*Proceedings of the 2018 Winter Simulation Conference*
*M. Rabe, A.A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.*

# SIMULATION BASED PREDICTION OF THE NEAR-FUTURE EMERGENCY MEDICAL SERVICES SYSTEM STATE

Tobias A. Granberg
Hien T.N. Nguyen

Div. Communication and Transport Systems
Linköping University, ITN
SE-60174 Norrköping, SWEDEN

## ABSTRACT

An ambulance dispatcher decides which ambulances to allocate to new calls, and how to relocate ambulances in order to maintain a good coverage. Doing this, it is valuable to have information about the future expected response times in different parts of the area of responsibility, as well as the expected number of available ambulances. We present a simulation model that can be used to predict this, and compare the results to a naïve forecasting model. The results show that while it is difficult to accurately predict the future system state, the simulation based prediction manages this better than the naïve model.

## 1 INTRODUCTION

Emergency medical services (EMS) providers try to find the most effective and efficient use of a limited set of resources in day-to-day operations, so that ambulances can respond to emergency calls in a time-efficient manner. The operative planning and control of the resources are done by ambulance dispatchers who decide e.g., how many and which ambulance(s) to dispatch to a specific call and also how to relocate (also called reposition) available ambulances to compensate for others that are busy. When making these decisions, the dispatchers have to consider the current state of the system including available ambulances and calls that need to be serviced. In addition, they need to take into account the probable future state of the system, which will be the result of the decisions made, as well as stochastic events. Some ambulances that currently are busy might become available in the near future, new calls will enter the system, and some available ambulances will become busy. Some parts of the future state are fairly easy to predict, e.g., when an ambulance that is currently transporting a patient to a hospital will become available, while other parts are more difficult, e.g., when and where a new call will occur. In particular, it is useful to have an estimate of the future expected response times to different parts of the area, as these will influence the operational decisions. For example, a dispatcher might consider relocating an ambulance to an area that currently has a long expected response time. However, it is possible that the situation will rectify itself, through ambulances currently en-route towards the area, either already relocated, on their way back to base, or on their way to a hospital to deliver a patient. In this case, an estimate of the future expected response times would support the dispatcher's decision.

In this paper, we focus on the problem of predicting the near-future EMS state. The study is motivated by a specific need for this kind of decision support expressed by the Swedish company SOS Alarm Sweden AB, who is responsible for all emergency medical call prioritization and ambulance dispatch in Sweden (at the time of the study). Furthermore the problem of predicting the operative near-future EMS state, and analyzing the quality of the predictions, has not been studied independently before.

Within this area, two types of previous research are particularly related to our work, where the first concerns models for ambulance location, relocation and dispatch, in which the system state evolvement is explicitly considered. The second type of related work consists of papers about computer simulation

models for EMS systems. There exist valuable reviews of simulation applications for EMS in Goldberg (2004), Henderson and Mason (2004), and Aboueljinane et al. (2013). Readers are referred to the paper of Aboueljinane et al. (2013) for a critical overview of the planning issues studied, the associated modeling assumptions, model development, verification, validation, and the results obtained. Furthermore, three recent reviews of operations research, planning and logistic studies within EMS, give a good insight into the area (Aringhieri et al. 2017; Reuter-Opperman et al. 2017; Bélanger et al. 2018).

Different descriptions and definitions of the EMS system state have been made since research within the area started to emerge. One of the most well-known is probably from Larson's hypercube queuing model (Larson 1974) where a system state is represented by a binary string where each 0/1 indicates if an ambulance is available or busy. A state transition occurs when a vehicle changes its status from available to busy (upward transition) or from busy to available (downward transition). The upward transition rate depends on the call arrival rate and the probability of dispatching the ambulance. The downward transition rate depends on the mean service time. A steady state is an equilibrium that the system is expected to reach after it has been operating for a sufficiently long time. One possible use of the hypercube queuing model is to evaluate ambulance station locations, or to analyze what happens when system changes occurs, e.g., decentralization of the operations like in Takeda et al. (2007) who use the same system state representation as in Larson (1974).

In an operational context, when deciding which ambulance to dispatch to a new call, or how to relocate available ambulances to improve the coverage, more detailed system state representations might be necessary. One nice example can be found in in Maxwell et al. (2010), where Approximate Dynamic Programming (ADP) is used to suggest where idle ambulances should be positioned. The current system state is described by the state of the ambulances and the waiting calls. Each individual ambulance state is given by the activity (e.g., *idle at base*, *serving at scene of call*, etc.), the origin and destination associated with the activity, as well as the starting time for the current activity. The calls are described by their status (*assigned to ambulance i* or *queued for service*), their location, time of arrival and priority. Having this amount of information in the system state description makes it possible to take into account e.g., remaining service times and travel times for ambulances that are not currently idle. When making relocation decisions, this information might be helpful. However, as argued by Jagtenberg et al. (2015), an elaborate state representation also gives a large state space, which might be computationally intractable as well as difficult to explain to EMS personnel. Instead, they define the state space as the destinations of all idle ambulances, and develop a relocation heuristic that gives better performance compared to a static solution. Schmid (2012) use a state description similar to the one in Maxwell et al. (2010), including both spatial and temporal information, to solve a relocation and dispatching problem in Vienna using ADP. ADP is also used by Lam et al. (2017), who develop a relocation model for Singapore. Like in Maxwell et al. (2010), they use simulation to train the value function approximation necessary in the ADP model, simulating the EMS operations until the next event. This is in contrast to Schmid (2012) who instead approximate the value function around the post-decision state (e.g., immediately after a relocation decision has been made). This way, it is not necessary to predict or evaluate the next state that the system will reach.

