Analysis of Ambulance Location Models Using Discrete Event Simulation

Pascal Lutter, Dirk Degel, Lara Wiesche and Brigitte Werners

Abstract The quality of a rescue service system is typically evaluated ex post by the proportion of emergencies reached within a predefined response time threshold. Optimization models in literature consider different variants of demand area coverage or busy fractions and reliability levels as a proxy for Emergency Medical Service quality. But no comparisons of the mentioned models with respect to their real-world performance are found in literature. In this paper, the influence of these different model formulations on real-world outcome measures is analyzed by means of a detailed discrete event simulation study.

1 Introduction

Rescue and Emergency Medical Services (EMS) are an important part of public health care. The quality of a rescue service system is typically evaluated ex post by the proportion of emergencies reached within a predefined response time threshold. Coverage is one of the most accepted a priori quality criteria in EMS literature [1]. Since 1971 [9], different covering models and various extensions of these models are used to support ambulance location planning. The main challenge in ambulance location planning is to provide an adequate service level with respect to accessibility of an emergency within a predefined response time threshold and availability of ambulances [4]. Optimization models in literature consider different variants of demand area coverage, such as single coverage [2], double coverage [6] and empirically required coverage [5]. Other models use busy fractions [3] and reliability levels [7] as a proxy criterion for EMS quality. All those models support the decision maker on the strategic and tactical level of ambulance location planning, but differ regarding the specification of the objective functions as well as concerning input parameters and model assumptions. To the best of our knowledge, no systematic comparisons

P. Lutter (✉) · D. Degel · L. Wiesche · B. Werners
Faculty of Management and Economics, Chair of Operations
Research and Accounting,
Ruhr University Bochum, Bochum, Germany
e-mail: pascal.lutter@rub.de

M. Lübbecke et al. (eds.), *Operations Research Proceedings 2014*,
Operations Research Proceedings, DOI 10.1007/978-3-319-28697-6_53

of different ambulance location models exist in literature. The aim of this paper is to provide a comparison of the mentioned models concerning their suitability for decision support in strategic and tactical ambulance location planning. A discrete event simulation is used to systematically evaluate the resulting solutions of each covering concept. It is analyzed, which of those covering concepts provides the best proxy criterion for the real world performance measure. The remainder of the paper is structured as follows: First a brief overview of the selected ambulance location models is given. Technical details of the discrete event simulation are described and results of a real world case study are presented afterwards.

2 Ambulance Location Models

In this paper, daytime-dependent extensions of five well known models for ambulance location are considered: The (1) *Maximal Covering Location Problem* (MCLP) [2], the (2) *Double Standard Model* (DSM) [6], the (3) *Maximum Expected Covering Location Problem* (MEXCLP) [3], the (4) *Maximum Availability Location Problem* (MALP I/II) [7], and the (5) *Empirically Required Coverage Problem* (ERCP) [5]. To compare these models, a consistent constraint set is used and model assumptions are briefly summarized. For additional descriptions of these models see e.g. [1]. The aim of these models is to maximize the total demand served within a legal response time threshold of r minutes, given a limited number of p_t ambulances in period t. i indicates the planning squares or *demand nodes* ($i \in \mathscr{I}$), while d_{it} denotes the demand of node i in period $t \in \mathscr{T}$. To be able to serve an emergency at demand node i, at least one ambulance has to be available within the response time threshold r, e.g. positioned at node $j \in \mathscr{N}_{it} := \{j \in \mathscr{J} \mid \text{dist}_{ijt} \leq r\}$, where dist_{ijt} describes the response time between node i and node j in period t. The integer decision variable $y_{jt} \in \mathbb{N}_0$ indicates the number of ambulances positioned at node j in period t, and the binary decision variable x_{it}^k is equal to 1 if demand node i is covered k times in period t. With the preceding notation, generic *covering constraints* are given by

$$\sum_{j \in \mathscr{N}_{it}} y_{jt} \geq \sum_{k=1}^{p_t} x_{it}^k \qquad \forall i \in \mathscr{I}, \forall t \in \mathscr{T}. \tag{1}$$

Setting the right hand side of constraints (1) equal to 1 ensures that each demand node i can be reached within the response time threshold at least once if an ambulance is available. This single coverage may become inadequate when several emergencies occur at the same time and the assigned ambulances become busy. To hedge against parallel operations resulting in unavailability of ambulances, the mentioned models use different concepts and objective functions. All models ensure a sufficient number of ambulances located in $\mathscr{N}_{it}$ to serve each demand node i. Table 1 compares the covering constraints, the objective functions, and the assumptions of a priori information of the models. In addition to covering constraints, further constraints are used to ensure correct relocations and to restrict the number of

