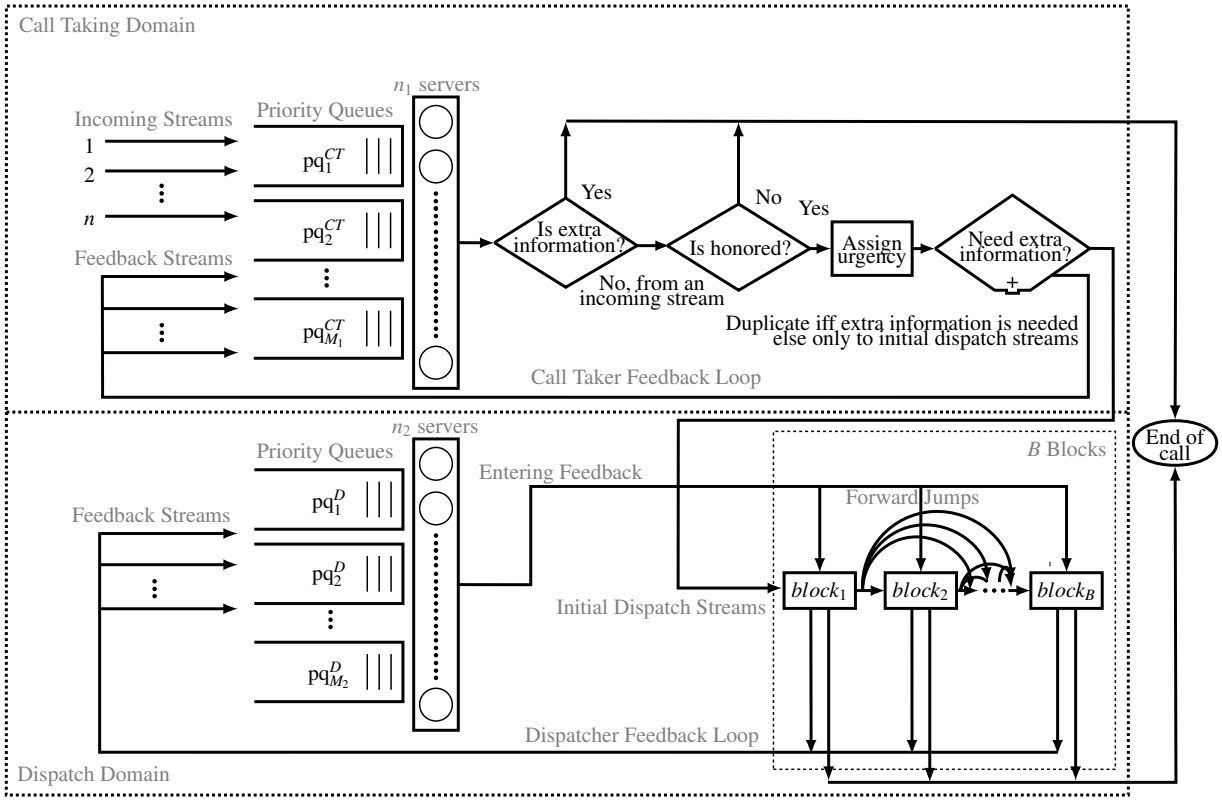


Figure 1: Queuing model representing the call taker and dispatcher domains of the EMS call center

waiting, possible contact with the dispatcher in a feedback contact, and continuation to a succeeding block or the end of the request. A closer description of the chain of blocks is found in Subsection 2.2. A special case is $block_1$. This block is always present and always leads to a feedback; this represents the dispatch of an ambulance where the priority and service time of the dispatch equal the priority and service time by the assigned server in the dispatch server pool.

2.2 Block Description

A model has a fixed number of $B \in \mathbb{N}$ blocks. We describe the structure of $block_b$, $b \in \{1,2,\ldots,B\}$, and illustrate its behavior for a request with urgency $u \in U$. The number of priority queues open to receive feedback from the blocks is denoted by M. Note that $M = M_2$.

