Patient-Aware Logistics in Emergency Medical Services

Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften
(Dr.-Ing.)
von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)
genehmigte

Dissertation

von

M.Sc. Sven Watzinger

Tag der mündlichen Prüfung:	12.06.2026
Referent:	Prof. Dr. Stefan Nickel
Korreferent:	Assoc. Prof. Dr. Derya Demirtas

Acknowledgements

This thesis is the result of a journey that numerous individuals enabled and supported. I am profoundly grateful for the opportunity of creating this thesis, and for the support I received along the way. I especially want to thank:

Prof. Dr. Stefan Nickel, for entrusting me with the project alongside which this thesis was created, for his continuous drive and motivation throughout supervising this thesis, and for sparking my interest in applying Operations Research in healthcare during my studies.

Assoc. Prof. Dr. Derya Demirtas, for her interest in my work, and for refereeing this thesis.

The team at the SQR-BW, for their continuous work in building, maintaining, and improving the treasure of real-world data, without which this thesis would not have been possible.

My colleagues at the group of discrete optimization and logistics, for making the office an enjoyable place to work, and lunch about much more than the intake of calories. My thanks go especially to Hannah Bakker for her guidance and detailed feedback during the creation and writing of this thesis. My first steps into the realm of EMS logistics were greatly supported by David Olave-Rojas, who helped me find my footing and with whom it was a great joy to discuss and refine ideas.

My friends, for their support, their patience when my energy went into this thesis, and for being there to celebrate.

My parents, Susanne and Jörg, and my brother Eric, for always being there for me, for fostering my curiosity, and for supporting me unconditionally throughout my educational path.

My partner Kirsten, for her incredible patience, her support, and her belief in me during all the highs and lows of this journey.

Sven Watzinger

Karlsruhe, August 2026

Abstract

Emergency Medical Services (EMS) are an integral part of healthcare systems. Their purpose is to respond to medical emergencies, treat patients on scene, and, if necessary, transport them to medical facilities for further care. This thesis studies EMS systems from a logistic perspective. Predictive and prescriptive models from the field of Operations Research can support making logistic decisions in EMS. Such quantitative models require quantitative evaluation criteria for logistic decisions. Evaluation criteria in EMS systems are typically based on response time targets. There is, however, an ongoing debate about how well response time targets can quantify the impact of logistic decisions on the quality of EMS, and specifically on patient outcomes. This debate leads to the fundamental question of this thesis: How can the impact on patient outcomes be considered when evaluating EMS systems from a logistic perspective, and what does this imply for making logistic decisions? This thesis proposes a novel approach to evaluating EMS logistics through utility functions defined for patient categories. It subsequently explores what this approach implies for the operational ambulance dispatching and the strategic ambulance station location and sizing decisions. We find that evaluating EMS systems from a logistic perspective should and can primarily be based on how logistic decisions impact patients. Patients have differing needs, and considering this heterogeneity in logistic decisions can improve patient utilities and reduce required resources.

We structure our research into three topics: Quality evaluation, system behavior, and logistic decisions. Quality evaluation addresses the question of how to evaluate EMS systems from a logistic perspective. System behavior refers to the question of how to assess the impact of logistic decisions on EMS system behavior. The topic of logistic decisions explores how to model logistic decisions in EMS systems. The EMS system in Baden-Württemberg, Germany, serves as the application case for this thesis, enabling us to conduct computational experiments with real-world data.

We approach the topic quality evaluation through a literature review on quality dimensions and quality indicators for EMS logistics. The review highlights the gap in quantifying how logistic decisions impact patient outcomes. We subsequently conduct an expert consensus process to define utility functions for patient categories. Utility functions for heterogeneous patients improve the quantification of how logistic decisions impact patient outcomes compared to evaluating response time targets.

We employ a hybrid simulation approach combining agent-based and discrete-event modeling paradigms to study the topic system behavior. Simulation enables modeling how logistic decisions impact EMS system behavior, and facilitates evaluating logistic decisions before implementing them in practice.

Several logistic decisions need to be made in EMS. We investigate what our findings in the topic quality evaluation imply for the operational ambulance dispatching and the strategic ambulance station location and sizing decisions. We conduct a scoping review of ambulance dispatching strategies that reveals a variety of modeling approaches with differing attributes. A simulation-based comparison of strategies with suitable attributes for our patient categories and application case enables investigating the effects of considering heterogeneous patients. The experiments indicate that differentiated dispatching strategies can significantly reduce driving times to prioritized patients, but appear sensitive to how well prioritized patients can be detected as such during the dispatching decisions. We model the ambulance station location and sizing decision using mixed-integer linear programming. Considering heterogeneous patients affects the validity of existing modeling assumptions beyond integrating patient categories. We formulate three mixed-integer linear programs to relax limiting assumptions and to consider heterogeneous patients. Computational experiments demonstrate that our modeling approach can avoid overcapacities induced by previous modeling approaches and provide insight into the instance-dependent effects of considering heterogeneous patients on required resources and system behavior.

Interdisciplinary approaches facilitate quantifying the connection between medicine and logistics. Patients in EMS have differing needs, and the modeling of logistic decisions should take these needs into account explicitly.

Contents

Acronyms

EMS Emergency Medical Services

EPV emergency physician vehicle

OHCA out-of-hospital sudden cardiac arrest

RTT response time target

SQR-BW Department for Quality Assurance in Emergency Medical Services in Baden-Württemberg (dt.: Stelle zur trägerübergreifenden Qualitätssicherung im Rettungsdienst in Baden-Württemberg)

BPLSCP Binomial Probabilistic Location Set Covering Problem

BRLSCP Binomial Reliability Location Set Covering Problem

MEPLP-HR Maximum Expected Performance Location Problem for Heterogeneous Regions

MESLMHP Maximal Expected Survival Location Model for Heterogeneous Patients

PRLSCP Poisson Reliability Location Set Covering Problem

QPLSCP Queuing Probabilistic Location Set Covering Problem

QRLSCP Queuing Reliability Location Set Covering Problem

QRUBUL Queuing Reliability Upper-Bound Unavailability Location Problem

QRUBUL-HR Queuing Reliability Upper-Bound Unavailability Location Problem for Heterogeneous Regions

QRUBUL-HR-HP Queuing Reliability Upper-Bound Unavailability Location Problem for Heterogeneous Regions and Heterogeneous Patients

UBUL Upper-Bound Unavailability Location Model

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PRISMA-ScR Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews

Chapter 1

Introduction

Wherever the art of Medicine is loved,
there is also a love of Humanity.

Hippocrates

1.1 Motivation

Compassion is a driving force behind the human need to help and alleviate suffering (Goetz et al., 2010). Societies have institutionalized helping others in medical matters through healthcare systems. EMS are an integral part of healthcare systems. Their purpose is to respond to medical emergencies outside medical facilities, treat patients on scene, and, if necessary, transport them to facilities for further care.

The EMS system in Germany has seen an increase in both emergency calls and costs in recent years. Sieber et al. (2020) found that call volumes increase yearly by approximately 5%, and the expenses of statutory health insurers for transporting patients, including EMS costs, have more than doubled from 2014 to 2024 (Federal Ministry of Health, 2025). A continuing staff shortage puts further pressure on the EMS (Habicht & Hahnen, 2024). EMS systems can be studied from multiple perspectives, such as: medical, financial, organizational, legal, or logistic. We study EMS from a logistic perspective in this thesis. The fundamental task of logistics is to ensure that the right resource is available at the right place and the right time. Logistics plays a complementary role to medicine in healthcare systems. Medicine provides the capabilities to heal and treat patients, and logistics makes these capabilities accessible to patients. The role of logistics is especially pronounced in EMS systems due to the correlation between time to treatment and outcome for some emergencies, such as sudden cardiac arrest (Bürger et al., 2018).

EMS systems employ several resources, such as ambulances or helicopters. Logistics in EMS encompasses several logistic decisions, for instance, on the locations of resource stations, on the number of resources, or on coordinating resources during operations. These decisions influence how the EMS system behaves in terms of logistic and operational processes. Evaluating the system behavior can, in turn, inform how logistic decisions should be made to result in a desirable system behavior. Fig. 1.1 depicts this approach to designing EMS systems from a logistic perspective.

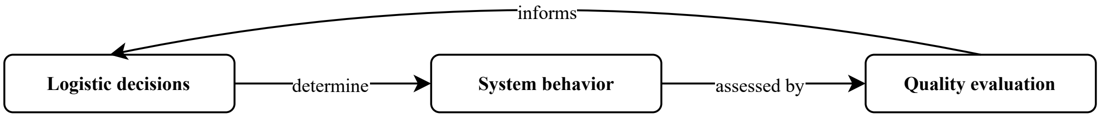


Figure 1.1: Designing EMS systems from a logistic perspective

Operations Research can support designing EMS systems with predictive and prescriptive analytics through mathematical models (Aringhieri et al., 2017). Prescriptive models recommend how to make a decision. Predictive models forecast how a decision impacts the system behavior. Evaluating the modeled system behavior enables iterating through the decision-behavior-evaluation loop before implementing a decision in the actual system. The role of Operations Research is therefore to provide quantitative decision support for decision makers in EMS systems.

Quantitative models require quantitative metrics. Response time targets are the basis for the most commonly used quality indicators in EMS systems (Pons et al., 2005; Reza Hosseini et al., 2017). The idea of the response time is to capture the time from the onset of a medical emergency until the first EMS resource arrives at the patient. A response time target sets an upper threshold for the response time. Existing modeling approaches for logistic decisions in EMS often reflect this by basing their objective on the proportion of calls reached within such a threshold (Bélanger et al., 2019). Al-Shaqsi (2010) and Murphy et al. (2016), among others, have challenged the use of response time targets as a quality indicator due to their binary nature and the missing correlation to patient outcomes for the majority of emergencies. Current metrics for EMS logistics, therefore, do not fully capture the patient perspective. This leads to the following fundamental question that we aim to address in this thesis:

How can the impact on patient outcomes be considered when evaluating EMS systems from a logistic perspective, and what does this imply for making logistic decisions?

1.2 Research Questions and Scope

Addressing this fundamental question touches upon all three parts of designing EMS systems from a logistic perspective depicted in Fig. 1.1. While the approach constitutes a loop, a causal chain exists from making *logistic decisions* to *system behavior* to *quality evaluation*. We propose traversing this chain in reverse for our research. Fig. 1.2 lists the research questions and the corresponding contributions developed in this thesis.

Quality evaluation

RQ1: How can the quality of EMS logistics be quantified?

Contributions:
- Development of a framework for quality indicators in EMS logistics
- Identification of a research gap in quantifying the effectiveness of EMS logistics
- Development of patient categories and utility functions for heterogeneous patients through an expert consensus process

System behavior

RQ2: How can the effects of logistic decisions on the quality of EMS be predicted?

Contributions:
- A generic concept model for EMS systems modeling emergency rescue and patient transport operations with an arbitrary number of involved resources
- Coordination concepts of task leaders and task backlogs, enabling coordination of multiple resources and integration of scheduled and ad hoc tasks
- An executable simulation model in AnyLogic combining agent-based and discrete-event modeling paradigms
- Validation of the simulation model for ground-based emergency rescue operations in Baden-Württemberg, Germany, with real-world data

Logistic decisions

RQ3: How can the ambulance dispatching decision consider heterogeneous patients?

Contributions:
- Scoping review of ambulance dispatching strategies
- Selection of strategies suitable for heterogeneous patients in the EMS in Baden-Württemberg
- Comparison of suitable strategies in simulation experiments with real-world data

RQ4: How can the ambulance station location and sizing decision consider heterogeneous patients?

Contributions:
- Formulation of a capacitated set covering model with system-optimal demand assignment for the ambulance station location and sizing decision
- Comparison of three existing models and the newly developed model using simulation with real-world data
- Extension of the base model to consider location-dependent service times
- Extension of the base model to consider heterogeneous patients with differentiated coverage thresholds and service levels

Figure 1.2: Research questions and contributions of the thesis.

This thesis arose from the research project EVRALOG-BW[1] in cooperation with the EMS in Baden-Württemberg, Germany. We use historical call data from the EMS system in Baden-Württemberg in several numeric experiments and analyses throughout this thesis, and the results presented in this thesis should be interpreted accordingly. The methodological approaches, however, aim to be transferable to other EMS systems.

[1]Funded by the Ministry of the Interior, Digitalisation and Local Government in Baden-Württemberg, Germany. Further information: https://healthcarelab.ksri.kit.edu/english/projects.php

1.3 Outline of the Thesis

The thesis is structured into five parts, framed by this introduction and a conclusion, as depicted in Fig. 1.3.