Table 1 Comparison of model constraints, objectives, and assumptions about required information

Model	Covering constraint	Objective	Required information
MCLP	$\sum_{j\in\mathcal{N}_{it}} y_{jt} \geq x_{it}^1$	$\max \sum_{i\in\mathcal{I}} \sum_{t\in\mathcal{T}} d_{it}x_{it}^1$	d_{it}
DSM	$\sum_{j\in\mathcal{N}_{it}} y_{jt} \geq x_{it}^1 + x_{it}^2$	$\max \sum_{i\in\mathcal{I}} \sum_{t\in\mathcal{T}} d_{it}x_{it}^2$	d_{it}
MEXCLP	$\sum_{j\in\mathcal{N}_{it}} y_{jt} \geq \sum_{k=1}^{p_t} x_{it}^k$	$\max \sum_{i\in\mathcal{I}} \sum_{t\in\mathcal{T}} \sum_{k=1}^{p_t} d_{it}(1-q_t)q_t^{k-1}x_{it}^k$	d_{it}, q_t
MALP I	$\sum_{j\in\mathcal{N}_{it}} y_{jt} \geq \sum_{k=1}^{p_t} x_{it}^k$	$\max \sum_{i\in\mathcal{I}} \sum_{t\in\mathcal{T}} d_{it}x_{it}^{K_t}$	d_{it}, α, q_t, K_t
ERCP	$\sum_{j\in\mathcal{N}_{it}} y_{jt} \geq \sum_{k=1}^{p_t} x_{it}^k$	$\max \sum_{i\in\mathcal{I}} \sum_{t\in\mathcal{T}} d_{it}x_{it}^{K_{\ell_i t}}$	$d_{it}, K_{\ell_i t}$

$K_t := \lceil \ln(1-\alpha)/\ln(q_t) \rceil$, $K_{\ell_i t}$ empirically required degree of coverage (see explanation below)

ambulances in use (see e.g. [5]). In the MCLP and the DSM a uniform single, respectively double coverage is maximized. Few information is needed, but the unavailability of ambulances due to parallel operations is ignored in the MCLP or simplified in the DSM by using a time and spatial fixed backup (double) coverage. In the MEXCLP it is assumed that each ambulance has the probability q, called *busy fraction*, of being unavailable and the expected covered demand is maximized. In the earliest version of the MEXCLP [3], it is assumed that each ambulance has the same probability of being busy. In order to compare the models on the basis of the same criterion, a time-dependent busy fraction q_t is used in the following analysis. In MALP I and MALP II [7], a chance constrained program is used to maximize the demand covered at least with a given probability α. The minimum number of ambulances required to serve demand node i with a reliability level of α in period t is determined by

$$1 - q_t^{\sum_{j\in\mathcal{N}_{it}} y_{jt}} \geq \alpha, \tag{2}$$

which can be linearized as $\sum_{j\in\mathcal{N}_{it}} y_{jt} \geq \lceil \ln(1-\alpha)/\ln(q_t) \rceil =: K_t$ (with system unique busy fraction q_t in MALP I). In MALP II, the assumption of identical busy fractions is relaxed and the busy fractions q_{it} are calculated for each demand node i and period t. The problem of using demand node specific busy fractions q_{it} is that these values depend on the output of the model and are unknown a priori [1]. To overcome this difficulty, a more direct and data-driven way to determine the required coverage is used in the ERCP [5]. The empirical distribution function representing the number of parallel EMS operations per time unit and district is calculated. The 95 % quantile of the stochastic demand per district l and time period t is determined empirically in order to derive the required degree of coverage $K_{\ell_i t}$ and thus the necessary number of ambulances. This assures that there is a sufficient number of ambulances to cover all parallel operations in at least 95 % of all cases. To compare the mentioned models, all relevant model parameters, like busy fractions, reliability levels and the empirically required coverage levels are calculated using an identical data base which relies on the same spatial and time-dependent partitioning.