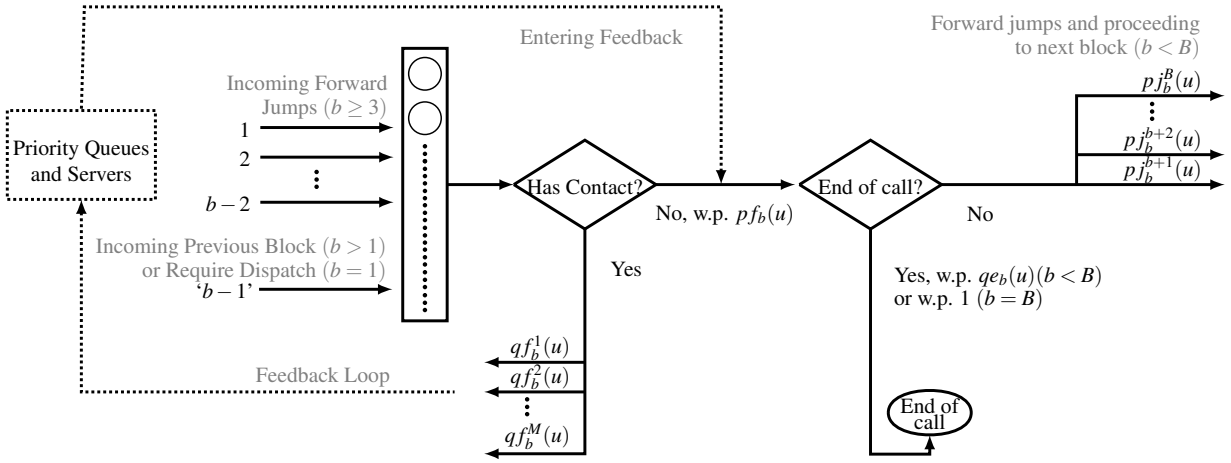


Figure 2: Block description

The incoming tasks originate from previous blocks (for $b > 1$) or the newly honored requests from dispatch (for $b = 1$); see Figure 2. To mimic the behavior of a time interval in which no communication occurs, tasks are handled by an infinite server pool with general service time. This service time can be interpreted as a delay by a driving time or contact moment with a patient at the incident location or hospital.

At this point in time a decision has to be made whether the ambulance team and centralist have a contact moment using priority queue $m \in \{pq_1^D, pq_1^D, \ldots, pq_{M_2}^D\}$. With probability $qf_b^m(u)$ a contact occurs through the feedback loop mechanism to priority queue pq_m, and with probability $pf_b(u)$ no contact occurs in this block and the request moves forward to the next decision moment. The service time is generally distributed, and the parameters may depend on b and u. Note that $pf_b(u) + \sum_{1 \leq m \leq M_1} qf_b^m(u) = 1, \ \forall\ b,u$. Note that *End of Call* can also be interpreted as a special case of a forward jump.

Right after service completion, the feedback task enters $block_b$ again. These two tasks from the feedback loop and tasks that had no feedback continue to the next decision moment. Now we determine whether there is an *End of Call*, with probability $qe_b(u) = 1 - pe_b(u)$, or, alternatively, if the task moves to a succeeding block. If $b = B$ we take $qe_b(u) = 1$; this leads to an *End of Call*. The last decision moment of $block_b, b \neq B$ tells us at which block the task continues, the so-called forward jump. With probability $pj_b^{b'}(u)$ we redirect it to $block_{b'}$. For $b' \leq b$ we have $pj_b^{b'}(u) = 0$, and $\sum_{b'>b} pj_b^{b'}(u) = 1\ \forall\ b \neq B,u$.

We end this section by some remarks on the difference between our model and Kozan and Mesken (2005). The main difference is that Kozan and Mesken assume a limited number of vehicles. When no vehicle is available the centralist has to retry the dispatch later. We only put bounds on the number of centralists and assume that a dispatch is always possible. The model of Kozan and Mesken (2005) stops at the moment of dispatch, while our model includes follow-up contact moments that also contribute to the workload of the centralists.

3 DATA AND PARAMETER ESTIMATIONS

In this section we discuss the data and the estimations of the parameters. We used a dataset of the EMS call center in the city of Utrecht in the Netherlands over a period of three months in 2011, together with an additional set with details of the ambulance services in this period.