Enayati et al. (2017) define the system state as the state of the individual ambulances, including their current location, if they are idle or busy, and accumulated busy time. The last factor is used to ensure that relocations are performed in a way that restricts and distributes the workload for the EMS personnel. In their work, as in many other relocation models, it is assumed that the ambulances are busy, and cannot accept new calls while relocating. This is in contrast to Zhang et al. (2010), where the relocation route is planned, taking into account the possibility that the ambulance will get a call while traveling. Thus, it may be beneficial to travel through high call density areas even if this makes the total route longer.

Van Barneveld et al. (2016) define the system state as the current location or destination and phase of the ambulances, where the possible phases are 0) at base or relocating, 1) traveling to scene, 2) busy at scene, 3) transport to hospital, and 4) busy at hospital. While temporal information is not explicitly stated as part of the state representation, the authors also keep track of the service times, at least for Phase 4

ambulances, while arguing that the time until it may be possible to relocate Phase 1-3 ambulance is too long or too uncertain. Using a penalty function, relocations can be selected in a way that minimizes the unpreparedness of the state. This is in contrast to Andersson and Värbrand (2007) who try to maximize the preparedness. In their work, an explicit system state representation is not given, but their preparedness measure implicitly, and heuristically, takes the future system state into account. Bandara et al. (2014) analyze dispatching strategies, using discrete event simulation. They define the system state as the status of all ambulances, who can be either *busy* or *available*, and evaluate different dispatching rules for an EMS system with calls of different priority, with the aim of improving the patient's survivability.

Also related to our work is the system Optima Live, described in Mason (2013). It includes a feature displaying the future coverage, based on the prediction of the future system state. The state is defined as the updated availability and position of all ambulances once their current activities (including relocation moves) are completed. The prediction horizon is not explicitly stated, but it can be assumed to be the time for all ambulances to complete their current activities. When an ambulance finishes its current activity, its subsequent activity is not predicted and simulated. It is possible, however, to calculate the expected availability for each ambulance. Future calls are displayed as predicted call rates, giving the user the ability to identify areas with inadequate preparedness, i.e., poor coverage and high expected call frequency. The difference from our work is that the prediction does not consider new calls entering the system during the forecasting period, and that only one single sample of the future state is produced. Furthermore, as far as we know, our study is the first to try to measure the accuracy of a future EMS state prediction model.

## 2 AN EMS FUTURE STATE PREDICTION MODEL

### 2.1 The Conceptual Prediction Model

We define the EMS system state at a particular time as the position, activity and time spent on the activity of the ambulances as well as the position and status of calls at that time. The state evolves stochastically over time and the main sources of randomness include the arrival of new calls, travel times, and service times (including time at site and time at hospital).

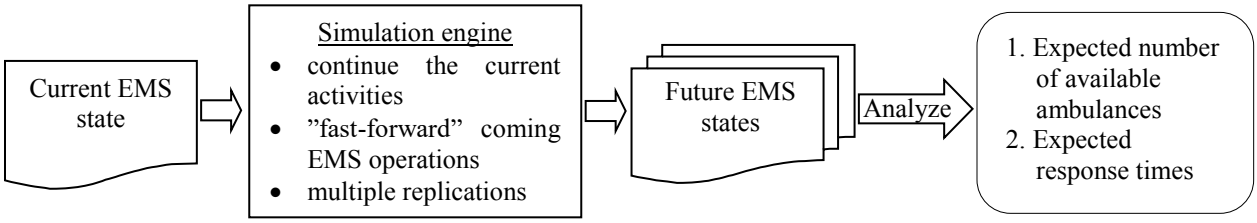


Figure 1: Conceptual model for predicting the future EMS state.

As the system state space grows quickly when the number of ambulances and calls increase, we base the prediction model on simulation (see Figure 1). The conceptual model for predicting the future system state starts from a given state and simulates the continuation of current activities, e.g., an ambulance on the way to a call site, or transporting a patient to a hospital. It also generates new calls, and dispatches ambulances to these calls, and keeps track of all related activities until the end of the prediction horizon, which is assumed to be 30-60 minutes.

Based on a snapshot of the current EMS state, the simulation engine provides a possible future system state. As the EMS system is dynamic and stochastic in nature, multiple replications have to be run from the same snapshot to capture various samples of the future state. These observations are then further analyzed to derive useful information for ambulance dispatchers, i.e., the expected number of available ambulances and the expected response times to zones or other locations in the area of responsibility.

## 2.2 A Discrete Event Simulation Model for Västra Götaland County, Sweden

### 2.2.1 System Characteristics and Input Data

Västra Götaland County covers an area of 25,388 km$^2$ on the western coast of Sweden with a population of 1,590,604 accounting for 17% of Sweden's total population in 2011. The same year, the EMS system handled about 200,000 calls, where 67% were emergency calls and 33% were non-emergency calls and patient transportations. An emergency call is labeled priority 1 or 2, where priority 1 is reserved for the most urgent and life threatening calls. A non-emergency or patient transport call is given priority 3 or 4, where priority 4 calls in general are handled by dedicated patient transport vehicles, i.e., not ambulances. A call is also given a description indicating the cause of the emergency or the health condition of the patient, e.g., trauma or allergy. A total of 90 ambulances were distributed over 47 stations. The operational fleet size varied between 60 and 90 ambulances in respect to the demand changes during day and night, weekday and weekend.