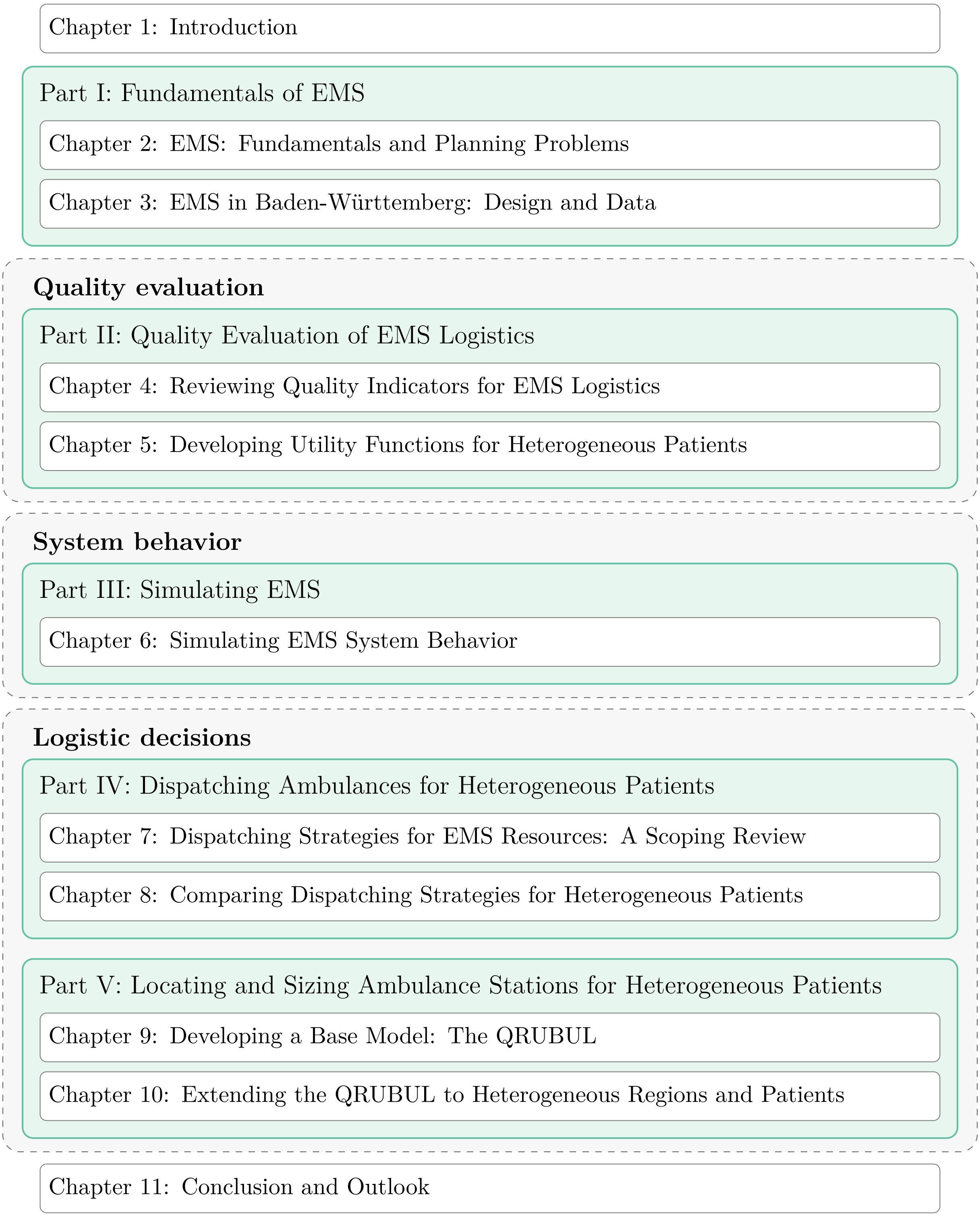


Figure 1.3: Outline of the thesis

Part I lays the foundation for the thesis. Chapter 2 introduces the fundamental tasks, elements, and planning problems of EMS systems. Chapter 3 describes the specific design, legislation, and available data of the EMS system in Baden-Württemberg, Germany, which serves as the application case for this thesis.

Part II addresses the first research question on quality evaluation. Chapter 4 reviews existing quality indicators and develops a framework for quality indicators in EMS logistics. Based on an identified research gap, Chapter 5 develops utility functions for heterogeneous patients using an expert consensus process to quantify the effectiveness of EMS logistics.

Part III focuses on system behavior. Chapter 6 presents the development and validation of a hybrid agent-based and discrete-event simulation model implemented in AnyLogic. The model is validated using real-world data from Baden-Württemberg.

Part IV investigates the operational decision of ambulance dispatching. Chapter 7 conducts a scoping review of dispatching strategies. Chapter 8 compares selected dispatching strategies for heterogeneous patients using simulation experiments to assess their impact on patient utilities and driving times.

Part V examines the strategic decision of ambulance station location and sizing. Chapter 9 introduces the QRUBUL, a capacitated set covering model based on a system-optimal demand assignment. Chapter 10 extends the model to consider location-dependent service times and heterogeneous patients. Both chapters evaluate the models through computational experiments based on real-world data.

Finally, Chapter 11 summarizes the findings of the thesis and provides an outlook on future research directions.

Part I

Fundamentals of EMS Logistics

Chapter 2

EMS: Fundamentals and Planning Problems

This chapter introduces EMS to provide the foundations of the application domain of this thesis. We start by presenting the fundamental tasks and elements of EMS systems in Section 2.1. The scope of this thesis encompasses the task of emergency rescue, and we describe the core process and central time intervals of emergency rescue operations in Section 2.2. Several logistic decisions need to be taken in EMS systems, and we provide an overview of them in Section 2.3. We conclude the chapter with an introduction to planning criteria in EMS in Section 2.4.

2.1 Tasks and Elements of EMS

EMS systems generally perform two major tasks: *emergency rescue* and *patient transportation*. *Emergency rescue* encompasses the prehospital care for patients with medical emergencies and, if necessary, the transport to a hospital. It is characterized by ad-hoc reactions to oftentimes time-critical situations. *Patient transportation*, on the other hand, refers to the non-emergency transport of patients. These transportation requests can generally be scheduled in advance, and the fulfillment of requested time windows plays a larger role than an immediate response to a transportation request. Transports can occur between medical facilities or between medical facilities and patients' homes. Kergosien et al. (2015) investigated using common resources for both tasks of emergency rescue and patient transportation from a logistic perspective, indicating that the two tasks are not necessarily independent.

The scope of this thesis encompasses the task of *emergency rescue* in EMS. Unless otherwise stated, we refer to this task when referencing EMS in the main chapters.

EMS systems are comprised of two major elements to perform these tasks: *Coordination centers* and *EMS resources*. *Coordination centers* are the point of contact for any incoming call. They are responsible for extracting the information, for instance, from an emergency call. Based on the information, they decide if and which resources are required. The coordination centers then dispatch and coordinate the required resources. They can additionally provide medical support via telephone, such as instructing first aid measures or resuscitation. There are two main roles in the coordination center: Call-takers answer calls and extract the relevant information, and dispatchers dispatch and coordinate the resources. Depending on the EMS system, these roles are fulfilled by one person or by separate persons (van Buuren et al., 2017). Fig. 2.1 schematically presents the interaction between calls, coordination centers, and EMS resources.

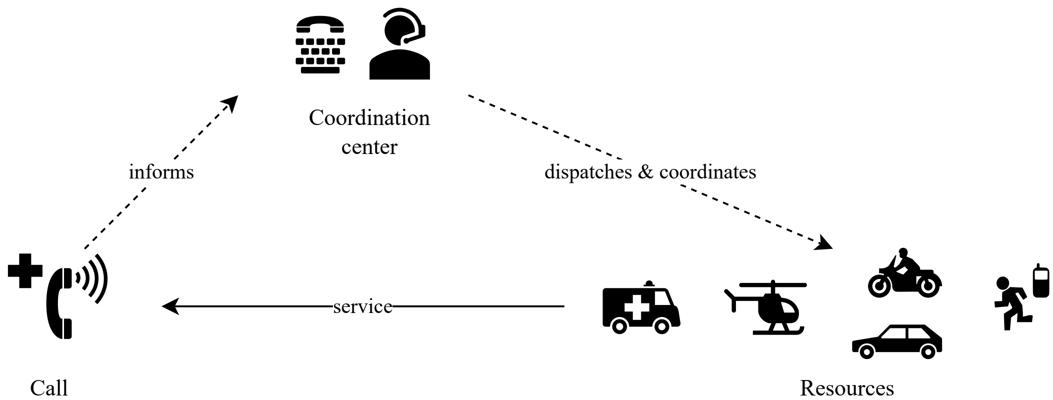


Figure 2.1: Fundamental elements of EMS systems.

The specific characteristics of the available *EMS resources* can vary between EMS systems, but they can be broadly classified as follows:

Ambulances are ground-based vehicles that have the capacity to transport patients. Oftentimes, multiple types of ambulances are present in an EMS system with either *separate* or *hierarchical* responsibilities. In some systems, such as the German one, there are different types of ambulances for patient transportation and emergency rescue, which are generally used separately for these tasks. In other systems, tiered ambulances, typically called Advanced Life Support (ALS) and Basic Life Support (BLS) ambulances, are used, which have a *hierarchical* capability (Yoon & Albert, 2020).

Helicopters are air-based vehicles and can have the capacity to transport a patient. The advantages of helicopters over ground-based vehicles are firstly their ability to reach remote locations which might either not be accessible by road at all or only with a significant delay (Jagtenberg et al., 2021). Secondly, helicopters are faster, especially over longer distances, which can reduce the transport times.

Rapid response vehicles describe vehicles used in emergency rescues to move personnel to the scene of the emergency, but without the capacity to transport patients. Their purpose is either to be able to respond faster to emergencies than ambulances or to transport additional personnel such as physicians to the emergency (Da Ros et al., 2024; Olave-Rojas & Nickel, 2021).

(Smartphone-based) first responders are primarily employed for patients with sudden out-of-hospital cardiac arrest and are dispatched if they can reach the location of the emergency faster than alternative resources (van den Berg et al., 2025). First responders can include firefighters, police officers, off-duty EMS personnel, and citizen responders. The use of smartphone-based systems to dispatch first-responders has gained increasing adoption in recent years (Metelmann et al., 2021).

2.2 The Core Process of EMS: The Rescue Chain

The rescue chain is the core process of an emergency rescue operation and contains several central time intervals. Fig. 2.2 depicts the rescue chain and its time intervals. The depicted process is a simplification of the possible process variants that occur in practice, and Part III dives deeper into the processes during emergency rescue operations.

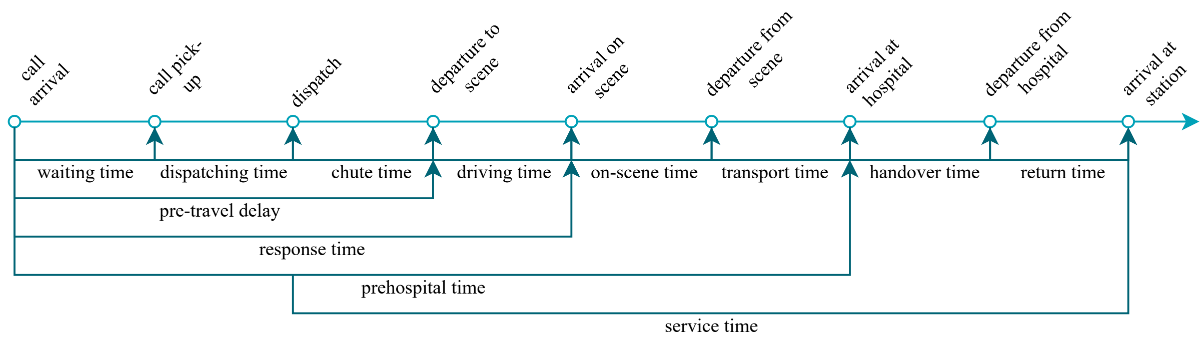


Figure 2.2: The rescue chain and its time intervals.

The rescue chain begins with the *call arrival* in a coordination center. Most commonly, a call is a telephone call on an emergency number, such as 112 or 911, depending on the region. Other types of calls include calls through emergency apps or eCalls from vehicles (Kubiak et al., 2024). The *call pick-up* occurs as soon as a call-taker is available in the coordination center. The call-taker extracts all relevant information from the call. If the call requires dispatching resources, the *dispatch* occurs, and an alarm is sent out to the resources. There can be a delay between receiving the alarm and the *departure to the scene* by the resources, for instance, if an ambulance crew is resting at the station and needs to get ready to start driving to the scene. After the *arrival on scene*, the EMS staff treats the patient on scene. If the patient requires transport to a facility for further care, most commonly a hospital, the *departure from the scene* marks the beginning of transporting the patient. If no transport is required, the EMS resources depart for their station. The transport ends with the *arrival at hospital*, where the patient is handed over, usually to the emergency department. The EMS resources start returning to their stations with the *departure from hospital*, and the rescue chain ends with the *arrival at station*.

Two time intervals along the rescue chain are especially important from a patient perspective: The *response time* and the *prehospital time*. The *response time* captures the time between the call arrival in the coordination center and the arrival of the first EMS resources on scene. The underlying idea of the response time is to capture a proxy for the time between the onset of the medical emergency and the initial treatment, and there are emergencies, such as sudden cardiac arrest, for which a correlation between the response time and the medical outcome exists (Bürger et al., 2018; Neukamm et al., 2011; Valenzuela et al., 1997). The *prehospital time* captures the time between the call arrival in the coordination center and the arrival at a hospital. This time interval is especially important for emergencies, such as strokes, for which an effective diagnosis and treatment are not possible on scene, but only in a hospital. In these cases, the prehospital time instead of the response time serves as a proxy for the time between the onset of the medical emergency and the (effective) treatment. We investigate the importance of time intervals for patients in Part II of this thesis.

The *service time* and the *pre-travel delay* primarily play a role from the logistic perspective on EMS systems. The *service time* refers to the time interval from dispatching EMS resources until they arrive back at their station. The interval captures the duration during which the resources are busy serving a call and can inform planning parameters such as the utilization or availability of resources. Depending on the EMS system, resources can also be dispatched to a new call while returning to their station or be redispatched from one call to another, but unless otherwise stated, the *service time* refers to the time interval from dispatch to arrival at the station in this thesis. The *pre-travel delay* captures the part of the response time that is not the driving time. This time interval is relevant since several logistic decisions primarily impact the driving times of EMS resources, while objectives are often based on the response time.