3 Discrete Event Simulation of EMS Systems

All previously described models consider different variants of demand coverage as a proxy criterion for EMS quality, defined as the proportion of calls served within the legal response time threshold. This real world outcome measure is mainly influenced by the positioning of ambulances implied by different objective functions and covering constraints. The solution quality, e.g. the quality of an EMS system can only be evaluated ex post. Discrete event simulation represents a common approach to analyze complex and dynamically changing environments like EMS systems. In this paper, a simulation approach is applied to compare the performance of the aforementioned models regarding the real world outcome measure. In the following, the main components of the simulation are described. The data generation process for the discrete event simulation consists of two main modules:

1. Generation of random events: A whole weekday is subdivided into 24 time intervals $t \in \{0, \dots, 23\}$ with a length of $\Delta = 1$ h. For a given demand node i and a time interval $[t, t + \Delta)$, in the following indicated by t, the number of emergencies occurring within t can be approximated by a Poisson distribution P_λ with parameter λ. The average number of emergency calls per time interval t at a given weekday D is $P_{\lambda_{it}^D}$ with $\lambda_{it}^D := (\alpha_D/365) \cdot \sum_{\ell=1}^{365} d_{it}^\ell$. The parameter λ_{it}^D is used as an estimator for the parameter of the Poisson distribution, where d_{it}^ℓ denotes the historical number of emergencies occurring in period t in demand node i at day ℓ. The scaling factor α_D is determined empirically and serves as a correction term for introducing day-related seasonality. This is necessary since the total demand fluctuates within different weekdays. For each t and i, the quantity of emergencies d_{it} is sampled from previously specified Poisson distributions. Then, the emergencies are distributed according to the realization of a uniform random variable within the time interval t.
2. Travel time generation: The travel time is not constant for different time intervals t of the day, cf. [8]. Typically, higher traveling speeds are achieved in the evening, while lower speeds are observed around noon and during rush hours. To incorporate realistic driving speeds, a time-dependent random variable $v_t \sim N(\mu_t, \sigma_t)$ is used, where μ_t and σ_t are determined empirically. For each generated emergency, the travel time is sampled from $N(\mu_t, \sigma_t)$ and stored in the corresponding variable.

During the simulation, an emergency event is characterized by the time of occurrence, the associated demand node, the traveling speed of the associated ambulance and the emergency duration. The duration of each operation is sampled from the empirical distribution function. The simulation process works as follows: An ambulance is characterized by the assigned EMS station and an availability indicator. An ambulance is available, if it is currently not serving an emergency. For each emergency occurring, the selection of ambulances is performed by a predefined nearest-distance strategy: For a given emergency position i, all ambulance locations are sorted by increasing distances to i. Note, that the traveling distance depends on the location of ambulances which are an outcome of the tested optimization model. If there is an ambulance available at the nearest station, this vehicle is selected. Otherwise, the next station in the list is checked. The process repeats until an available ambulance has been found or all stations are checked. If no ambulance is available, a queue of unfulfilled requests is being built. Whenever an ambulance is assigned to serve an emergency, the travel time is generated and the vehicle is blocked for the duration time of the operation. The simulation process terminates after serving all emergencies. Finally, dividing the overall number of emergencies served on time by all emergencies occurring gives the desired real world quality measure.

4 Case Study and Results

A real world case study for evaluating model performances is conducted by specifying all required model parameters (demand, busy fractions and empirically required coverage) on the basis of a data set from a German city containing more than 20,000 operations per year. In all models the number of ambulances in time period t is given by the parameters p_t and all demand points are considered as potential ambulance locations. The average emergency demand over 1 year is visualized in the first picture of Fig. 1. The first objective is to maximize the model specific objective function. The second objective is to cover a maximal number of demand areas at least once. The third objective aims at minimizing the number of vehicle locations. To hedge against dual degeneracy in location models, a lexicographic approach is applied. The coverage induced by the solutions of the models are visualized in Fig. 1. Demand nodes are colored from light to dark gray and visualize the number of zero (light gray)