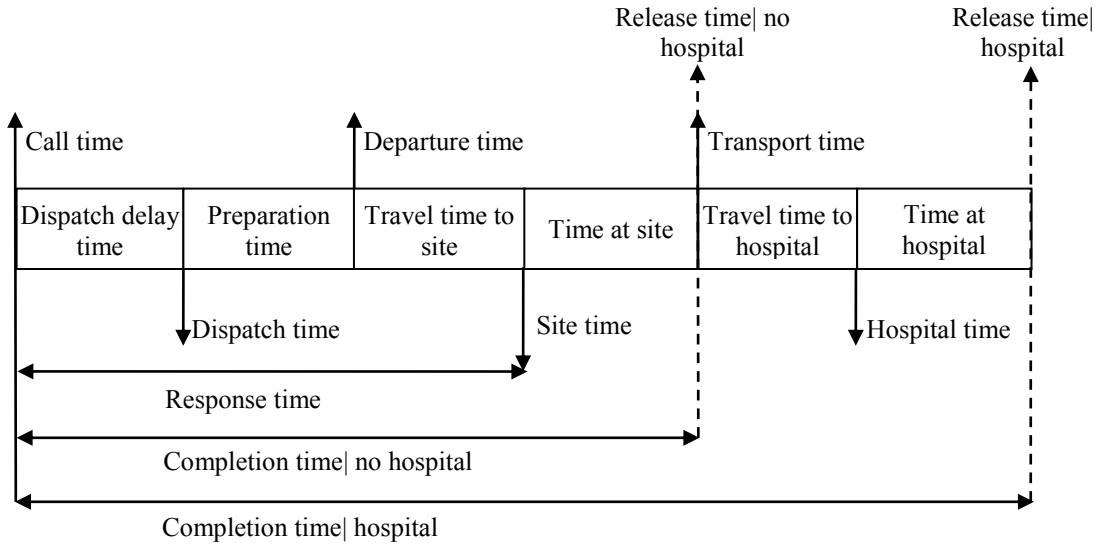


Figure 2: Time stamps and derived information from the call data.

The company SOS Alarm is responsible for receiving emergency calls and performing the central dispatch service. The dispatch center is equipped with a communication and call handling system called CoordCom (http://www.carmenta.com/en/products/carmenta-coordcom/). One part of CoordCom is a geographical information system (GIS) called ResQMap. In ResQMap, the county has been divided into a grid of 1710 cells, where each cell has the size of 4x4 km. The system is capable of displaying, in real time, the positions and statuses of vehicles and the positions of calls that have been received or served. It can also visualize the coverage and the preparedness for each zone (for more information, see Andersson and Värbrand 2007). Directly from this system, it is possible to extract the expected number of calls per hour for each zone, deterministic and static travel times between any pair of zones (i.e., the travel times do not vary by time of day or day of the week), and a large number of service logs and historical call data.

The logged call data includes the following information: call arrival time, geographical coordinates, priority, call description, dispatched ambulance, station of dispatched ambulance, hospital destination (if available), when the ambulance receives the mission, when it leaves for the incident, when it reaches the patient, when it transports the patient to a hospital (if available), when it arrives at the hospital (if available), and when it is released from the mission. Figure 2 displays the time stamps, service periods, and travel periods in the call data. Depending on if a call requires a patient transportation to a hospital or not, there are two different release times, and consequently two different completion times.

The system operations are continuously logged, and can at any point in time be extracted to a *snapshot file*. The snapshot provides information about the fleet size, the status and position for the vehicles, and status and position for the active calls.

### 2.2.2 Components of the Simulation Model

The simulation model was developed in C#. It generates a stream of priority 1, 2 and 3 calls, to which ambulances are dispatched. The model is discrete event based and the possible event types in the system are *a call arrives*, *an ambulance is assigned to a call*, *an ambulance departs for an incident site*, *an ambulance arrives at the site*, *an ambulance leaves the site for a hospital*, *an ambulance arrives at a hospital*, *an ambulance finishes at site or a hospital* and *an ambulance arrives at a base*. The simulation keeps track of event occurrences, updates the system state, and creates the new consequent event as a result of the current one. A run is terminated when the prediction horizon has been reached.

The simulation model has the following components:

- A snapshot parser that analyzes a system state given to the model by the ResQMap GIS, and determines the next event for each ambulance.
- A stochastic call generator that predicts when and in which zone an emergency call occurs. The priority (1, 2 or 3) of the call is also determined.
- A deterministic traveling model that determines the route and the travel time between any pair of origin and destination.
- Models and rules for dispatching, handling waiting calls, and hospital selection.
- Stochastic service time estimations (for the delay time before dispatch, the preparation time, the time on site, and the time at hospital) and the probability for patient transportation.