2.3 Logistic Decisions in EMS

The lengths of several intervals in the rescue chain, such as the waiting time, driving times, transport times, or return times, are mainly a result of logistic decisions taken in EMS. Logistic decisions additionally influence the cost a society has to bear for an EMS system by impacting the amount of resources. Logistic decisions in EMS have been studied in the field of Operations Research, and the following reviews provide comprehensive overviews of modeling approaches: Aringhieri et al. (2017), Bélanger et al. (2019), Gholami-Zanjani et al. (2018), Mukhopadhyay et al. (2022), Neira-Rodado et al. (2022), and Reuter-Oppermann et al. (2017). The decisions can be structured by their time horizon along the *strategic, tactical* and *operational* level, as shown in Fig. 2.3.

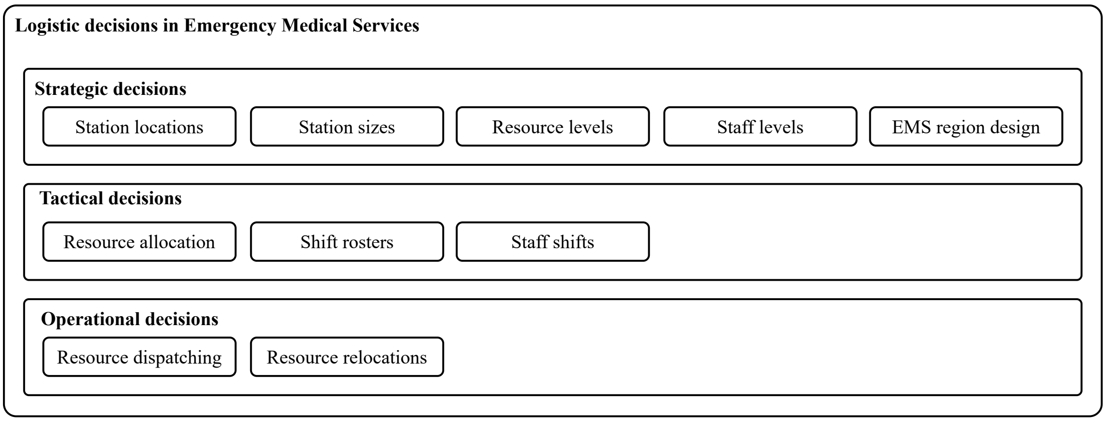


Figure 2.3: Logistic decisions in EMS.

Decisions on the *strategic* level are characterized by a long-term planning horizon of several years. The *station location* decision determines the permanent location of resource stations, for instance, ambulance or helicopter stations. The *station size* decision determines how many resources can be allocated to a station and needs to be taken when constructing a new station. We investigate the station location and sizing decisions in Part V of this thesis. The decision on *resource levels* determines the amount of resources of different types in an EMS system and should consider back-up capacities, for instance, for disaster situations, training, or resource maintenance. Resources are operated by EMS crews, deciding on *staff levels* closely related to the decision on resource levels. EMS systems are typically organized in EMS regions. An EMS region is a geographical area, and logistic decisions in EMS systems are often taken for a single EMS region. The *design of EMS regions*, therefore, influences every other logistic decision. In practice, EMS regions typically follow administrative boundaries.

The *tactical* level contains medium-term decisions, usually with a time horizon of up to one year. The *resource allocation* decision defines how many resources are allocated to which station, and might vary for the days of the week and the times of the day. *Shift rosters* define the shift patterns for the EMS crews. A 24/7 ambulance, for instance, could be crewed with three eight-hour shifts or two twelve-hour shifts. Assigning individual staff to specific shifts, both for EMS resources and in coordination centers, constitutes the decision on *staff shifts*.

Decisions taken daily during EMS operations fall into the *operational* level. The *resource dispatching* decision determines which resource to dispatch to a call. The decision can encompass the decision on which resource type to dispatch and which specific resource to dispatch. We investigate the dispatching decision in Part IV of this thesis. EMS resources are typically allocated to one station but do not necessarily remain at this station during operations. The *resource relocation* decision either actively relocates resources between stations or temporary locations, or determines which location a resource should move to after it finishes serving a call. The objective of relocations is generally to maintain a spatial distribution of resources with which the EMS system can respond quickly to calls.

2.4 Planning Criteria in EMS

Planning criteria determine the quantitative objectives that prescriptive models aim to maximize or minimize through the logistic decisions. Planning criteria should therefore ideally quantify the system's objective well. Response time targets (RTT) are the most commonly used planning criteria in EMS systems (Pons et al., 2005; Reza Hosseini et al., 2017). A RTT consists of a maximum response time, for instance 10 minutes, in conjunction with a percentage, for instance 95%, that determines the proportion of calls for which the maximum response time must not be exceeded. The focus on RTT as the central planning criterion for EMS has been challenged. The main argument is the missing correlation with patient outcomes (Murphy et al., 2016). While the correlation between time and outcome for patients with sudden cardiac arrest has been shown, there is a lack of evidence to suggest a general RTT for all emergencies (Al-Shaqsi, 2010; Valenzuela et al., 1997). The binary nature of RTT is additionally deemed unsuitable for providing insight into the quality of care or patient outcomes according to Murphy et al. (2016). We further explore planning criteria in Part II of this thesis.

Chapter 3

EMS in Baden-Württemberg: Design and Data

The research in this thesis accompanied the project EVRALOG-BW[1] in cooperation with the EMS in Baden-Württemberg, Germany, and we conducted several experiments and analyses with historical data from that EMS system. This chapter introduces the EMS system in Baden-Württemberg to provide the foundation for the following parts of this thesis. The EMS in Germany is organized and regulated separately in each of the 16 federal states, and the contents of this chapter apply specifically to the EMS in the federal state of Baden-Württemberg.

Section 3.1 introduces the organization of the EMS in Baden-Württemberg with its EMS regions and central elements. The legislation for EMS in Baden-Württemberg underwent several changes during the development of this thesis, and Section 3.2 presents its evolution. Lastly, Section 3.3 describes the data available for this thesis.

3.1 Organization and Elements of EMS in Baden-Württemberg

The EMS in Baden-Württemberg is organized in the 35 EMS regions depicted in Fig. 3.1. The geographic boundaries of EMS regions follow administrative boundaries. For each EMS region, a regional committee is responsible for making logistic decisions for the ground-based EMS, for instance, decisions on ambulance station locations and allocations. With two exceptions, each EMS region is coordinated by one coordination center, and the coordination centers coordinate both EMS and fire brigade operations.

[1]Funded by the Ministry of the Interior, Digitalisation and Local Government in Baden-Württemberg, Germany. Further information: https://healthcarelab.ksri.kit.edu/english/projects.php

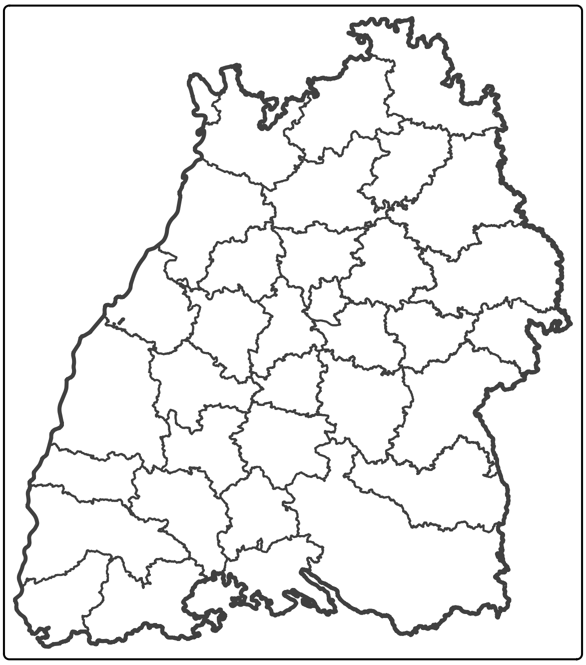


Figure 3.1: EMS regions in Baden-Württemberg.

The EMS in Baden-Württemberg is responsible for providing both emergency rescues and patient transportations. Several resources are available to fulfill these tasks, as depicted in Fig. 3.2. *Coordination centers* coordinate both emergency rescue and patient transportation operations.

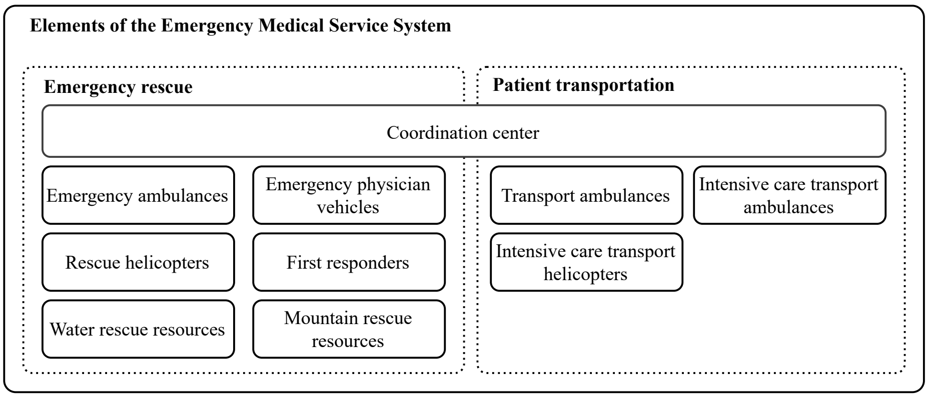


Figure 3.2: Elements of EMS in Baden-Württemberg within thesis scope.

Emergency ambulances conduct the majority of emergency rescue operations. They are equipped to both treat patients on scene and transport them to a facility for further care, if necessary. This thesis focuses on emergency rescue operations, and we use the term ambulance in the following to refer to emergency ambulances, unless stated otherwise. Emergency physicians are dispatched to critical emergencies alongside ambulances. They use *emergency physician vehicles* to drive to the scenes of emergencies. Emergency physician vehicles are equipped to treat patients on scene, but do not have the capacity to transport patients. *Rescue helicopters* are also staffed by emergency physicians and have the capacity to transport patients. They can reach remote locations and reduce transport times to facilities for further care. *First responders* aim to reduce response times to especially time-critical emergencies, such as sudden cardiac arrest. Different volunteer-based first responder systems exist in Baden-Württemberg, and smartphone-based first responder systems have gained increasing traction in recent years (Metelmann et al., 2021; Müller et al., 2024). Special *water rescue* and *mountain rescue resources* serve to reach, treat, and transport patients in water bodies and mountainous terrain, respectively.

Transport ambulances are the primary resource to conduct patient transportations. Their equipment and the required qualifications of the crew differ from those of emergency ambulances. *Intensive care transport ambulances* enable transporting patients who require continued medical treatment and/or supervision beyond the capabilities of transport ambulances. Especially for longer transport durations, *intensive care helicopters* are additionally available for patient transportation operations.

Table 3.1 provides indicators of the structure of the EMS system in Baden-Württemberg to facilitate a basic comparison to further EMS systems the reader might be familiar with.

Table 3.1: Basic indicators of the EMS in Baden-Württemberg

Indicator	Value
Inhabitants[2]	11.2 mil.
Area[2]	35.748 km^2
Emergency ambulance trips 2024 (SQR-BW, 2025)	1.1 mil.
Emergency physician vehicle trips 2024 (SQR-BW, 2025)	0.29 mil.
Emergency ambulances[3]	~ 480
Emergency ambulance stations[3]	~ 350
Emergency physician vehicles[3]	~ 210
Emergency physician vehicle stations[3]	~ 200

3.2 Legislation in Baden-Württemberg

The legislation for EMS Baden-Württemberg operates on two levels. The Act on Emergency Medical Services “RDG” (2024) defines the legal framework. The provisions of the Act are detailed further and operationalized in an EMS plan. Both the EMS Act and the EMS plan evolved during the creation of this thesis. The most relevant provisions for this thesis refer to the definition of planning criteria. Table 3.2 lists the evolution of Acts and plans with the contained planning criteria. The most recent EMS Act defined the formal status of the EMS plan as a regulation[4].

[2] Data retrieved from https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/Administrativ/02-bundeslaender.html on 10.12.2025

[3] Data retrieved from https://www.sqrbw.de/rettungsdienst/notfallrettung on 10.12.2025.

[4] https://www.landesrecht-bw.de/perma?d=jlr-RettDPlVBWrahmen. Access: 14.07.2026.

Table 3.2: Planning criteria for emergency rescue operations in legislation of EMS in Baden-Württemberg. Note: The technical definition of the response time in the legislation differs slightly from the definition in Fig. 2.2.

Published	Type	Planning criteria
2009	Act	• response time should not be more than 10, and no higher than 15 minutes
2014	Plan	• response time should not be more than 10, and no higher than 15 minutes in 95% of cases
2022	Plan	• response time of 12 minutes in 95% of cases for emergency rescue operations, where utmost urgency is required to save human lives or prevent serious damage to health
2024	Act	• response time of 12 minutes in 95% of cases • prehospital time of 60 minutes in 80% of cases • further categories and planning criteria can be defined in an additional regulation
2026	Regulation	• response time of 12 minutes in 95% for category 1 • prehospital time of 60 minutes in 80% for category 2 • response time of 30 minutes in 80% for category 3

Two major changes concerning the planning criteria occurred during the evolution of Acts and plans. The first change was introduced with the EMS plan in 2022. It changed the response time target to 12 minutes, and it explicitly specified that the response time target only applies to emergencies, for which utmost urgency is required to save human lives or prevent serious damage to health. The EMS Act in 2024 introduced the second major change. It firstly defined multiple categories with differentiated planning criteria. Secondly, it defined a planning criterion based on the prehospital time instead of the response time for one of the categories. The EMS Act in 2024 was published after the research in Part II of this thesis had been conducted and published by Watzinger et al. (2024).