3.1 Data Sets

The EMS call center of ambulance region Utrecht, the Netherlands, shared a database with data of all their telephone records over a period of the perio April–June 201, the so-called arbi data. This database contains the timestamps of the telephone communication, i.e. the moment the centralist lifted the handset, when it was hung up, and if it was an in- or outbound call. This arbi data set consists of 109,000 inbound and outbound telephone calls, of which 79% are inbound. The arbi data set did not include information on the content of the calls, such as the applicant class (112, police, or hospital), and whether the request was honored. The records do not indicate if the call was a new incoming call or a follow-up call by an ambulance team.

Additional information was added to the call center records by matching the arbi database with the database of the ambulance service. In the considered period there were 41,000 ambulance services. The matching is probabilistic, in the sense that we coupled the beginning of the ambulance service to a telephone contact in the arbi data. In the matching process, usually one inbound call was matched to one request and thereby to one trip. In some cases an inbound call was matched to multiple trips, we omitted these multiple matches. All together, 92% of the ambulance services were matched to an arbi record. Indirectly we used the fact that a centralist manually creates a new trip detail record at the same time he picks up the telephone. From these matched data records we identified applicant classes, urgencies, priorities, and service times. Unmatched arbi records were classified as follow-up calls. Parameters that could not be determined from the data sets were estimated through expert opinion.

3.2 Parameter Estimations

Every applicant class is mapped onto one incoming stream. Probability distributions are grouped by incoming stream and urgency. The grouping is based on applicant classes with similar service time distributions and priorities. Only civilian requests qualify for not being honored; this happens with probability 21% regardless of the urgency. Every honored request gets an urgency assigned. In Table 1 we see the incoming streams, which applicant classes every group contains, and the probability that a new inbound request is from a certain incoming stream and urgency couple. We use three urgency levels and three priorities, which we call high, medium, and low in both cases.

Hospitals and GPs both have two lines to reach the EMS call center and are able to prioritize their request. We used the distribution of the high and low urgency services to estimate the distributions of the telephone services of the call center. We take a lognormal distribution for the service time at the call taker for each incoming stream and urgency couple; see Table 2. Bolotin (1994) shows that a mixture of lognormal distributions are a good fit for call center service times. It is justified to use the urgency in the call taker's service time because the conversation leads to an urgency. Standard deviations are omitted in this paper for simplicity.

With probability 10% a request from a civilian request gets extra information, which is independent of the urgency. The time needed for the extra information is lognormally distributed, similar to the urgency independent chute time. The chute time is the time it takes the EMS team to leave the base location. We took the urgency dependent chute time as a proxy for the dispatch time distribution.

The service time distribution of the dispatcher's follow-up contacts is obtained from unmatched arbi records. Because we were unable to distinguish between the status in the process, priority or urgency, we used the same best fit lognormal distribution for every follow-up contact with a mean of 43 seconds.

Table 1: Probability distribution of incoming stream and urgency of a request. The handling mechanism of starred applicant classes is discussed in the main text.

Incoming Stream	Applicant Classes	Stream Priority	Urgency Distribution		
			High	Medium	Low
Civilian	112	High	14.64%	9.21%	0.51%
Police	Police, Fire Fighters	Medium	1.24%	1.27%	0.04%
Hospital - High	Hospitals*	Medium	0.52%	0.81%	0.00%
Hospital - Low	Hospitals*	Low	0.00%	0.00%	17.39%
GP - High	GPs*, GP Centers*	Medium	7.39%	2.78%	0.00%
GP - Low	GPs*, GP Centers*	Low	0.00%	2.78%	8.96%
Health Care Institutions	Psychiatric, Midwifes, Homecare	Medium	0.23%	0.58%	2.45%
EMS call center	EMS call center Centralists	Medium	0.72%	11.69%	1.50%
Others	Unspecified, Others	Medium	3.08%	9.87%	2.34%
		Totals	*27.81%*	*38.99%*	*33.20%*

Table 2: Average service time at the call taker (in min:sec)

Incoming Stream	Urgency		
	High	Medium	Low
Civilian	2:07	2:04	2:11
Police	1:52	1:45	2:10
Others	2:06	2:10	2:02
Health Care Institutions	1:46	2:21	2:01
Hospital - High	1:44	2:04	N.A.
GP - High	1:50	2:04	N.A.
GP - Low	N.A.	2:04	2:09
Hospital - Low	N.A.	N.A.	2:05
EMS call center	2:55	2:09	2:01
Weighted Average	*2:01*	*2:06*	*2:05*

3.3 Routing of Follow-up Calls

In the simulation runs of our model there are $B = 6$ blocks. Let us describe the follow-up contact, end of call, and forward jump probabilities for each of the blocks.