The *snapshot parser* analyzes a snapshot file to get data about active ambulances and their current activities, as well as about active cases. The model then predicts subsequent events by estimating the expected service times or travel times of the present activities. For example, if an ambulance just has been assigned to a call, the next event is *an ambulance departs for an incident site*. The simulation model in this case estimates the preparation time for the ambulance crew. As all relevant time stamps are known, i.e., the starting time for the simulated period, when the present activities started, and when the position of the ambulance was updated, it is easy to determine how large part of the service time or travel time that has elapsed. Using a suitable probability distribution, it is then possible to deduce the expected remaining time of the current activity.

The *call generator* gets predicted call rates from the ResQMap GIS. The call rate (i.e., the expected number of priority 1, 2, 3 calls per hour) is available for every zone, every hour of the week (or 24x7 hourly slots). The interval between call arrivals in the county is assumed to follow an exponential distribution with the mean equal to the sum of the call rates over the zones. The ratio between the call rate at a zone and the call rate in the county is used as the probability that a call originates in that zone. The call priority is determined according to the priority probability calculated from the historical data for the whole county.

The *traveling model* receives the travel times from the ResQMap GIS. In order to identify the position of an en-route vehicle (which is necessary e.g., when deciding which ambulance to dispatch to a new call), the route between the origin and the destination has to be known. Vehicles are assumed to travel along the shortest paths that are pre-calculated from the grid representation of the county. Whenever an en-route ambulance is a candidate in the dispatch process, its position is calculated and updated.

*Waiting calls* are calls that have not been assigned to ambulances, and are in queue to be served. While it is rare that this occur, should it happen, they are handled in the order of priority. Priority 1 calls are processed first, in a first-in-first-out manner.

The *Dispatching procedure* will select the closest (i.e., shortest response time) ambulance candidate. An ambulance is a candidate for dispatch if satisfying one of the following conditions:

- Available at the station.
- Available, on its way to the station.
- Soon available, i.e., the ambulance is delivering a patient at a hospital, and the predicted remaining time at the hospital is less than 5 minutes.
- Unavailable and on the way to an incident site, but the new call has a higher priority. In this case the old call is interrupted and the ambulance is reassigned to the new one. The interrupted call will be assigned to a new ambulance, with a deterministic preparation time.

We modeled the dispatch delay time, preparation time, time at site, and time at hospital by probability distributions that were fitted to historical data from 2010 and 2011. EasyFit version 5.5 was used to find well fitted distributions. The probability that patients are delivered to hospitals from a call site and the probability distribution of call priorities were also calculated based on the analysis of the data. More detailed information of the distribution fitting and can be found in Nguyen (2015).

Regarding hospital selection, the simplest rule is selecting the hospital closest to the incident. The data analysis, however, showed that 31.6% of cases were not delivered to the closest hospital. Thus, we estimated destination probabilities for each demand zone and for each call priority using the historical data.

After the event *an ambulance finishes at site or a hospital*, the ambulance will return to the home station if there are no calls in queue for being served.

### 2.2.3 Verification

A verification was performed to ensure that there were no logical faults, bugs or errors in the simulation model. The distributions of generated random values were checked against the input distributions through graphical comparisons. We used scatter plots to show the correlation between simulated and historical call volumes, as well as between simulated and empirical hospital workloads. The snapshot parser was tested with various snapshots extracted from service log files. The parser not only retrieves information about ambulances, calls and current activities. It also normalizes the information according to the scope of the simulation, detects conflicts e.g., when one ambulance is assigned to two different calls, and detects out-of-date information e.g., when the status and location of an ambulance have not been updated although it has finished a mission. Detailed tracing of entities and events in the simulation was also performed for a number of snapshots, using a simulation period of one hour.

To check that the simulation was able to represent the Västra Götaland EMS system, the model was tested in a traditional long term scenario, running the model for 50 months of operations. Differences between the simulation results and historical data in number of calls, travel times, services times and the resulting response times and completion times were analyzed, and it was concluded that the model was valid for use as an engine for the near future state prediction.

## 3 VALIDATION OF THE FUTURE STATE PREDICTION MODEL

### 3.1 Experiment Design

To test the forecast accuracy of the simulation model, we extracted snapshots from the system log files and performed two sets of experiments; one where the state 30 minutes in the future was predicted, and the equivalent for 60 minutes. The simulated states were compared to the real states, e.g., a 60 minute simulation starting with the snapshot at 10:00 a.m. will give a prediction of the system state at 11:00 a.m. This is compared to the real system state in the snapshot at 11:00 a.m.

The demand is typically highest during daytime on weekdays. On other times, the frequency of calls is lower, and it is easier for the dispatcher to manually manage the operations. Thus, the prediction model would be most useful during high demand. Therefore, snapshots were extracted for daytime 7:00 – 18:00, for 23 weekdays, Monday-Friday, in October and November 2010. Figure 3 shows some facts about the real system in the validation period: the locations of stations and hospitals, estimations of the response times to individual zones (minutes), and the real call volumes for high demand zones. During the validation period, the county had a total of 5234 calls at an average rate of about 23 calls per hour. The median response times were estimated based on the travel times for the closest ambulance to each zone in the snapshots. Mean response times were also calculated, but are affected by large outlier values; thus the median values provide a better illustration of the typical response times in the county. Figure 3 also shows a high call frequency in the south west, where Göteborg, the largest city in the county is located.