A further change during the evolution of Acts and plans concerns the dispatching decision. The EMS plan in 2014 stated that the suitable EMS resource has to be dispatched to a call that reaches the call location quickest. The EMS plan in 2022 explicitly requires this dispatching strategy exclusively for calls for which the coordination center enables the use of sirens for ambulances. The EMS regulation published in 2026 requires dispatching the closest available ambulance to calls, for which utmost urgency is required to save human lives or prevent serious damage to health. If this is not the case, the dispatching decision should consider the response time to the call at dispatch, and additionally, the effect on the availability of resources, and how well the remaining resources can reach calls. This enables not dispatching the closest available resource to every call, which we investigate in Part IV.

3.3 Thesis Data

The Department for Quality Assurance in Emergency Medical Services in Baden-Württemberg (dt.: Stelle zur trägerübergreifenden Qualitätssicherung im Rettungsdienst in Baden-Württemberg) (SQR-BW) provided the historical data we use in this thesis. The SQR-BW is responsible for the quality assurance of EMS in Baden-Württemberg. It collects the data from every EMS operation in Baden-Württemberg, both from coordination centers and documentation from EMS resources. It harmonizes the data and publishes a yearly quality report with a wide range of both logistic and medical quality indicators. SQR-BW (2025) is the most recent quality report at the time of writing.
This thesis utilizes two datasets, one from 2021 and one from 2023. The datasets contain call response data from eleven EMS regions. Calls from an EMS region include all calls to which an ambulance and/or emergency physician vehicle stationed within the EMS region was dispatched, irrespective of the geographic location of the call. For each call response, the data includes the timestamps along the rescue chain of each call response, as well as information regarding the dispatched resources, call locations, and transport destination locations. Table 3.3 provides key statistics regarding the size and completeness of the datasets. The SQR-BW developed empirical validity requirements for evaluating time intervals (SQR-BW, 2025). The proportions of available time intervals in Table 3.3 refer to available valid time intervals. The SQR-BW additionally provided the historic locations and allocations of ambulances, as well as hospital data, including locations and facilities available in the hospitals.

Table 3.3: Data quality of the historical emergency call dataset for the years 2021 and 2023. Prehospital time availability is evaluated for calls with transport.

	2021	2023
calls	390,597	423,202
call coordinates available	99.9%	99.9%
pretravel delay available	85.2%	85.5%
driving times available	91.6%	91.8%
response time available	82.7%	82.7%
prehospital time available	83.6%	97.6%

We analyze the spatial and temporal distributions of calls in the following to facilitate a fundamental understanding of the call demand to be served by the EMS in Baden-Württemberg. Fig. 3.3 displays the spatial distribution of call locations in the eleven EMS regions. The EMS regions form two clusters, which we will refer to as study regions in this thesis. The call densities vary significantly within the EMS regions, with higher densities in urban and lower densities in rural areas. The datasets contain EMS regions with both urban and rural areas, primarily in the northern study regions, and mainly rural EMS regions, primarily in the southern study region. The experiments conducted in this thesis indicate that the effect of logistic decisions can depend on the characteristics of the region for which they are taken. Novel approaches to taking logistic decisions should therefore be studied in regions with different characteristics, a requirement met by the diverse mix of EMS regions in our datasets.

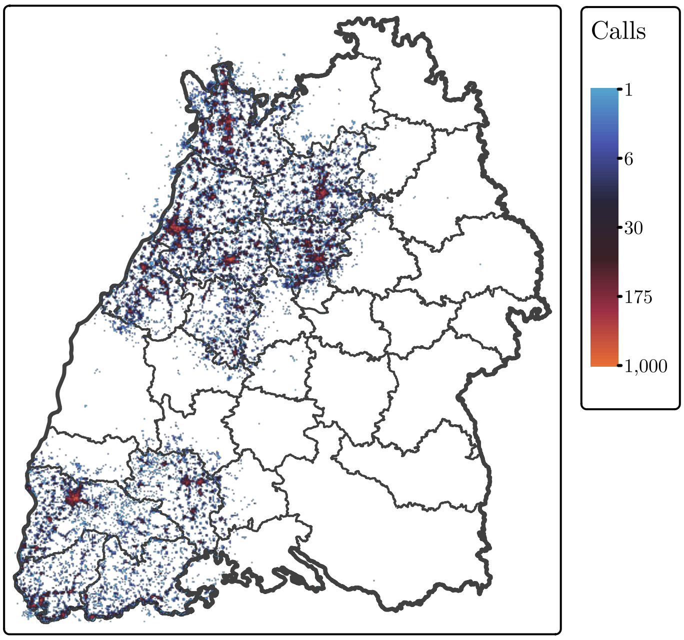


Figure 3.3: Spatial distribution of calls in study regions. Raster resolution: 500m.

Fig. 3.4 depicts the temporal distribution of calls over the hours of the day. We separate the analysis by weekdays, Saturdays, and Sundays to highlight that the spatial distribution can vary depending on the day.

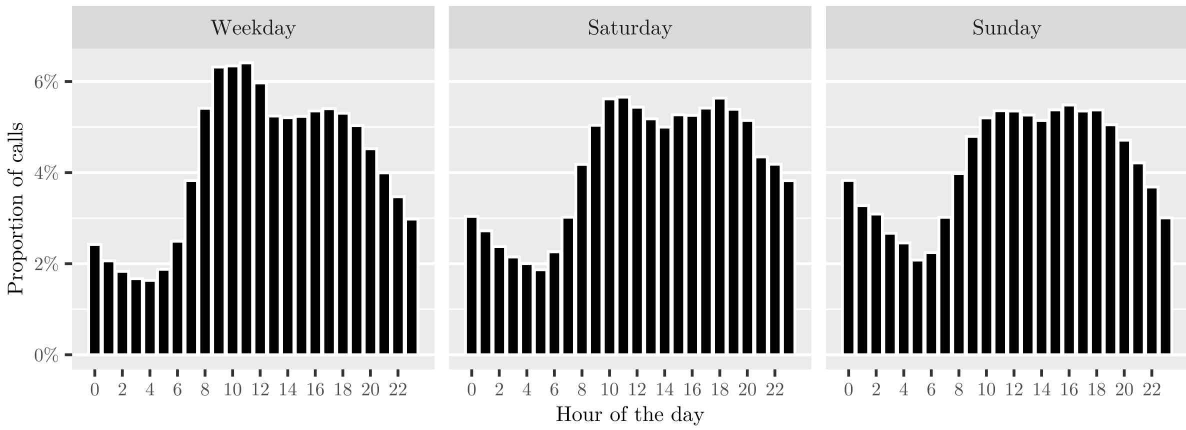


Figure 3.4: Temporal distribution of calls per hour of the day.

Part II

Quality Evaluation of EMS Logistics

Summary of Part II

Research Question and Objectives

Research Question: How can the quality of EMS logistics be quantified?
Research Objectives:

- Review the concept of quality and existing quality indicators in EMS
- Develop a framework for quality indicators in EMS logistics
- Develop a novel measure to quantify the effectiveness of EMS logistics for heterogeneous patients

Scope and Methodology

The tasks of EMS are patient transportation and emergency rescue. We focus on quantifying the quality of emergency rescue operations. The aim of investigating the quality of EMS logistics is informing logistic decisions. We therefore limit the scope to quality dimensions that are influenced by logistic decisions. Fig. 3.5 displays the quality dimensions that we deem as influenced by logistic decisions. We identify and address a gap in quantifying the effectiveness of EMS logistics.

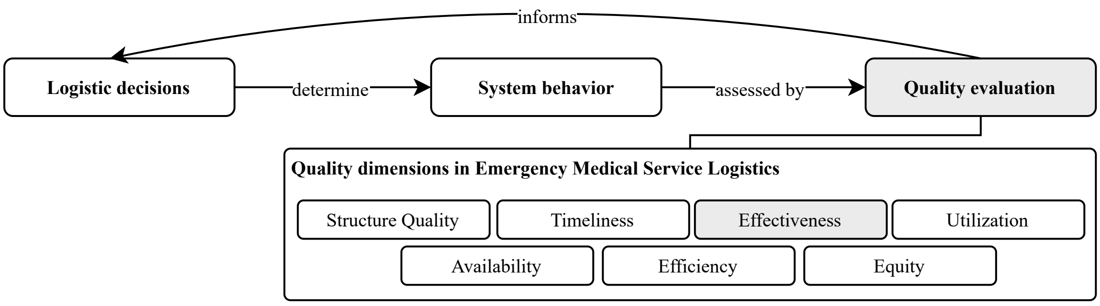


Figure 3.5: Part II investigates how the quality of EMS logistics can be quantified and aims to close an identified gap in quantifying the effectiveness of EMS logistics.

We conduct a *literature review* to identify existing quality indicators and quality dimensions in EMS. Based on the review, we develop a *framework for quality indicators* in EMS logistics. To address an identified gap in quantifying effectiveness, we employ a *consensus process* with medical experts to develop patient categories and utility functions that quantify the importance of time for different patient categories.

Main Contributions

- Development of a framework for quality indicators in EMS logistics
- Identification of a research gap in quantifying the effectiveness of EMS logistics
- Development of patient categories and utility functions for heterogeneous patients through an expert consensus process

Key Findings

- Existing quality indicators in EMS logistics focus on response times and do not sufficiently capture the effectiveness of logistic decisions.
- The importance of time and the relevant time interval along the rescue chain depend on the patient's needs.
- The current medical evidence and availability of connected outcome and process data is insufficient to quantify the effectiveness of EMS logistics for the majority of patients data-based.
- Utility functions defined by medical experts for patient categories are a step towards quantifying the effectiveness of EMS logistics for heterogeneous patients.

Chapter 4

Reviewing Quality Indicators for EMS Logistics

> You can't improve what you don't measure.
>
> Harbour (1997)

Decision makers assess logistic decisions by how they impact cost and by how they impact quality. Quantifying both cost and quality enables an objective assessment of this trade-off. Cost can usually be quantified monetarily. Quality, on the other hand, is much harder to quantify. How a society perceives the quality of a healthcare system is subject to the values of the society. This makes it difficult to define and even harder quantify quality objectively and universally. Any quantification of the quality of a healthcare system, including EMS, should therefore be continually challenged. Quantifying quality is, nevertheless, necessary to evaluate logistic decisions as objectively as possible, and defining and measuring quality indicators enables quantifying quality.

Internationally, EMS systems focus on the measure of response time as a quality indicator. As a statistic, mainly the percentage of response times within a specified threshold, the RTT, is used (Pons et al., 2005; Reza Hosseini et al., 2017). There is, however, no consensus on what threshold and what percentage should be used. (Schehadat et al., 2017) found the legal thresholds in Germany to vary between the federal states, with values between 10 and 15 minutes. In Italy, 8 minutes are used in urban areas (Aringhieri et al., 2016), in the Slovak Republic, Switzerland, and Portugal, 15 minutes (Carvalho et al., 2020; Jánošíková et al., 2019; Strauss et al., 2021) and in parts of France 20 minutes (Aboueljinane et al., 2014). Some EMS systems use differentiated thresholds. In the Netherlands, 15 minutes are used for the most urgent and life-threatening calls, and 30 minutes are used for urgent, but not life-threatening calls (Reuter-Oppermann et al., 2017). England differentiates between four categories with values ranging from 15 to 180 minutes, depending on the type of emergency (Department of Health & Social Care, 2023). Poland differentiates between urban and rural areas, using two thresholds each, with 8 and 15 minutes in cities with more than 10,000 inhabitants and 15 and 20 minutes for rural areas (Luxem et al., 2016). Typically, RTTs are defined as a target in conjunction with a percentage, for example, 95% of the most urgent and life-threatening calls in the Dutch system must be reached within 15 minutes (Reuter-Oppermann et al., 2017). This target is then used as the central planning indicator, especially for logistic decisions, and the achieved percentage of calls reached within the RTT is used as a quality indicator.

The focus on RTTs as the central quality indicator for EMS has been challenged. The main argument against the exclusive use of RTTs as a quality indicator is the missing correlation with patient outcomes (Murphy et al., 2016). Although the correlation between time and outcome for patients with out-of-hospital sudden cardiac arrest (OHCA) has been shown (Bürger et al., 2018; Neukamm et al., 2011; Valenzuela et al., 1997), there is a lack of evidence to suggest a general RTT for all emergencies (Al-Shaqsi, 2010; Myers et al., 2008; Pons & Markovchick, 2002). Especially the binary nature of RTTs is not suitable for providing information on quality of care or patient outcomes (Murphy et al., 2016). Therefore, the quality indicators currently used in practice do not fully reflect the multidimensional nature of quality in EMS. The following chapter will review the concept of quality in EMS and existing quality indicators to investigate how quality in EMS logistics can be quantified in addition to RTTs.

Related work on quality indicators in EMS is presented in Section 4.1. The concept of quality in healthcare and its dimensions are discussed in Section 4.2 and the quality dimensions relevant to EMS logistics are extracted. Existing quality indicators are examined in Section 4.3 and a framework for quality indicators is proposed based on quality dimensions. Categorizing quality indicators in the framework leads to the identification of a gap in quantifying the *effectiveness* of EMS logistics.

4.1 Related Literature on Frameworks for Quality Indicators in EMS

Existing literature on quality indicators either focuses on medical indicators or covers all aspects of EMS. To the best of our knowledge, there is no literature relating specifically to quality indicators for the logistic aspects of EMS.

El Sayed (2012) reviews quality indicators and expands on clinical quality indicators. Identified indicators are sorted into the types: structure, process, and outcome. Pap et al. (2018) present a scoping review on prehospital care quality indicators, differentiating between clinical quality indicators and system/organizational quality indicators. Haugland et al. (2019) concentrate on quality indicators for physician-staffed EMS in their review. Quality indicators are grouped into the dimensions of structure, process, outcome, and patient satisfaction.