The probabilities $qf_b^{\cdot}(u)$ for a follow-up contact are listed in Table 3; they are all based on expert opinion. The probability distribution for the duration between two adjacent blocks is obtained from the trip detail records; these are the durations for the travel times, treatment duration, transfer duration at the hospital, etc. We fitted a lognormal, normal, and exponential distribution and used the mean squared error to determine a best fit. These distributions depend only on the urgency and status, and in particular not on the incoming stream. Because of its extent we omit the details of these distributions in this paper.

Forward jumps only occur from *During Patient Treatment* to *Transfer at the hospital*; that is when a patient does not require transport to a hospital. The probability that a forward jump $pj_4^6(u)$ occurs is 42% for high urgency, 40% for medium urgency, and 15% for low urgency requests. We do not use a redirect to *End of Call* in the calculations for this paper.

We look at each of the blocks individually and set the remaining parameters. Notice that the contacts occur directly after the duration where the block is named after, with $Block_1$: *Dispatch* as the only exception.

Table 3: Probability the EMS has to serve a task.

Block	Status	Urgency		
		High	Medium	Low
Block_1	Dispatch	100%	100%	100%
Block_2	Leaving the Base Location	15%	15%	15%
Block_3	Driving to Patient	30%	10%	10%
Block_4	During Patient Treatment	10%	0%	0%
Block_5	Driving to Hospital	0%	0%	0%
Block_6	Patient Transfer at Hospital	100%	100%	100%

$Block_1$: *Dispatch* This contact moment between the ambulance and EMS call center occurs for every honored request. The EMS call center contacts the ambulance and dispatches it to the incident. It often happens digitally by sending a notification to their pager. Note that the block's infinite servers have a zero service time, because there is no time delay between call taking and dispatching other than the queues of the dispatcher pool. There is a feedback loop which directs to the queue of the same urgency: $qf_1^{index(p)}(index(u)) = 1 \ \forall u \in U$. By the assumption that enough ambulances are available, a request cannot end at this stage: $pe_1(u) = 1$. When an ambulance is already on the road, there is no departure from base, but the EMS team has motivation for similar questions to the EMS call center and $block_2$ will not be skipped. Notice that in this case the block name is not fully accurate. There is a redirect to $block_2$: $pj_1^3(u) = 1 \ \forall u \in U$.

$Block_2$: *Leaving the Base Location* When an ambulance team read the incident description on the on-board monitor, there may be pressing questions as regards medical uncertainties. Another reason for contact is that the incident location might be unclear to the EMS team. The feedback loop in this block applies to the medium priority queue only. Under the assumption that a request will not be canceled at this stage, the request is passed to $block_3$: $pe_2(u) = 1, pj_2^3(u) = 1 \ \forall u \in U$.

$Block_3$: *Driving to Patient* When the ambulance arrives at the incident location, there can be a contact moment with the EMS call center. The feedback loop is for the medium priority queue only. When an ambulance arrives at a patient location it may contact the EMS call center again. The *Transfer at hospital* relates to the situation when a patient cannot be found or is not yet ready for transportation, an end of request can occur. The contact in this case has similar behavior to leaving a hospital. When there is no cancellation, the patient is always treated, i.e. the forward jump is to $block_4$: $pj_3^4(u) = 1 \ \forall u \in U$.

$Block_4$: *During Patient Treatment* Depending on the findings at the incident locations, there are multiple possible outcomes. When a patient is treated and needs transportation to a hospital, the crew can contact the EMS call center to ask them to notify the hospital's emergency department. When a patient is treated and the crew is available again, they can update the EMS call center and go to a new base location. When a patient needs transport to a hospital we redirect to $block_5$. If a patient is treated at the incident location, forward jumps occur with earlier discussed probabilities $pj_4^6(u)$.