In total we ran $N_{30} = 379$ experiments with a 30 minutes time horizon and $N_{60} = 274$ experiments with 60 minutes. We evaluated the prediction of a future state using the following output:

1. $a_n^f$= forecasted number of available ambulances in experiment *n*, compared to $a_n^r$= real number of available ambulances in experiment *n*.
2. $t_{ni}^f$ = the forecasted expected response time to zone *i* in experiment *n*, compared to $t_{ni}^r$= the real expected response time. The response time to zone *i* is estimated by the travel time for the closest available ambulance.

In a validation experiment, the simulation was run with multiple replications to get different samples of the future state. For a set of replications, the expected number of available ambulances is calculated as the mean value. The expected response time for a single zone in one replication is estimated as the expected response time for the closest available ambulance. Over a set of replications, the response time for the zone is estimated as the median of the response times in each replication:

$t_{ni}^f = med(t_{ni1}^f, t_{ni2}^f, \ldots, t_{niK}^f)$, where $t_{nik}^f$= the forecasted expected response time to zone *i* in experiment *n,* replication $k = 1, \ldots, K$.

To determine the necessary number of simulation replications, we studied how the standard deviations of the expected region-wide response time and the expected number of available ambulances varied with the number of replications. The region-wide response time is calculated as the simple average of the individual zones' expected response times (i.e., not a weighted average). We concluded that $K = 100$ replications are more than enough to capture the uncertainty in forecasting the future system state for the selected system measures.

Through the validation, we wanted to investigate the accuracy of the simulation model in predicting the abovementioned system measures. The forecast

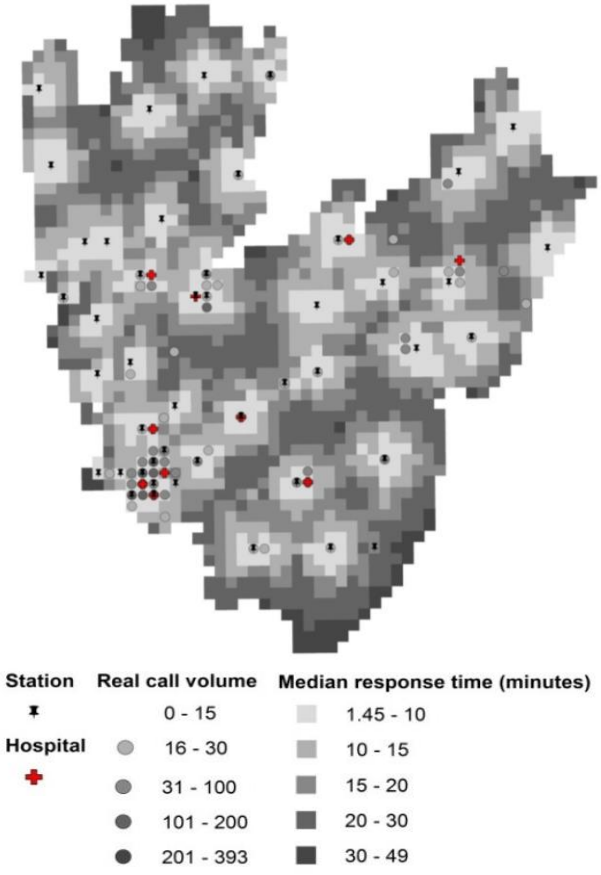


Figure 3. Historical, median response times and call volumes for the validation period.

accuracy was evaluated by the mean absolute error (MAE), the absolute percentage error (APE), the mean absolute percentage error (MAPE), and the root mean squared error (RMSE). We also compared the predictions by the simulation model to a naïve prediction that assumes that the system is static during the forecast horizon and uses the current system state as the forecast of the future system state. When considering the two selected evaluation measures, the naïve prediction is not necessarily a bad one. The number of available ambulances can be expected to remain the same, as long as the call frequency does not change significantly from one point in time to the next (some ambulance will become busy while others will become available). Also, given the highly stochastic nature of the system, it might be similarly accurate to assume that no new calls will appear during the prediction period, as trying to predict where the new calls will appear.

Studying the forecast errors, it is possible to determine how the prediction horizon affects the prediction capability, and if the model can be used to support operational EMS decision making. The comparison between the simulation prediction and the naïve prediction can be used to determine if an advanced simulation based prediction model is necessary, or if simpler predictions work just as well.

### 3.2 Validation Results

We ran the validation experiments on a computer with an Intel Core Duo 3.0 GHz processor, 4 GB RAM, running under Windows 7. In the experiments with 60 minutes simulation time, it took 3.5 seconds on average and a maximum of 5 seconds to run 100 simulation replications and perform the calculation of the output measures. The 30 minute simulation runs, also configured with 100 replications, spent 2.4 seconds on average performing the prediction.

Table 1: Errors when predicting the number of available ambulances.

| | Naïve-60 | Simulation-60 | Naïve-30 | Simulation-30 |
|---|---|---|---|---|
| Mean number of available ambulances | 46.7 | 48.7 | 47.6 | 46.3 |
| MAE | 4.7 | 4.3 | 3.6 | 3.6 |
| RMSE | 5.9 | 5.4 | 4.6 | 4.5 |

The results (see Table 1) show that the simulation based predictions are slightly better than the naïve predictions when looking at the MAE and the RMSE (see (1) and (2), where N is the number of experiments).

$$MAE = \frac{1}{N}\sum_{n=1}^{N} |a_n^r - a_n^f| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N} (a_n^r - a_n^f)^2} \tag{2}$$

This advantage for the simulation based predictions seems to increase with the prediction horizon, i.e., the difference between the naïve and the simulation model is more distinct when the prediction horizon is 60 minutes.