Next to dedicated reviews, overviews of quality indicators can also be found in studies designed to either create novel quality indicators or to select a set of quality indicators for a specific purpose. Myers et al. (2008) develop a set of evidence-based quality indicators tailored to the medical conditions of STEMI (ST-Segment Elevation Myocardial Infarction), Respiratory Distress, Status Epilepticus, and Trauma. Using a Delphi process, Murphy et al. (2016) define a set of key performance indicators for prehospital emergency care in Ireland, where some indicators result from an initial literature review and others are introduced by the panelists in the Delphi process. Haugland et al. (2017) focus on physician-staffed EMS and develop quality indicators using an expert panel method. The development of quality indicators is based on a conceptual framework combining the dimensions of structure, process, and outcome with the dimensions of patient satisfaction, safety, effectiveness, efficiency, equity, and timeliness. Lastly, Coster et al. (2018) employ a consensus process to arrive at quality indicators grouped into patient outcome measures, clinical management measures, and whole system measures.
The absence of research specifically addressing quality indicators for EMS logistics motivates the following review of quality dimensions and quality indicators to gain an improved understanding of how the effects of logistic decisions can be evaluated.

4.2 Reviewing Quality Dimensions for EMS Logistics

Section 4.2.1 reviews dimensions of quality in healthcare to arrive at an understanding of quality in EMS. Next, Section 4.2.2 investigates which of those quality dimensions are relevant for logistic aspects of EMS.

4.2.1 Reviewing Quality Dimensions in Healthcare

Different approaches to defining quality in healthcare exist. Campbell et al. (2000) divide the approaches into *generic* and *disaggregated.*
Generic approaches try to define quality as a whole. The "most widely cited" (El Sayed, 2012) and "most applicable" (Moore, 1999) generic definition of quality in healthcare was introduced by Institute of Medicine (1990, p.4) as: "Quality of care is the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge." This definition implies that the major focus of quality in healthcare lies on health outcomes. It also hints at a differentiation between an individual and a population perspective.
Disaggregated approaches aim at defining quality by dividing it into disaggregated dimensions and then defining those dimensions of quality. Donabedian (1980) differentiates between *structure quality*, *process quality* and *outcome quality* (s. Fig. 4.1). Pap et al. (2018) provide concise definitions of those dimensions, where *structure* refers to "the attributes of the setting in which care is provided", *process* encompasses "the activities that contribute

to healthcare being carried out by healthcare practitioners", and *outcome* reflects the "effects of healthcare on individuals and populations". Especially, the outcome quality seems to have been a basis for the generic definition of Institute of Medicine (1990), who adopts the differentiation between the effects on individuals and on populations. Looking at the dimensions of structure, process, and outcome, a directional relationship exists in the sense that a good structure quality supports a good process quality, which in turn supports a good outcome quality. From this observation follows that the requirements for process quality should be derived from its effect on outcome quality, and the requirements for structure quality should be derived from its impact on process quality.

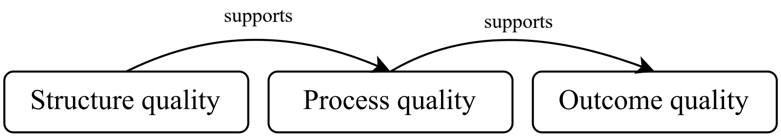


Figure 4.1: Dimensions of quality in healthcare according to Donabedian (1980).

Campbell et al. (2000) expand on the observation that quality can be viewed from an individual perspective and a population level perspective (s. Fig. 4.2). Quality on an individual level is disaggregated into *access* and *effectiveness*. *Access* reflects whether individuals have access to those parts of a healthcare system that meet their needs, while *effectiveness* reflects whether the received care in a healthcare system is effective at fulfilling the individuals' needs. Quality on a population level is disaggregated into the dimensions *equity* and *efficiency*. *Equity* captures the level of difference in access to effective care between individuals within a population. *Efficiency* refers to the amount of resources used to provide a population with access to effective care.

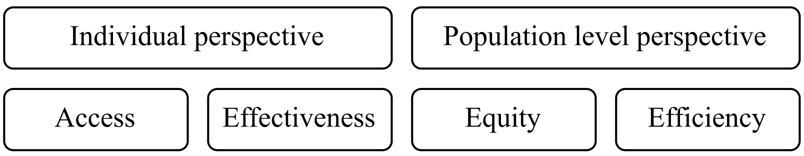


Figure 4.2: Dimensions of quality in healthcare according to Campbell et al. (2000).

The Joint Commission on Accreditation of Healthcare Organizations introduced a further approach to a disaggregated definition of quality in healthcare (Dunford et al., 2002). It distinguishes between the dimensions of *efficacy, appropriateness, availability, timeliness, effectiveness, continuity of care, safety, efficiency* and *respect and caring* (s. Fig. 4.3).

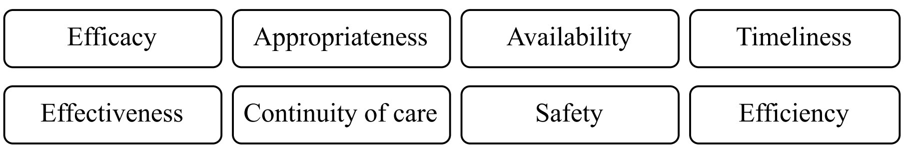


Figure 4.3: Dimensions of quality in healthcare according to Dunford et al. (2002).

Similarly, the World Health Organization (WHO, 2006) defines quality in healthcare along the dimensions of *effectiveness, efficiency, accessibility, acceptability/patient-centeredness, equitability* and *safety* (s. Fig. 4.4).

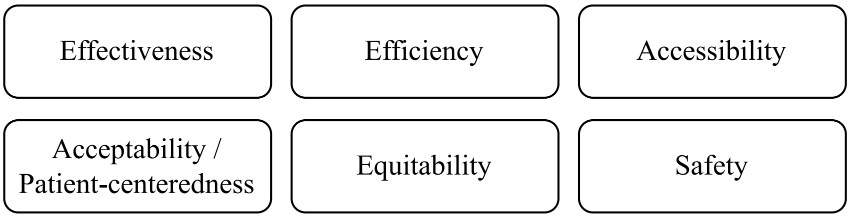


Figure 4.4: Dimensions of quality in healthcare according to WHO (2006).

Availability and *timeliness* can be seen as facets of the dimension of *access* in the context of logistic decisions. *Availability* reflects whether there are fundamental barriers to access. *Timeliness* is a necessary condition to access, since medical conditions often deteriorate over time, making effective treatment only possible in case of timely access.

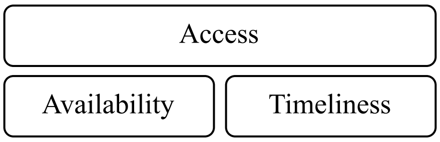


Figure 4.5: Connection between the dimensions of access, availability, and timeliness.

Both approaches of Dunford et al. (2002) and WHO (2006) incorporate the dimensions of *access*, *effectiveness* and *efficiency*, while *equitability* is only incorporated by WHO (2006). While those dimensions can be seen as part of an economic perspective, both Dunford et al. (2002) and WHO (2006) also explicitly incorporate the medical perspective by considering the *efficacy*, *appropriateness*, and *safety of care* and the fundamentally human nature of healthcare with the dimensions of *respect and caring* and *acceptability/patient-centeredness*. Dunford et al. (2002) finally reflect the often sectoral design of healthcare systems by including *continuity of care* as a quality dimension.

The different approaches to defining quality in healthcare share common themes while emphasizing different aspects. Generic approaches like Institute of Medicine (1990) focus on health outcomes as the primary measure of quality, while disaggregated approaches explicitly divide quality into dimensions. Despite their differences, all approaches recognize quality as a multidimensional concept that cannot be captured by a single measure, as the use of RTT in EMS would suggest. The disaggregated approaches are particularly relevant for EMS logistics, as they enable the identification of specific quality dimensions that might be influenced through different logistic decisions.

4.2.2 Identifying Quality Dimensions for EMS Logistics

EMS are a part of healthcare, and the presented quality dimensions all fundamentally apply to EMS. The focus of this thesis lies on the logistic aspects of EMS. We therefore reduce the discussed quality dimensions of healthcare to the dimensions that can be influenced through logistic decisions in EMS.

Starting with the dimensions of *structure quality*, *process quality*, and *outcome quality*, structure quality can be influenced by logistic decisions on the strategic and tactical level, with decisions on ambulance base locations or resource levels. Similarly, process quality can be influenced directly on the operational level, for instance, through the ambulance dispatching decision. While the structure and process quality influence outcome quality, outcome quality is not influenced by logistic decisions directly, but rather by the quality of the medical treatment.

The dimension of *access* can be influenced by logistic decisions in both its facets of *timeliness* and *availability*. Timeliness is reflected by the different time intervals along the rescue chain, while availability is reflected by sufficient resource availability. The dimension of *effectiveness* links outcome quality and process quality by capturing the degree to which the logistic processes enable a good outcome quality. This interpretation is supported by a literature review on logistics in EMS by Aringhieri et al. (2017), who defined effectiveness as comparing output and outcome. From the population level perspective, logistic decisions can influence both *equity* and *efficiency*. Decisions on the temporal and spatial distribution of resources influence equity. Additionally, according to Aringhieri et al. (2017), equity can be measured vertically and horizontally, meaning whether the same treatment is provided for the same needs and different treatments are provided for different needs. The operational decision that can be taken according to the differing needs of patients is the dispatching decision, showing that equity can be influenced by logistic decisions. Efficiency reflects how many resources are needed for a certain quality level, and the resource level is a direct consequence of logistic decisions on the strategic and tactical levels.

While the additional dimensions of *efficacy, appropriateness, safety, respect and caring, acceptability/patient-centeredness* and *continuity of care* reflecting the medical, human, and healthcare system design aspects of healthcare apply to EMS in general, no direct influence of logistic decisions on them is apparent.

To conclude, quality in EMS is a multidimensional concept that should be viewed holistically. The quality dimensions deemed relevant for logistic aspects of EMS are *structure quality*, *process quality*, *access* with its facets of *timeliness* and *availability*, *effectiveness*, *efficiency* and *equity* (s. Fig. 4.6).

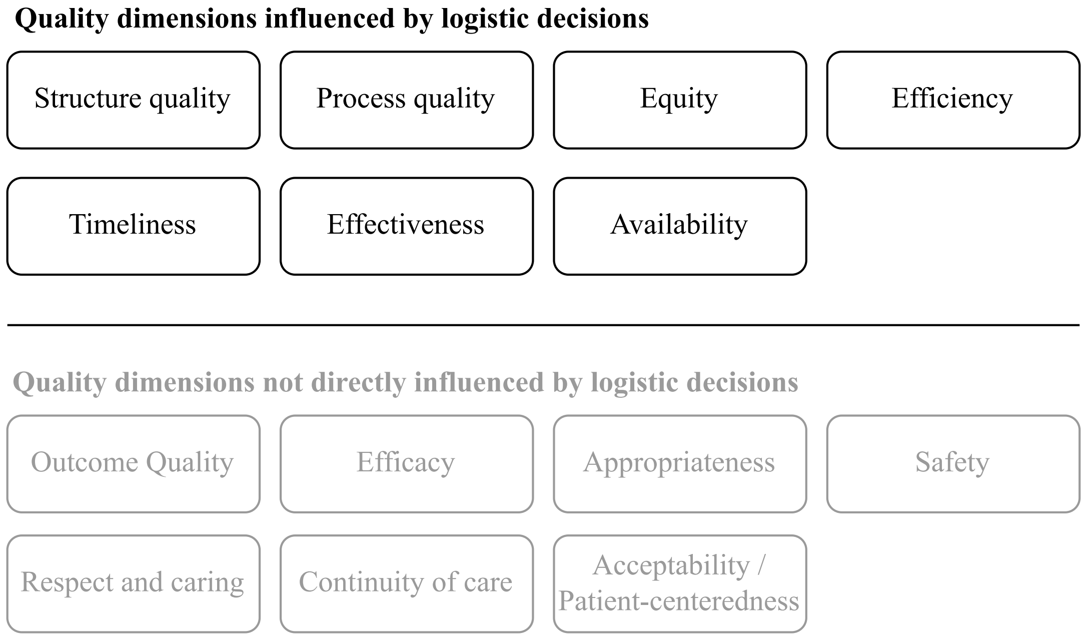


Figure 4.6: Dimensions of quality in EMS.

4.3 Developing a Framework for Quality Indicators in EMS Logistics

We have identified several quality dimensions applicable to EMS logistics. These quality dimensions allow structuring quality indicators for EMS logistics. In the following section, we develop a framework for quality indicators in EMS to connect individual quality indicators to the quality dimensions identified in Section 4.2.2. A quality indicator consists of a measure that is observed, for instance, for each call response, and a statistic that is used to aggregate the observations. The purpose of the framework is to facilitate a holistic evaluation of how logistic decisions impact the quality of EMS. We categorize quality indicators identified in studies of logistic decisions into our framework to gain an overview of how well the quality dimensions are represented.

4.3.1 Defining Framework Dimensions

The dimensions of the framework are based on the quality dimensions identified in Section 4.2.1. The framework is built along two axes. The horizontal axis is used to categorize the measure used in a quality indicator. The vertical axis categorizes the statistics used to aggregate the individual measures. Fig. 4.7 depicts the framework.