$Block_5$: *Driving to Hospital* It is unlikely that a contact occurs at the arrival at a hospital. We include this stage as a delay, although it could be merged with $Block_6$. The only non-zero parameters are $pe_5(u) = 1$ and $pf_5(u) = 1 \ \forall u \in U$.

$Block_6$: *Transfer at the hospital.* When the ambulance has delivered the patient, the EMS call center is notified that they are available for dispatch again . This contact also has a medium priority. Since this is the last block in line, it results by definition in an *End of Call*: $qe_5 = 1$.

The simulation engine is written in a C++ and MariaDB using the TIFAR-framework that we developed at the CWI in Amsterdam. The correct working of the software was validated by intensively tracing individual tasks and call center agents. We validated the input distributions and parameters against the output database.

4 SIMULATIONS AND RESULTS

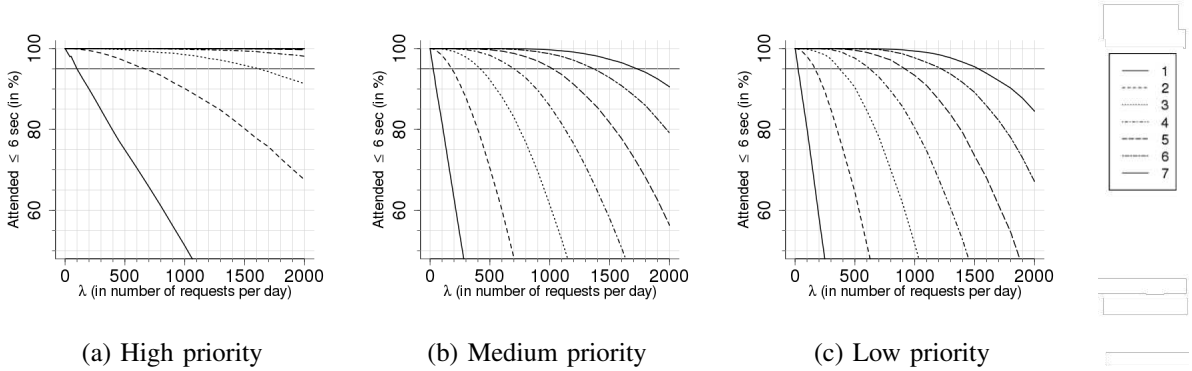
We simulated the model for various request arrival rates λ up to 2,000 calls a day, corresponding to approximately 200,000 ambulance services a year. As policies we take combinations of one to seven call takers and one to four dispatchers. Omitting very unlikely combinations gives 20 combinations. We run the model for every combination of policy and call arrival rate. In total we simulate 50,000 requests for each combination. We define the following two performance indicators:

- The fraction α_1 of the requests that are picked up by the call taker within at most r_1 time units.
- The average call center time. This is the time between a request entering the call center and the moment that an ambulance team receives their instructions to drive to the incident. Only honored requests are considered in this calculation.

4.1 Performance Indicator: Waiting Time Exceeding Threshold at Call Taker

The waiting time is the duration between the moment a request arrives at the system and the moment the call taker picks up the telephone. A request is said to be attended in time when the waiting time does not exceed 6 seconds. The ratio of requests that are attended in time and the total call volume ranges from 0 to 100%. The value for this key performance indicator should be at least 95% for high priority streams in real EMS call centers. Notice that due to the separation of call takers and dispatchers, the pool of dispatchers has no influence on this indicator. Figure 3 shows the percentage of requests that attended in time as a function of the number of requests per day.

Tasks of a certain priority are not influenced by tasks of a lower priority. This effect is clearly visible in the graphs. If there is only one call taker, the system breaks already at 100 requests per day due to simultaneity of calls. By adding a second call taker this value increases to 650 requests per day, and a call center with three call takers can handle up to 1,600 requests per day within the standards. The fourth call taker lets us handle all scenarios within the standards, however, a fifth, sixth, or seventh call taker has a noticeable effect on the medium and low priority calls.