The APE for Experiment *n* is calculated as:

$$APE_n = \frac{|a_n^r - a_n^f|}{a_n^r}. \tag{3}$$

Figure 4a compares the distributions of APEs by showing the ratio of experiments (y-axis) that have APEs less than or equal to a value on the x-axis, when predicting the number of available ambulances.

For example, an APE less than or equal to 10% is found in 56% of the naïve-60 predictions, in 65% of the simulation-60 predictions, and in 70% of the naïve-30 and the simulation-30 predictions.

For evaluating response times to individual zones, we investigated the MAEs and the MAPEs, calculated as

$$MAE_i = \frac{1}{N}\sum_{n=1}^{N} |t_{ni}^r - t_{ni}^f|, \tag{4}$$

$$MAPE_i = \frac{1}{N}\sum_{n=1}^{N} \frac{\left|t_{ni}^r - t_{ni}^f\right|}{t_{ni}^r} \tag{5}$$

for each zone *i*. We compared the prediction performances for the different scenarios by the ratio of zones that have MAPEs less than or equal to 5%, 10%, 15%, etc. as shown in Figure 4b. For example, the simulation-30 scenarios produced a MAPE ≤ 10% for 853 zones (50% of the zones) and a MAPE > 50% for 56 zones (3% of the zones). This is better than the naïve-30 predictions, at least when considering scenarios with a MAPE less than 35%. The difference between the two prediction models is even more apparent when looking at the 60 minute predictions, where the simulation-60 scenarios in general have significantly lower MAPEs than the naïve-60 scenarios. Note that the overlap of the simulation-60 and naïve-30 results is just a coincidence.

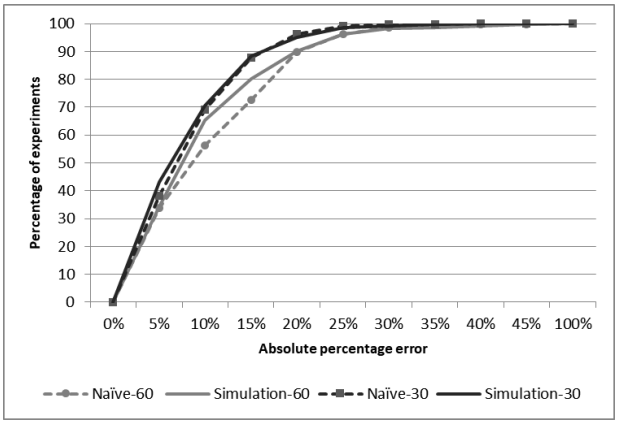


Figure 4a:. Distributions of APE when predicting the number of available ambulances.

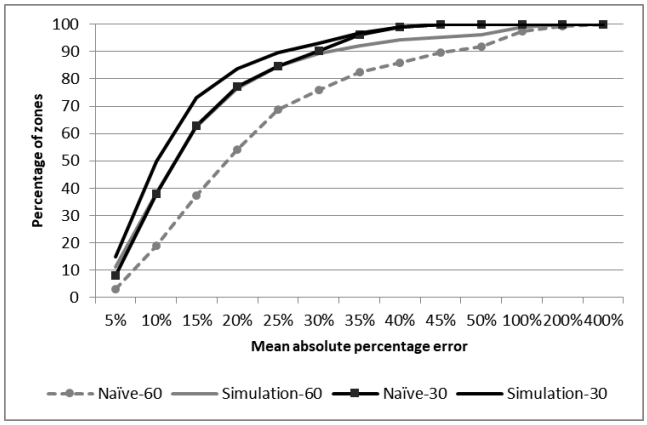


Figure 4b: Distributions of MAPE when predicting individual response times to zones.

The spatial distribution of individual MAPEs is illustrated in Figure 5 for the 30 and 60 minutes predictions. Only MAPEs > 10% are visualized, since predictions with MAPEs below 10% are considered as good. Many zones where the MAPEs are above 30% have ambulance stations or hospitals located in them, or nearby. This is not surprising, since the response times to these zones in general are short (which can be observed in Figure 3), giving small actual values, $t_{ni}^r$. Thus the MAPEs will become large.

Looking at the MAE, it might be reasonable to expect higher values further away from the ambulance stations, where the response times usually are longer. However, like the MAPEs, the MAEs are larger close to the stations. This might be explained by the fact that sometimes the ambulance is available at the station, in which case the response time to zones around the station will be short. But when the ambulance is busy, the response times will be significantly longer. A predicted response time for a zone, is calculated as the median of 100 replications. For a zone close to an ambulance station the median will in many experiments be short. However, not necessarily as short as the expected response time when the

ambulance is available at the station, since in some replications, the ambulance will be busy. This is compared to the estimated real response time for the zone which will be short if the ambulance in reality was available at the station, or long, if the ambulance was busy. Thus, in many cases, it is likely there will be an error when comparing the response times close to the stations. Zones with a low MAE is often found between ambulance stations, since if one ambulance is busy, an ambulance from another station might reach the zone in roughly the same time. It is also apparent from Figure 5 that zones at the edge of the area have high MAEs, which can be explained by the same reasoning as above.