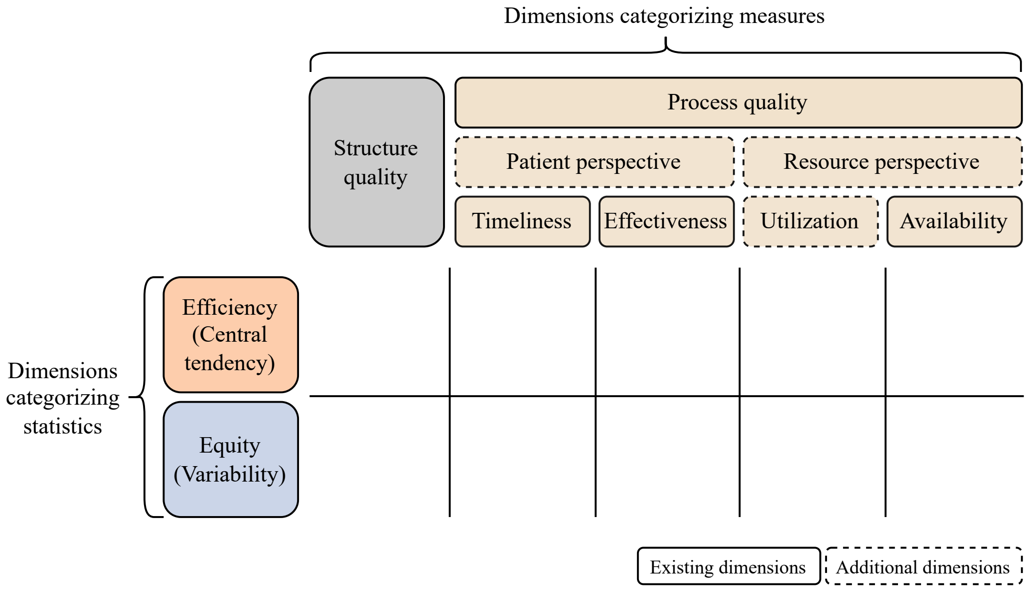


Figure 4.7: Framework for quality indicators for logistics in EMS.

Dimensions to Categorize Measures

The basic dimensions along the first axis are *structure quality* and *process quality*. As mentioned in Section 4.2.2, *outcome quality* is not deemed relevant for logistic quality indicators in this thesis. The exclusion of outcome quality is supported by the fact that no quality indicators in the context of logistic decisions were found that could be categorized as quantifying *outcome quality*. Within the framework, *structure quality* encompasses all measures that can be captured without a call response having been performed.

Within the dimension of *process quality*, a first distinction is made between quality indicators capturing the *patient perspective* and quality indicators capturing the *resource perspective*. The sub-dimensions *timeliness* and *effectiveness* differentiate the *patient perspective*. *Timeliness* could be seen as a component of *effectiveness*. Next to the convenience of short waiting times, the importance of *timeliness* stems primarily from the fact that some medical conditions deteriorate over time, making effective treatment more likely, the shorter the time to treatment is. The distinction between *timeliness* and *effectiveness* is that *timeliness* considers only time-based measures, which need not necessarily be derived from medical outcome data, while *effectiveness* interprets time in a way that can be directly linked to patient outcomes.

The *resource perspective* is divided into the sub-dimensions *utilization* and *availability*. From an economic perspective, a high *utilization* of resources is desirable. On the other hand, a high *availability* is desirable from a quality of care perspective. The possible conflict between those dimensions is the fact that a high *utilization* makes it more difficult to achieve a high *availability* of resources.

Dimensions to Categorize Statistics

The statistics used to aggregate individual measures for evaluating an EMS system are categorized into the dimensions of *efficiency* and *equity*. According to Aringhieri et al. (2017), *efficiency* compares input and output. Within the framework, statistics used to calculate the central tendency of individual measures are categorized within the *efficiency* dimension. The statistics categorized into the *equity* dimension, on the other hand, calculate the variability of individual observations. The underlying assumption is that *equity* can only be assessed by the differences between observations.

4.3.2 Categorizing Quality Indicators

We extract quality indicators from studies on logistic decisions in EMS and categorize them into the framework in the following. The detailed results of the considered studies and applied quality indicators of each study can be found in Section A.1. The categorization of the measures along the first axis is shown in Table 4.2 and the categorization of the statistics along the second axis in Table 4.1. Table 4.2 uses the following abbreviations:

- S: Structure quality
- P-P-T: Process quality - Patient perspective - Timeliness
- P-P-E: Process quality - Patient perspective - Effectiveness
- P-R-U: Process quality - Resource perspective - Utilization
- P-R-A: Process quality - Resource perspective - Availability

The iso-elastic social welfare function is a parametrized function. Depending on the parameter choice, it can represent an *efficiency* or an *equity* focused measure. For this reason, it is included in both dimensions in Table 4.1.

Table 4.1: Description of quality statistics.

Dimension	Statistic	Description
Efficiency	$\% < T$	percentage or proportion of observations below a threshold T
Efficiency	x_p	p-percentile of observations
Efficiency	mean	arithmetic mean of observations
Efficiency	mean | t > T	arithmetic mean of those observations, that are larger than a threshold T
Efficiency	max	maximum of observations
Efficiency	min	minimum of observations
Efficiency	sum	sum of observations
Efficiency	prop	proportion of observations, equals the arithmetic mean in case of binary observations
Efficiency	count	count of observations
Efficiency	utilitarian social welfare	calculates the social welfare as the mean of the individual utilities (Jagtenberg & Mason, 2020b)
Efficiency	iso-elastic social welfare	parametrized function which calculates the social welfare as a trade-off between the utilitarian and the Bernoulli-Nash social welfare (Jagtenberg et al., 2021)
Equity	sd	standard deviation of observations
Equity	cv	coefficient of variation of observations
Equity	Bernoulli-Nash social welfare	calculates the social welfare as the product of the individual utilities (Jagtenberg & Mason, 2020b)
Equity	iso-elastic social welfare	parametrized function which calculates the social welfare as a trade-off between the utilitarian and the Bernoulli-Nash social welfare (Jagtenberg et al., 2021)

Table 4.2: Description of quality measures.

Dimension	Measure	Description
S	Coverage	A demand node (emergency location) is defined as covered if it can be reached from a supply node (ambulance station) within a certain response time.
S	Response Time	Measures the time it takes an ambulance to reach a demand node from a supply node.
S	Driving distance to nearest base	For each demand node, the driving distance to the nearest supply node is measured.
S	EMS units per 100,000 inhabitants	
S	EMS units per km^2	
S	EMS stations	
S	Cost of stations and ambulances	
P-P-T	Call wait time	Measures the time between call arrival in a dispatch center and the call pick-up by a call-taker.
P-P-T	Dispatch time	Measures the time between the call pick-up by a call taker and the alarm of an (available) ambulance.
P-P-T	Response time	Measures the time between call arrival in a dispatch center and arrival of an ambulance at the scene.
P-P-T	On scene time	Measures the time spent at the scene, either until the transport to a hospital begins or the treatment at the scene is concluded, if no transport is necessary.
P-P-T	Service time	Measures the time between when the ambulance starts driving to the scene and the arrival at a hospital.
P-P-T	Scene to hospital time	Measures the time between leaving the scene and arriving at a hospital.
P-P-T	Prehospital time	Measures the time between call arrival in a dispatch center and arrival at a hospital.
P-P-T	Alarm to patient handover time	Measures the time between alarming an ambulance and handing the patient over at a hospital.
P-P-E	Survival probability	Measures survival probability as a function of response time (Bürger et al., 2018; Maio et al., 2003; Valenzuela et al., 1997).
P-R-U	Round-trip time	Measures the time between leaving the station and arriving back at the station.
P-R-U	Mileage	Measures the distance driven by an ambulance.
P-R-U	Utilization rate	Measures the proportion of shift time spent on call responses.
P-R-U	Ambulances simultaneously utilized	Measures the number of ambulances that are responding to calls at a given moment in time.
P-R-U	Time for which no ambulance is busy	Measures the time, where every ambulance is available for a new call response.
P-R-U	Unused ambulances	Measures the number of ambulances that were not needed for any call response.
P-R-U	Overtime for paramedics	Measures the time during which ambulances are responding to a call or driving back to their base outside their shift times.
P-R-U	Calls per ambulance	Measures the number of calls responded to by each ambulance.
P-R-U	Calls per base	Measures the number of calls responded to by each ambulance station.
P-R-A	Ambulance available at call	Measures, whether an ambulance is available, when one is needed for a call response.
P-R-A	Dispatch from a station, which is not the closest	Measures, whether an ambulance is dispatched from a station, which is not the closest to the emergency, e.g. because no ambulance is available at the closest section.
P-R-A	Call queue length	Measures the number of calls waiting for an ambulance to become available.
P-R-A	System preparedness	Measures the system preparedness as calculated by Carvalho et al. (2020) and S. Lee (2017)

Fig. 4.8 shows two counts as a result of the literature review to get an overview of how often which quality dimension is assessed. The first count indicates how many distinct quality indicators were found in each cell of the framework. The second count shows how often those quality indicators were used throughout the literature. These counts in combination provide an overview of the relevance of the different quality dimensions in the literature.

	Structure quality	Process quality			
		Patient perspective		Resource perspective	
		Timeliness	Effectiveness	Utilization	Availability
Efficiency (Central tendency)	(15 | 22)	(18 | 117)	(3 | 10)	(12 | 33)	(6 | 9)
Equity (Variability)	(2 | 3)	(3 | 11)	(0 | 0)	(2 | 5)	(0 | 0)

(Count of distinct indicators | Sum of usages in the literature)

Figure 4.8: Framework for quality indicators for logistics in EMS.

4.4 Quality Indicators in EMS Logistics: Discussion

The differing counts in Fig. 4.8 lead to several conclusions. Firstly, the majority of indicators focus on *efficiency* rather than *equity*. This could be due to the fact that the focus in practice is on response time targets, which fall into the *efficiency* dimension. Another possible reason is that notions of *equity* are more difficult to quantify than notions of *efficiency*. While the importance of *equity* in the context of EMS has been recognized, quality indicators reflecting this quality dimension are seldom used.

A second observation is that most authors only evaluate the *process quality* and not the *structure quality* of the investigated EMS systems. This complicates comparing the different systems and judging the transferability of approaches between systems.

Looking at the indicators capturing the patient perspective, a clear emphasis can be seen on indicators measuring the *timeliness* in comparison to indicators measuring the *effectiveness* of the system. Additionally, according to Campbell et al. (2000), "definitions of access must emphasize a notion of access according to need and the timely use of services". While the indicators do measure how timely the access of the patients is, there is little to no relation to the different needs of the patients. The only measures used to quantify the *effectiveness* are survival functions. They are based solely on patients with OHCA, which constitute around 1–2% of EMS responses according to Myers et al. (2008). This is in line with results in Baden-Württemberg, where they constituted less than 2% of the emergency volume in 2022 (SQR-BW, 2023).

The most commonly used indicators in the literature focus on *efficiency* and *timeliness*. The single indicator with the highest usage is the percentage of calls reached within a certain RTT. This shows that the identified approaches mainly focus on improving the quality indicators present in current legislation and often do not challenge their validity.

The variety of identified quality indicators demonstrates that quality in EMS is multidimensional. Except for evaluating RTT, however, there seems to be no consensus yet on which set of quality indicators should be used to quantify the different quality dimensions.

Existing studies on quality indicators in EMS have focused on either purely clinical indicators or on a combination of clinical and whole system/organizational quality indicators. To the best of our knowledge, the presented framework presents the first approach focusing on quality indicators for the logistic aspects of EMS. This was achieved by connecting quality indicators employed in studies of logistic decisions to quality dimensions that originated in the medical literature.

Haugland et al. (2017) proposed a conceptual framework created by combining two sets of quality dimensions to develop quality indicators for physician-staffed EMS. This approach of creating a framework by combining quality dimensions was adapted in this work to create a framework for quality indicators for EMS logistics.

The framework can support decision-makers in practice by providing a multidimensional overview of the quality of their EMS, which is influenced by logistic decisions. This can inform improvement efforts as well as support the understanding of the effects of changes implemented in EMS. Researchers in the field of EMS logistics can use the framework to evaluate the impact of planning approaches more holistically. This can support in creating an understanding of how to model which decision to influence the quality dimensions.

The main limitation of the presented work is the omission of a systematic literature review to identify quality indicators. Each quality indicator could be assigned to exactly one cell in the framework, and each dimension of the framework contains at least one quality indicator, but basing especially the counts of the indicators in Fig. 4.8 on a systematic literature review would further enhance the validity of the results.

One avenue for further research would be to investigate whether a minimal set of quality indicators can be identified that is still able to reflect the range of quality dimensions. This would constitute a step towards a more consistent and comparable evaluation of logistic decisions in research and in practice.

Regarding the overarching dimensions of structure quality, process quality, and outcome quality, the direction of impact leads from structure quality to process quality to outcome quality, and the direction of requirements leads from outcome quality to process quality to structure quality. The definition of process quality indicators should therefore be derived from their impact on outcome quality. The quality dimension of *effectiveness* most closely links process quality to outcome quality, as it compares output to outcome. The only quality measure found in this quality dimension is the survival probability measured with survival functions. For the vast majority of emergencies, there exists no indicator linking outcome quality and process quality. The weaknesses of response time targets have therefore not been fully addressed, and a gap remains in quantifying the *effectiveness* of EMS logistics.

4.5 Quality Indicators in EMS Logistics: Conclusion

This chapter aimed to create a quantifiable understanding of quality in EMS logistics through quality indicators. We reviewed quality dimensions in health care have, and derived quality dimensions relevant for EMS logistics. We subsequently reviewed quality indicators and proposed a framework for quality indicators. The framework consists of sets of quality dimensions along two axes, one to categorize the measures and one to categorize the statistics of quality indicators. We assigned the quality indicators to the dimensions of the framework, and derived several conclusions from the resulting counts.

Quality in EMS should be viewed as a multidimensional concept. A range of quality indicators exists, but the main focus remains on quality indicators capturing the *efficiency* of *timeliness*. This is consistent with the response time target-based quality indicators used in practice and especially in regulations. Outside the evaluation of response time targets, there appears to be no consensus across studies of logistics in EMS on which quality indicators should be used to evaluate logistic decisions.