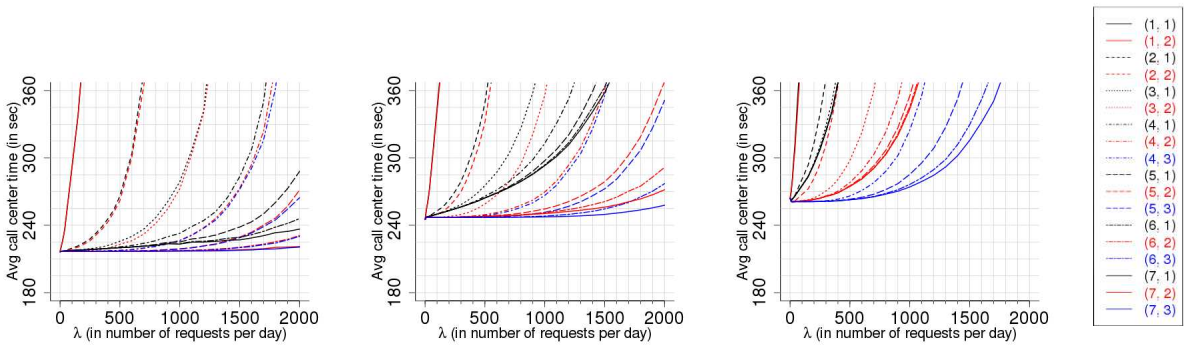
(a) High priority (b) Medium priority (c) Low priority

Figure 3: Call taker waiting time exceedings for various number of dispatchers

4.2 Performance Indicator: Average Duration until Dispatch

This indicator refers to time the EMS call center takes in the incident handling. It starts when the request arrives at the system, and ends when the dispatcher dispatches an ambulance to the incident, incorporating the two delays caused by call taker and dispatcher.

In Figure 4a we see that the curve that corresponds to the (7, 1) scenario is only slightly increasing. This means that one dispatcher is enough to handle all high urgency requests, and latency is due to the call takers. At the medium urgency we see that adding an extra dispatcher decreases the response time by one minute, at an intensity of 1,000 requests a day. A third dispatcher only has added value when low urgency response times become an issue.



(a) High urgency (b) Medium urgency (c) Low urgency

Figure 4: Average call center durations for various policies for (number of call takers, number of dispatchers)

5 CONCLUSION, AND FUTURE RESEARCH

In this paper we presented a performance and capacity simulation model for EMS call centers. The model includes two classes of centralists: call takers and dispatchers. A key feature of the model is that it includes follow-up calls from EMS teams and hospitals. The model also discriminates between multiple types of applicants which differ in priorities. The model enables EMS planners to better understand the impact of these features on the response time performance of EMS call centers.

We assume that the service time distributions of both call takers and dispatchers are independent of the workload. In reality, the time used for servicing a call may depend on the workload and service time decreases when workload increases. The inclusion of workload dependent service time distributions may have a significant impact on the response time performance of EMS call centers.

The input originates from actual call center databases. There are some estimations in our results; they cannot directly be used for management decisions without further research. Probability distributions and parameters may differ for various ambulance regions. Our model has been made for general EMS call centers, but it can also be used for other applications such as firefighter call centers.

Our model assumes switching occurs instantaneously. However, in practice switching between tasks takes time, which is a motivation for the six-second response time threshold. Inclusion of non-negligible switching times is an interesting subject.

An alternative form of EMS call centers exists, having one type of centralist who does both call taking and dispatching. It is interesting to compare the performance of this type of call center with our model. Evidently, using these generalists the overall performance improves, but leads to higher personnel cost. Further elaboration on the quantification of the tradeoff between cost and performance gain is a challenging topic for follow-up research.

ACKNOWLEDGMENTS

We thank Regionale Ambulance Voorziening Utrecht (RAVU) for sharing their EMS call center data. We thank numerous EMS drivers and nurses participating for their expert guesses. This research is supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs.

REFERENCES

Aboueljinane, L., E. Sahin, Z. Jemaï, and J. C. Marty. 2014. "A simulation study to improve the performance of an emergency medical service: Application to the French Val-de-Marne department". *Simulation Modelling Practice and Theory* 47:46–59.

Bhulai, S. 2009. "Dynamic routing policies for multi-skill call centers". *Probability in the Engineering and Informational Sciences* 23:75–99.

Bolotin, V. A. 1994. "Telephone circuit holding time distribution". Volume 1, 125–134: Elsevier.