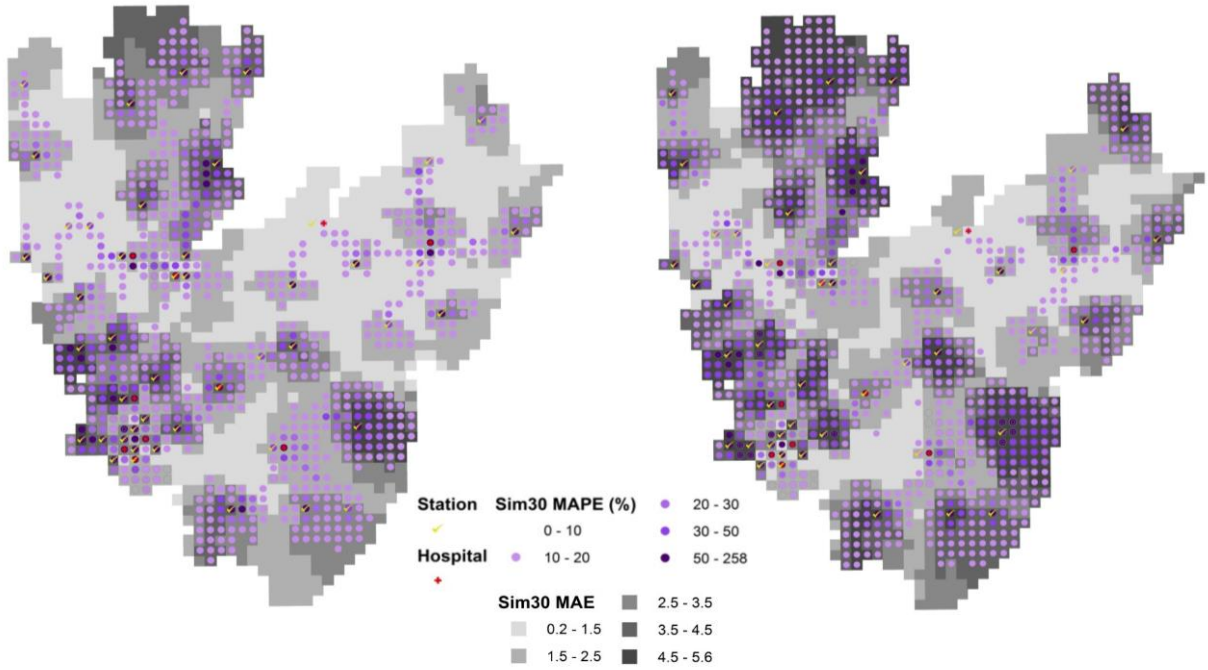


Figure 5: MAEs and MAPEs of response times to individual zones for the 30 minutes simulations (left) and 60 minutes simulations (right).

## 4 DISCUSSION - PRACTICAL USE OF THE PREDICTION MODEL

The estimated number of available ambulances in e.g., 30 minutes or one hour, gives the ambulance dispatchers information about the expected system load. Of course, the dispatchers would like to know exactly where every available (and busy) ambulance will be located in the future. While this prediction is fairly easy for some ambulances, e.g., the ones currently assigned to a call, it is extremely difficult for others, e.g., currently available ambulances in high or medium demand areas. Predicting the expected availability of single ambulances is easy; it can be estimated as the number of replications in a simulation run that an ambulance has been available, divided by the number of replications. This information is however not particularly valuable without a good estimate of the location, something that is much more difficult, at least for ambulances that have a high probability of becoming allocated to new calls during the simulation period. A simulation run with 100 replications, gives 100 possible future outcomes. Worst case, this means 100 different possible future locations for a single ambulance – information that is particularly difficult to translate to useful decision support.

Since it is too difficult to predict where all individual ambulances will be located, it is better to concentrate on presenting the expected zone response times rather than the spatial distribution of available ambulances. This information can be used to visualize whether the system can provide adequate response times to all or some potential demand zones in the near future. In a prototype implementation in

ResQMap of the prediction model, the expected response times are visualized as red zones having an expected response time longer than 30 minutes, orange zones between 15 and 30 minutes, and green zones less than 15 minutes. The application can, for example, be used by dispatchers to manually plan relocations, to improve coverage of zones with long expected response times. Specifically, it is useful for evaluating manual dispatch and relocation decisions; for example, a relocation is usually made to improve the preparedness for an area that currently has poor coverage. However, it is not always easy for the dispatchers to keep track of all ambulances that are en-route, or soon will become available. Thus, before executing a relocation decision, a dispatcher may find it useful to first check the predicted future state, to make sure that the area will not likely be covered in the near-future by ambulances that currently are busy or elsewhere.

## 5 CONCLUSION

There exist numerous studies where simulation has been used to predict the long term (steady state) performance of EMS systems. A typical purpose is to evaluate a system configuration by estimating the ratio of emergency calls that will be serviced within a time limit (e.g., 10 minutes). In this paper, we use simulation modeling to create support for the operational planning and control of ambulances. Starting with a snapshot of the current system state, a discrete event simulation model is used to predict the evolvement of the system for a limited time into the future. From the prediction of the future state, information about the number of available ambulances and expected response times to individual demand zones can be extracted.