We identified A gap in quality indicators measuring the *effectiveness* of EMS logistics. *Effectiveness* relates output to outcome, and is the quality dimension through which the effects of logistic decisions can be connected to patient outcome. This makes *effectiveness* a crucial quality dimension and the development of quality indicators capturing it a promising field for further research.

Chapter 5

Developing Utility Functions for Heterogeneous Patients

Chapter 4 identified a research gap in measuring the effectiveness of EMS logistics. The quality dimension of effectiveness relates output in terms of time intervals to outcome in terms of patient outcomes. Improving the quantification of effectiveness can therefore lead to a better understanding of how logistic decisions should be taken to improve patient outcome, which is a central objective of EMS.
Currently, the sole measures of effectiveness are survival functions that map response time to a survival probability. Survival functions have been used to model the ambulance location problem by Erkut et al. (2008) and Knight et al. (2012) (s. Fig. 5.1) and relate the time interval of the response time to patient outcome in terms of survival probability.

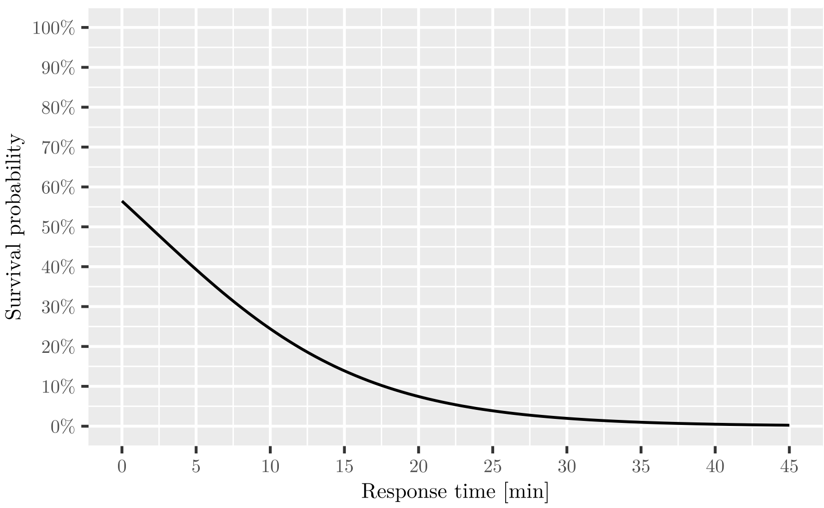


Figure 5.1: Survival function used by Knight et al. (2012)

The drawback of survival functions is that they are based solely on patients with OHCA, which constitute around only 1–2% of EMS responses according to Myers et al. (2008), and it is questionable whether they apply to all emergencies. The medical recommendation, for instance, for a stroke is to reach a hospital within 60 minutes (Fischer et al., 2016).

This implies that there are patients for whom the prehospital time is more important than the response time, and the effectiveness of EMS logistics should therefore be measured based on the prehospital time instead of the response time for those patients.

Our fundamental approach for addressing the research gap is to investigate if and how survival functions can be generalized and defined for the different patients encountered in EMS. As not all EMS patients are in an immediately life-threatening situation, the term *survival function* appears unsuitable for the generalization to different patient categories, and a more abstract term is needed. Jagtenberg and Mason (2020b) introduced the term *utility functions* to the literature on EMS logistics. The term originates from social welfare theory. The basic idea is that a utility function maps the output of a system to a utility value for each individual within the system. Since the outcome of a patient can be seen as the utility that an EMS system generates for the patient, we deem this term to be more suitable and use it in the following to describe the quantification of the importance of time for different patient categories.

The development of utility functions constitutes a new measure as a step towards new quality indicators. According to Stelfox and Straus (2013), two broad approaches to developing quality indicators exist: a deductive approach (from concept to data) and an inductive approach (from data to concept). Deductive approaches start with known concepts and existing scientific evidence and then deduce quality indicators that are able to capture those concepts. An example would be to start with the scientific evidence that the brain is damaged if it is not supplied with oxygen and that the damage increases over time (Larsen et al., 1993). From this, we can deduce that measuring the time from the onset of an OHCA until cardiopulmonary resuscitation is initiated could be a suitable quality indicator. Inductive approaches, on the other hand, start with existing data and try to induce quality indicators by analyzing variation and correlation in the data. Applied to the example of OHCA, one could analyze the differences in survival rates and the correlation between time until initiating cardiopulmonary resuscitation and survival probability to infer that measuring the time from the onset of an OHCA until cardiopulmonary resuscitation is initiated could be a suitable quality indicator.

The necessary data to choose an inductive approach for this work would have to include the outcome data of EMS patients. Since this data is not available, a deductive approach based on expert opinion appears as a suitable method. As stated by Stelfox and Straus (2013), expert opinion can be used in deductive approaches if the scientific evidence base is limited. As reasoned in Section 4.4, this is the case for patients with conditions other than OHCA.

The remainder of this chapter is structured as follows. Section 5.1 presents related literature on developing quality indicators. The design of the consensus process is explained in Section 5.2. The methodological background on how we derive utility functions from expert estimations is given in Section 5.2.3. The results of the consensus process are presented in Section 5.3.1, detailing the patient categories, and in Section 5.3.2 examining

the utility functions. Applying the results to historical data in Section 5.3.3 leads to several insights regarding possible applications of the research in practice. The findings are discussed in Section 5.4 and the chapter is concluded in Section 5.5.
The results of the consensus process have been published by Watzinger et al. (2024).

5.1 Related Literature on Developing Quality Indicators for EMS

Pap et al. (2018) identified the consensus process as the most frequent method in the development of quality indicators. Murphy et al. (2016) developed a set of quality indicators for prehospital emergency care in Ireland using a two-stage Delphi consensus approach. Concentrating on physician-staffed EMS, Haugland et al. (2017) employed a four-step expert panel method to develop quality indicators. Coster et al. (2018) combine nominal group technique, a modified Delphi study, and a patient and public involvement consensus workshop in three phases to identify and prioritize ambulance performance measures. Blozik et al. (2018) followed a six-step informal consensus process to arrive at quality indicators for primary care in Switzerland. Howard et al. (2019) used a three-round modified online Delphi study to identify, refine, and review quality indicators for prehospital emergency care in South Africa. Within the Australian EMS, Pap et al. (2020) developed prehospital care quality indicators in three phases. The first phase consists of a scoping review, the second phase of an aggregation and rapid reviews of identified indicators, and lastly the third phase of an expert consensus process. Azami-Aghdash et al. (2021) looked at quality indicators for EMS specific to road traffic injury. Following a literature review, semi-structured interviews were conducted, and a two-round Delphi process was performed to confirm the final indicators. In Germany, Schüttig et al. (2022) developed quality indicators for ambulatory-care-sensitive conditions using a consensus approach consisting of six stages. After an initial literature review, several rounds of online and offline group discussions led to a concretization of indicators. The indicators were tested in a pilot study and finally refined in a concluding discussion. Vianen et al. (2025) define outcome parameters for physician-staffed Helicopter Emergency Medical Services in the Netherlands using a Delphi consensus study.
Existing studies focus on the selection and prioritization of a set of indicators for a specific purpose using consensus processes. This work differs from existing literature insofar as the aim is not to define a comprehensive set of quality indicators but to develop a novel measure as a step towards quality indicators that are able to quantify the effectiveness of EMS logistics.

5.2 Designing a Consensus Process to Develop Utility Functions

The following sections describe our methodological approach to developing utility functions using an expert consensus process. Section 5.2.1 presents the composition of the expert panel and the stages of the consensus process. Section 5.2.2 describes the consensus mechanisms employed to define the patient categories and the utility functions. We use spline interpolation and non-linear regression to calculate the utility functions based on expert estimations and introduce these methods in Section 5.2.3.

5.2.1 Expert Panel and Workshops

The expert panel consists of nine experts from the EMS in Baden-Württemberg, divided into two groups. The first group consists of five physicians from the fields of anaesthesiology and internal medicine with qualifications in emergency medicine. This group focuses on defining patient categories, mapping the patient categories to medical diagnoses, and estimating utility functions. A potential downside of choosing exclusively emergency physicians is that physicians are only involved in less than 30% of emergency responses (SQR-BW, 2024). Emergency physicians are called to the most critical emergencies, and, therefore, might have a bias when tasked with categorizing all emergency patients. On the other hand, only physicians are qualified to make full medical diagnoses for all patients. This competence is necessary when mapping the patient categories to diagnoses. The second group consists of four experts with a background in the operational side of EMS. Two experts are the heads of coordination centers, one the head of the EMS of a relief organization in Baden-Württemberg, and one has qualifications in quality assurance of EMS and is the deputy head of the SQR-BW. This group focuses on mapping the patient categories to keywords used in coordination centers. The presence of non-physician experts reduces the possible bias in the results, as the experience of the second group of experts is based on the whole range of possible emergency patients and could therefore validate the results of the first group. Designing and moderating the consensus process based on the author's background in Operations Research ensures an interdisciplinary perspective combining the medical expertise with a focus on the effects of logistic decisions in EMS. The consensus process is organized as a series of workshops. The overview of the conducted workshops is presented in Table 5.1. Apart from the last workshop, the two groups work independently. This reduces the risk of influence between the group and mitigates the risk of biases.

Table 5.1: Description of workshops in the consensus process

Name	Participants	Date	Input	Output
Workshop 1	Group 1	17.08.2022	-	Patient Categories
Workshop 2	Group 1	19.09.2022	Patient Categories and individual mappings of categories to diagnoses	Final Mapping of Patient categories to medical diagnoses
Workshop 3	Group 2	09.11.2022	Patient Categories	Preliminary Mapping of patient categories to keywords
Workshop 4	Group 2	20.12.2022	Preliminary mapping of categories to keywords	Modified version of the mapping of categories to keywords
Workshop 5	Group 2	01.03.2023	Modified mapping of categories to keywords	Final version of the mapping of categories to keywords
Workshop 6	Group 1	21.03.2023	Individual estimations of utility functions	Preliminary version of utility functions
Workshop 7	Group 1 & Group 2	26.04.2023	Final utility functions; Patient Categories; Final mapping of categories to keywords	Patient Categories; Final utility functions; Final mapping of categories to keywords

5.2.2 Designing the Consensus Mechanism

We design the consensus mechanism based on the following four basic elements, introduced by Jones and Hunter (1995), that need to be considered in a consensus process:

- *Anonymity*: private rankings by participants
- *Iteration*: processes occur in rounds, allowing individuals to change their opinions
- *Controlled feedback*: participants are shown the distribution of the group's response
- *Statistical group response*: judgment is expressed using summary measures of the full group response

The expert panel defined the patient categories through open discussion to ensure a common understanding of the categories. For each category, the panel unanimously deemed either the response or the prehospital time as being relevant from a patient perspective. The corresponding utility function is then defined for each patient category with the following steps:

1. Estimating $(time, utility)$ points for each category: Each expert in the panel is asked to independently estimate at least three $(time, utility)$ points for each patient category. The *utility* is a value in the interval $[0, 1]$ and represents the *utility* that a patient receives, if a certain *time* is achieved. Depending on the patient category, *time* represents either the response or the prehospital time (s. Table 5.2).

2. Calculating splines based on estimated points: To create continuous functions from the estimated points, we calculate spline interpolations using Hyman filtering (Hyman, 1983) to ensure monotone functions (s. Section 5.2.3).

3. Aggregating splines based on individual estimations: To arrive at utility functions as a summary of the individual estimations, we use non-linear regression. We calculate the average of the individual splines for each minute. We then derive the parameters a, b and c of the basis-function $f(time) = \frac{1}{a+e^{b+c*time}}$ with non-linear regression, resulting in a utility function for each category (s. Section 5.2.3).

4. Discussing and adapting estimations: After calculating the utility functions, the results of the individual estimations and the aggregation are shown to the experts. Based on those results and a discussion on the differences in estimations, steps 1–3 are repeated. As all experts were content with the results after one repetition, the process was concluded after one iteration.

By following these steps, the elements defined by Jones and Hunter (1995) are integrated. Anonymizing the sources of the estimations of the utility functions in step 4 ensures *Anonymity*. By allowing all participants to update their estimations after showing and discussing the preliminary results, *iteration* and *controlled feedback* are considered in the process. The final utility functions can be interpreted as a *statistical group response* as they are the result of aggregating the individual estimations within the group.

5.2.3 Introducing Methods for Estimating Utility Functions

To provide the methodological background of how we derive utility functions from point estimates of the experts, the following sections present a brief introduction to spline interpolation, Hyman filtering, and non-linear regression.

5.2.3.1 Spline Interpolation and Hyman Filtering to Interpolate Point Estimations

Given a set of data points $P = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $x_1 < x_2 < \cdots < x_n$, spline interpolation in general refers to the method of interpolating the data points by defining a polynomial for each neighboring pair of points $(x_i, y_i), (x_{i+1}, y_{i+1})$, $i \in 1, \ldots, n-1$. In the context of this thesis, monotonic cubic spline interpolation is of specific interest. In the following, we firstly present the cubic spline interpolation applied in this work. Hyman filtering ensures monotonic splines, and we explain it secondly.