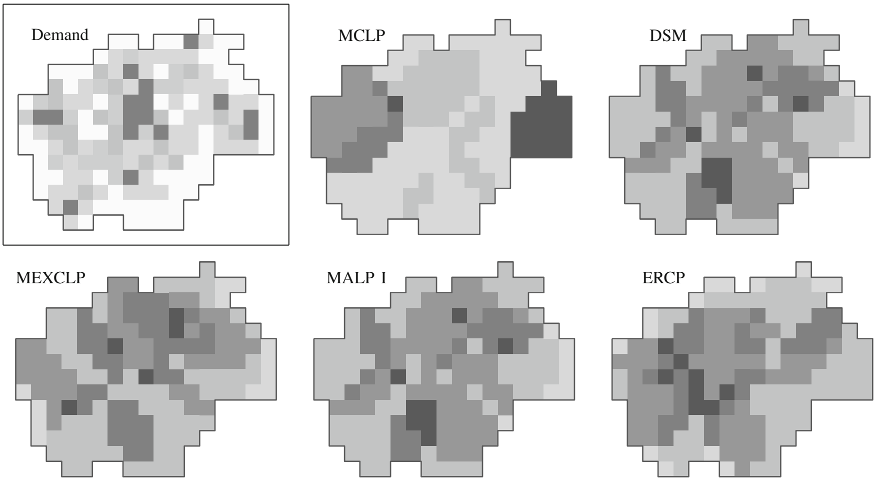

Fig. 1 Emergency demand and degree of coverage induced by the solutions of different models

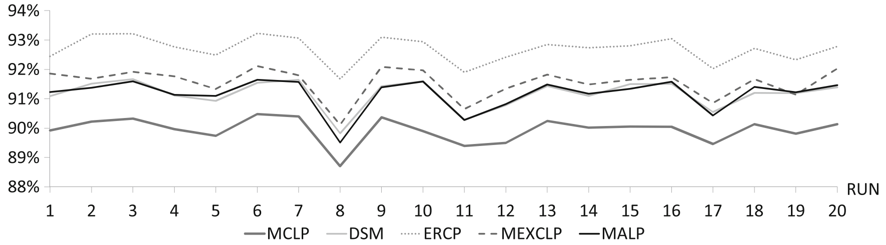

Fig. 2 Proportion of calls served within the time threshold of 8 min during 1 year (20 simulation runs)

to sixfold (dark gray) coverage. Results, i.e., the proportion of calls served within the legal response time threshold of 8 min for 20 simulation runs for each model solution are shown in Fig. 2. Data driven covering models like MEXCLP, MALP and ERCP outperform fixed covering models (MCLP, DSM) with respect to the real world EMS performance measure. Fixed covering models provide inadequate coverage of areas with high, resp. low, number of parallel operations due to disregarding the availability of ambulances. Instead, data driven approaches locate ambulances as needed by considering demand volume as well as criteria for ambulance unavailability.

5 Conclusion and Outlook

In this paper a discrete event simulation study is conducted to evaluate different ambulance location models. Based on the simulation study exemplary results of different coverage concepts concerning their influence on real world performance measures are shown. All analyzed concepts differ concerning the input parameters and model assumptions. Exemplary results suggest that models requiring detailed information (for example the MEXCLP and the ERCP) perform better than models ignoring these information. In the next step, studies are extended systematically to different typical city and demand structures.

Acknowledgments This research is financially supported by *Stiftung Zukunft NRW*. The authors thank staff members of *Feuerwehr und Rettungsdienst Bochum* for providing detailed insights.

References

1. Brotcorne, L., Laporte, G., Semet, F.: Ambulance location and relocation models. Eur. J. Oper. Res. **147**(3), 451–463 (2003)
2. Church, R., ReVelle, C.: The maximal covering location problem. Pap. Reg. Sci. **32**(1), 101–118 (1974)

3. Daskin, M.: A maximum expected covering location model: formulation, properties and heuristic solution. Transp. Sci. **17**(1), 48–70 (1983)
4. Daskin, M., Dean, L.: Location of health care facilities. In: Operations Research and Health Care, pp. 43–76. Springer (2004)
5. Degel, D., Wiesche, L., Rachuba, S., Werners, B.: Time-dependent ambulance allocation considering data-driven empirically required coverage. Health Care Manage. Sci. 1–15 (2014)
6. Gendreau, M., Laporte, G., Semet, F.: Solving an ambulance location model by tabu search. Location Sci. **5**(2), 75–88 (1997)
7. ReVelle, C., Hogan, K.: The maximum availability location problem. Transp. Sci. **23**(3), 192–200 (1989)
8. Schmid, V., Doerner, K.: Ambulance location and relocation problems with time-dependent travel times. Eur. J. Oper. Res. **207**(3), 1293–1303 (2010)
9. Toregas, C., Swain, R., ReVelle, C., Bergman, L.: The location of emergency service facilities. Oper. Res. **19**(6), 1363–1373 (1971)