Dwars, R. P. 2013. "Capacity planning of emergency call centers". MSc thesis, VU University Amsterdam.

Gans, N., G. Koole, and A. Mandelbaum. 2003. "Telephone call centers: tutorial, review, and research prospects". *Manufacturing Science and Operations Management* 5:79–141.

Henderson, S., and A. Mason. 2005. "Ambulance service planning: simulation and data visualisation". *Operations Research and Health Care*:77–102.

Koole, G., and A. Mandelbaum. 2002. "Queueing Models of Call Centers: An Introduction". *Annals of Operations Research* 113 (1-4): 41–59.

Kozan, E., and N. Mesken. 2005. "A Simulation Model for Emergency Centres". In *Proceedings of the International Congress on Modelling and Simulation. Advances and Applications for Management and Decision Making*, edited by A. Zerger and R. Argent, 2602–2608. Australia, Victoria, Melbourne: Modelling & Simulation Society of Australia & New Zealand Inc.

Puts, J. 2011. "Emergency Call Center: Finding a balance between costs and quality of service when dealing which emergency calls". MSc research paper, VU University Amsterdam.

Ross, E. 2007. "Simulation Analysis of Toronto Emergency Medical Service's Communications Centre". BSc thesis, University of Toronto.

Roubos, D., and S. Bhulai. 2012. "Approximate dynamic programming techniques for skill-based routing in call centers". *Probability in the Engineering and Informational Sciences* 26:581–591.

Savas, E. S. 1969. "Simulation and Cost-Effectiveness Analysis of New York's Emergency Ambulance Service". *Management Science* 15 (12): B–608–B–627.

van Buuren, M., R. van der Mei, K. Aardal, and H. Post. 2012. "Evaluating Dynamic Dispatch Strategies for Emergency Medical Services: TIFAR Simulation Tool". In *Proceedings of the Winter Simulation Conference*, WSC '12, 46:1–46:11: Winter Simulation Conference.

AUTHOR BIOGRAPHIES

MARTIN VAN BUUREN is a scientific programmer and a Ph.D. candidate at Centrum Wiskunde & Informatica (CWI) and VU University Amsterdam. His research focuses on operation research and simulation in the context of EMS. He develops strategic decision making software for emergency services. Having finished mathematical studies at TU Delft, he now participates in the REPRO ambulance planning project. His e-mail address is martin.van.buuren@cwi.nl

GEERT JAN KOMMER is a scientific researcher and a Ph.D. candidate at the Dutch National Institute of Public Health and the Environment and the VU University Amsterdam. He holds a M.Sc. in applied mathematics from the University Twente, Enschede in the Netherlands. His current research is concerned with the operational and economic aspects of acute care and ambulance services in particular. He is working on stochastic models for ambulance services in relation to health care performance at macro level. His e-mail address is geertjan.kommer@rivm.nl.

ROB VAN DER MEI is the leader of the research theme Logistics and the Industrial Liaison Officer at CWI, and a full professor at the VU University, Amsterdam. Before going to academia, he had been working for over a decade as a consultant and researcher in the ICT industry, working for PTT, KPN Telecom, AT&T Bell Labs USA, and TNO ICT. His research interests include capacity planning of EMS systems, performance modeling and scalability analysis of ICT systems, logistics, grid computing, revenue management, military operations research, sensor networks, call centers, queueing theory, and applications of big data. He is the co-author of over 125 papers in journals and refereed proceedings. He is also the leader of the REPRO ambulance planning project. His e-mail address is mei@cwi.nl.

SANDJAI BHULAI received his M.Sc. degrees in Mathematics and in Business Mathematics and Informatics, both from the VU University Amsterdam, the Netherlands. He carried out his Ph.D. research on *Markov decision processes: the control of high-dimensional systems* at the same university for which he received his Ph.D. degree in 2002. After that he has been a postdoctoral researcher at Lucent Technologies, Bell Laboratories as NWO Talent Stipend fellow. In 2003 he joined the Mathematics department at the VU University Amsterdam, where he is an associate professor in Applied Probability and Operations Research. His primary research interests are in the general area of stochastic modeling and optimization, in particular, the theory and applications of Markov decision processes. His e-mail address is s.bhulai@vu.nl.