For the EMS system in Västra Götaland, the validation results show that it is possible to use simulation to predict the near-future state, and that a simulation based model is more accurate than other simpler prediction methods. However, the results also indicate that it is extremely difficult to accurately predict the future EMS system state, due to the highly stochastic processes that affect the ambulance movements. Thus, for future research, additional effort should be invested into the spatial and temporal prediction of ambulance calls. Another possibility is to try to divide the ambulances into groups, depending on the uncertainty in their future status and location, and use different models, and different ways of presenting the results to the ambulance dispatchers, for each group of ambulances.

## ACKNOWLEDGEMENTS

This research was partly funded by SOS Alarm Sweden AB and Lindholmen Science Park through TUCAP (http://tucap.lindholmen.se/).

## REFERENCES

Aboueljinane, L., E. Sahin, and Z. Jemai. 2013. “A Review on Simulation Models applied to Emergency Medical Service Operations.” *Computers & Industrial Engineering* 66(4):734-750.

Andersson, T. and P. Värbrand. 2007. “Decision Support Tools for Ambulance Dispatch and Relocation”. *Journal of the Operational Research Society* 58(2):195–201.

Aringhieri, R., M. E. Bruni, S. Khodaparasti, and J. T. van Essen. 2017. “Emergency Medical Services and Beyond: Addressing New Challenges through a Wide Literature Review”. *Computers & Operations Research* 78:349-368.

Bandara, D. M., L. Mayorga, and L. A. McLay. 2014. “Priority Dispatching Strategies for EMS Systems”. *Journal of the Operational Research Society* 65(4):572-587.

Bélanger, V., A. Ruiz, and P. Soriano. 2018. ”Recent Optimization Models and Trends in Location, Relocation, and Dispatching of Emergency Medical Vehicles. *European Journal of Operational Research.* In press. https://doi.org/10.1016/j.ejor.2018.02.055, accessed 23.07.2018.

Enayati, S., M. E. Mayorga, H. K. Rajagopalan, and C. Saydam. 2017. "Real-time Ambulance Redeployment Approach to Improve Service Coverage with Fair and Restricted Workload for EMS Providers". *Omega* 79:67-80.

Goldberg, J. B. 2004. "Operations Research Models for the Deployment of Emergency Services Vehicles". *EMS Management Journal* 1(1):20-39.

Henderson, S. G. and A. J. Mason. 2004. "Ambulance Service Planning - Simulation and Data Visualization". In *Handbook of Operations Research and Health Care Method and Applications, vol 70,* edited by F. Sainfort et al., 77-102. US: Springer.

Jagtenberg, C. J., S. Bhulai, and R. D. van der Mei. 2015. "An Efficient Heuristic for Real-time Ambulance Redeployment". *Operations Research for Health Care* 4:27-35.

Lam, S. S. W., C. B . L. Ng, F. N. H. L. Nguyen, Y. Y. Ng, and M. E. H. Ong. 2017. "Simulation-based Decision Support Framework for Dynamic Ambulance Redeployment in Singapore". *International Journal of Medical Informatics* 106:37-47.

Larson, R. C. 1974. "A Hypercube Queuing Model for Facility Location and Redistricting in Urban Emergency Services". *Computers & Operations Research* 1: 67-95.

Mason, A. J. 2013. "Simulation and Real-time Optimized Relocation for Improving Ambulance Operations". In *Handbook of Healthcare Operations Management: Methods and Applications,* edited by B. T. Denton. International Series in Operations Research & Management Science 184:289-317.

Maxwell, M. S., M. Restrepo, S. G. Henderson, and H. Topaloglu. 2010. "Approximate Dynamic Programming for Ambulance Redeployment". *INFORMS Journal on Computing* 22(2):266-281.

Nguyen, T.N.H. 2015. "Quantitative Analysis of Ambulance Location-allocation and Ambulance State Prediction". Licentiate Thesis No. 1699. Linköping Studies in Science and Technology, Linköping University.

Reuter-Oppermann, M.., P. L. van den Berg, and J. L. Vile. 2017. "Logistics for Emergency Medical Service Systems". *Health Systems* 6(3):187–208.

Schmid, V. 2012. "Solving the Dynamic Ambulance Relocation and Dispatching Problem using Approximate Dynamic Programming". *European Journal of Operational Research* 219(3):611-621.

Takeda, R., J. Widmer, and R. Morabito. 2007. "Analysis of Ambulance Decentralization in an Urban Emergency Medical Service using the Hypercube Queueing Model". *Computers & Operations Research* 34(3):727–741.

van Barneveld, T. C., S. Bhulai, and R. D. van der Mei. 2016. "The Effect of Ambulance Relocations on the Performance of Ambulance Service Providers". *European Journal of Operational Research* 252(1): 257-269.

Zhang, L., A. Mason, and A. Philpott. 2010. "The Optimisation of a Single Ambulance Moveup". Report No. 682, Faculty of Engineering, University of Auckland, Australia.

## AUTHOR BIOGRAPHIES

**TOBIAS ANDERSSON GRANBERG** is Associate Professor in Quantitative Logistics at the Div. Communication and Transport Systems, at Linköping University. His research interests are in transportation planning in general, and emergency response planning and air transportation in particular. Tobias is one of the main founders of CARER (Centre for advanced research in emergency response), where he is a board member and assistant director. His email adress is tobias.andersson.granberg@liu.se.

**HIEN T.N. NGUYEN** has a Licentiate in Infrainformatics from Linköping University. Her research is focused on resource management within emergency medical services, particularly using simulation modeling. Her email address is ng.ngoc.hien@gmail.com.