Cubic Spline Interpolation

Cubic spline interpolation is a type of spline interpolation where each part of the spline is a third-degree polynomial. The utilized stats R package (R Core Team, 2023) is based on the calculation of splines presented in Forsythe et al. (1977, p.70–76). For each interval $[x_i, x_{i+1}]$, $i = 1, 2, \ldots, n-1$ the spline is represented by:

$$s(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \quad x_i \leq x \leq x_{i+1}.$$

For each of the $n-1$ sections between the data points, 4 parameters therefore need to be determined, leading to $4n-4$ parameters to be determined overall. At each of the $n-2$ interior data points (x_i, y_i), $2 \leq i \leq n-1$, the spline function s has to fulfill the following three conditions: s is continuous and has continuous first and second order derivatives. Additionally, $s(x_i) = y_i$ has to hold for each of the n data points. This gives $4n-6$ conditions, making it necessary to define two additional conditions in order to calculate the needed parameters. Forsythe et al. (1977) define the missing two conditions as follows: The unique cubic $c_1(x)$, which passes through the first four data points, and the unique cubic $c_n(x)$, which passes through the last four data points, are calculated. The third derivatives of $s(x)$ at the first and last data point are then matched to the third derivative of c_1 and c_n respectively, namely $s'''(x_1) = c_1'''$ and $s'''(x_n) = c_n'''$.
With these additional conditions, it is possible to formulate a system of linear equations and calculate the necessary parameters for the polynomials defined on each interval between adjacent data points.

Hyman Filtering

Calculating splines in the presented way ensures continuous and smooth, but not necessarily monotonic functions. If monotonicity of the spline is required, and if smoothness is not a requirement, Hyman-filtering can be applied (Hyman, 1983). For Hyman filtering, a spline is defined as

$$s(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \quad x_i \leq x \leq x_{i+1}$$

and

$$\begin{aligned} a_i &= y_i, & b_i &= \dot{f}_i, \\ c_i &= \tfrac{3S_{i+1/2} - \dot{f}_{i+1} - 2\dot{f}_i}{\Delta x_{i+1/2}}, & d_i &= -\tfrac{2S_{i+1/2} - \dot{f}_{i+1} - \dot{f}_i}{\Delta x_{i+1/2}^2} \end{aligned}$$

with:

- $\dot{f}_i$: a numerical approximation of the slope at x_i, $1 \leq i \leq n$. If a spline has already been calculated, the actual slope of the existing spline at x_i can be used.
- $S_{i+1/2} = \frac{\Delta y_{i+1/2}}{\Delta x_{i+1/2}}$
- $\Delta x_{i+1/2} = x_{i+1} - x_i$
- $\Delta y_{i+1/2} = y_{i+1} - y_i$.

Monotonicity can then be achieved by changing the values of $\dot{f}_i$ according to the following restraining functions and recalculating the parameters of the spline with the altered $\dot{f}_i$.

$$\dot{f}_i \leftarrow \begin{cases} \min[\max(0, \dot{f}_i), 3\min(|S_{i-1/2}|, |S_{i+1/2})], & \dot{f}_i \geq 0 \\ \max[\min(0, \dot{f}_i), -3\min(|S_{i-1/2}|, |S_{i+1/2})], & \dot{f}_i < 0 \end{cases}$$

Fig. 5.2 illustrates Hyman filtering effects by showing spline interpolation applied to five monotonically decreasing data points with and without filtering.

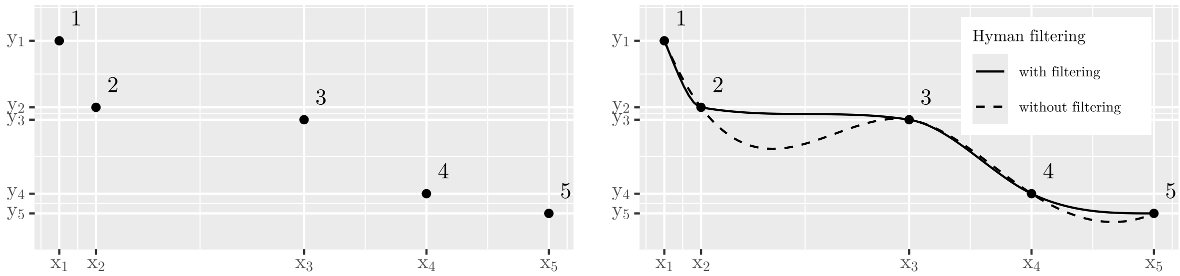


Figure 5.2: Example of cubic spline interpolation and Hyman filtering

5.2.3.2 Non-Linear Regression to Aggregate Interpolated Estimations

Spline interpolation is used to calculate continuous functions based on the point estimates of the experts. To arrive at utility functions, the splines then need to be aggregated to calculate the statistical group response. The fundamental idea is to aggregate the estimations through the central tendency of the estimations using point-wise statistical aggregation. For each minute, the mean of the values of the spline interpolations is calculated. To interpolate the resulting points, spline interpolation could be employed again. The downside of spline interpolation is that the resulting functions are cumbersome to work with in further computations due to their piecewise nature.

An alternative is using regression to fit a basis function to the aggregated points. This allows using a basis function with specific properties. As utility functions are a generalization of survival functions, which are calculated using logistic regression with the basis function $f(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$, we use a similar basis function. Preliminary experiments showed that increasing the degrees of freedom by one to $f(x) = \frac{1}{\beta_0+e^{-\beta_1+\beta_2 x}}$ led to utility functions that more closely resembled the form imagined by the experts in the consensus process. The non-linear model is fitted for each category using the minpack.lm R package (Elzhov et al., 2023), which is based on a modification of the Levenberg-Marquardt algorithm (Moré, 1978). At its core, the algorithm iteratively solves constrained linear least squares problems with local linearizations of the basis function to update the parameters of the non-linear basis function. The parameter values of the logarithmized regression model $\ln(y) = \ln(f(x))$ were used as a starting point for the algorithm.

5.3 Analyzing the Utility Functions for Heterogeneous Patients

The following section describes and analyzes the results of the consensus process. Section 5.3.1 describes the patient and call categories defined by the expert panel. Section 5.3.2 analyzes the utility functions and their properties. Section 5.3.3 analyzes historical utilities of patients in Baden-Württemberg based on mapping categories to diagnoses and keywords.

5.3.1 Describing the Patient and Call Categories

The medical experts defined six patient categories. Two main factors distinguish the categories. The first factor is whether it is possible to substantially improve the condition of the patient on site or whether the necessary treatment is only possible in a hospital. There are emergencies, such as strokes, where it is not possible to substantially improve the condition of the patient on site, for instance, if the necessary equipment for diagnosis and/or treatment is only available in a hospital. In this case, the prehospital time is more relevant than the response time from a patient's perspective. The second factor is how quickly the patient's condition deteriorates and how strong the correlation between time and outcome is. The discussion made clear that not all call responses of EMS result in a contact with a patient from one of these categories. To retrospectively categorize such call responses, two additional call response categories were defined. Table 5.2 lists the categories which are described subsequently.

Cat.	Description	Type	Relevant time
1	Life-threatening with the possibility of intervention on site.	patient category	response time
2	Life-threatening without the possibility of intervention on site.	patient category	prehospital time
3	Emergencies with promptly required therapy. Risk of damage to health cannot be ruled out. Possibility of intervention on site.	patient category	response time
4	Emergencies with promptly required therapy. Risk of damage to health cannot be ruled out. No possibility of intervention on site.	patient category	prehospital time
5	Progressive worsening of an existing medical condition.	patient category	response time
6	Psychosocial emergencies, socio-medical emergencies.	patient category	response time
7	Resource provision (e.g. fire, police situations, hazard alarm systems), without patient contact.	call response category	-
8	Patient transportation.	call response category	-

Table 5.2: Patient and call response categories

Category 1 includes patients who are acutely ill or injured in a life-threatening manner and for whom interventions carried out by the rescue team at the scene can lead to a decisive causal change in the course of the emergency. These include cardiovascular arrest, status epilepticus, or life-threatening trauma to the chest or extremities. For category 1 patients, time has a major influence on the utility. The relevant time period from a patient's perspective is primarily the response time.

Category 2 includes patients with acute life-threatening illnesses or injuries for whom the course of the illness cannot be significantly influenced by interventions at the scene. The primary quality objective for these patients is to stabilize them on site as required and to transfer them to a hospital for definitive treatment as quickly as possible. This category includes, for example, patients with stroke, ST-elevation myocardial infarction, or life-threatening abdominal trauma. For patients in category 2, time has a major influence on the utility. The relevant time period is primarily the prehospital time.

Category 3 includes patients who are not suffering from an acute life-threatening illness or injury, but who nevertheless require prompt treatment to counteract damage to their health. Patients in this category are also eligible for on-site intervention. The distinction to the higher priority categories is essentially based on the absence of life-threatening impairment of vital functions. This category includes, for example, exacerbated chronic obstructive pulmonary disease with spasticity (without severe respiratory distress, without pronounced signs of hypoxia) or hypoglycemia with stable vital signs (especially without unconsciousness). For patients in category 3, the relevant time period is the response time, but the time has less influence on the utility than in category 1.

Category 4 also describes patients without acute danger to life, but who require prompt intervention. However, the underlying illnesses and injuries in this category cannot be decisively influenced on-site. These include an acute abdomen without severe pain, a febrile infection with stable vital signs, or epistaxis. For this category, the relevant time period is the prehospital time, but the influence of time on the utility is less than for category 2.

Category 5 includes patients with a progressive worsening of an existing medical condition. For these patients, the focus is more on guiding the care pathway and providing care in an outpatient setting than on the fastest possible intervention at the scene or transportation to a hospital. The relevant time period for this category is the response time, but with less urgency than for categories 1 to 4.

Category 6 includes patients with psychosocial and socio-medical emergencies without immediate somatic danger. For these patients, EMS resource use is primarily justified because no alternative care system is available. The relevant time period for this category is the response time; the influence of this time on the utility is lower than in categories 1 to 4.

Category 7 describes call responses without patient contact. These include, for example, standby operations for police interventions or responses to fire alarm systems without patient contact. No link between time and utility can be defined, as there is no patient present during these operations.

Category 8 covers patient transports. If an emergency ambulance carries out such a transport, a resource of the emergency rescue is used without contact to an emergency patient. In addition to scheduled patient transports, this category covers patients without any condition that would fall into categories 1 to 6. Such patients could be cared for by a patient transport ambulance and its crew. Patient transport ambulances are equipped differently from emergency ambulances, and the crew of patient transport ambulances requires fewer qualifications than the crew of an emergency ambulance.

The categorization enables a differentiated interpretation of time intervals from a patient perspective by distinguishing between patients based on their medical urgency and the possibility of effective intervention on-site versus in hospital. Categorizing patients with respect to how different time intervals impact them during a call response is the first step to an improved measurement of the effectiveness of EMS logistics.

5.3.2 Analyzing the Utility Functions

The second step is to define utility functions for each patient category. The following sections show the estimated utility functions in Section 5.3.2.1 and investigate their properties in Section 5.3.2.2.

5.3.2.1 From Expert Estimations to Utility Functions

The utility functions quantify the importance of time for the patients. This links outcome quality to process quality and enables considering the differences in patients' needs in quantitative methods. The utility functions are based on the estimations of medical experts. Fig. 5.3 shows the *(time, utility)* estimations of one expert, including the splines based on the estimations for each of the six patient categories. It is apparent that the utilities for all patients decline with increasing time, but with varying slopes. Non-linear regression aggregates the five individual estimation splines as Fig. 5.4 shows. Table 5.3 lists the analytical formulations.

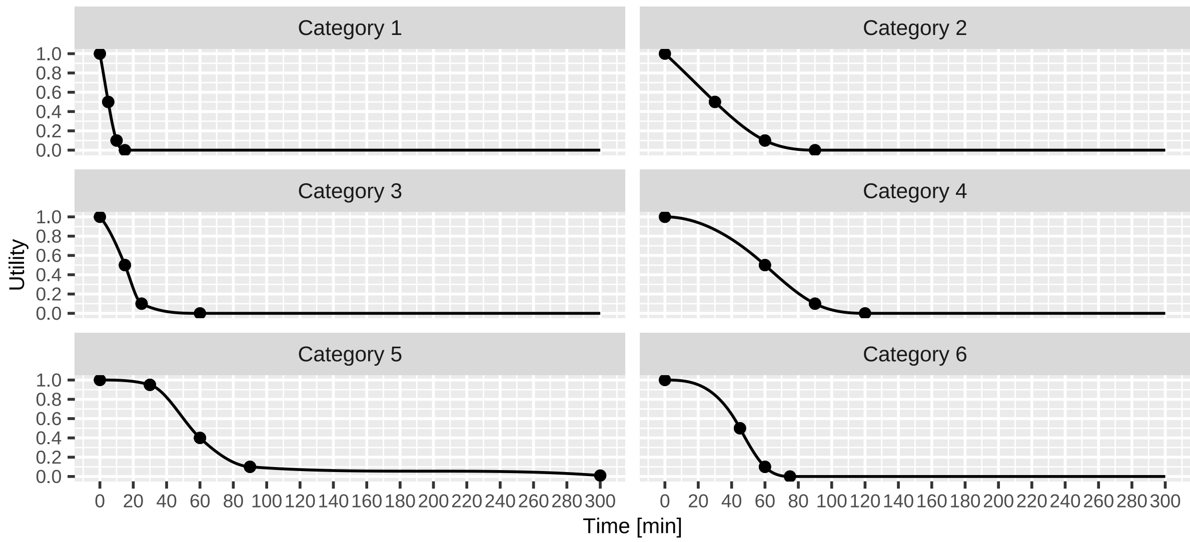


Figure 5.3: (*time*, *utility*) estimations (points) from one of the experts in the workshops and calculated splines (lines) for each patient category.

The panel submitted only four estimations for category 6 due to time constraints, but all workshop participants accepted the final utility functions. The underlying estimations of the experts vary significantly. The experts noted that patient needs differ within each category. Additionally, most medical conditions lack empirical evidence explicitly linking the time to treatment to outcomes, hindering quantitative analysis. Lastly, the concept of utility is abstract, and the individual interpretations of how a given utility value corresponds to experiences with treatments of patients can vary. All experts, however, agreed that the resulting utility functions sufficiently capture the differences between the patient categories.

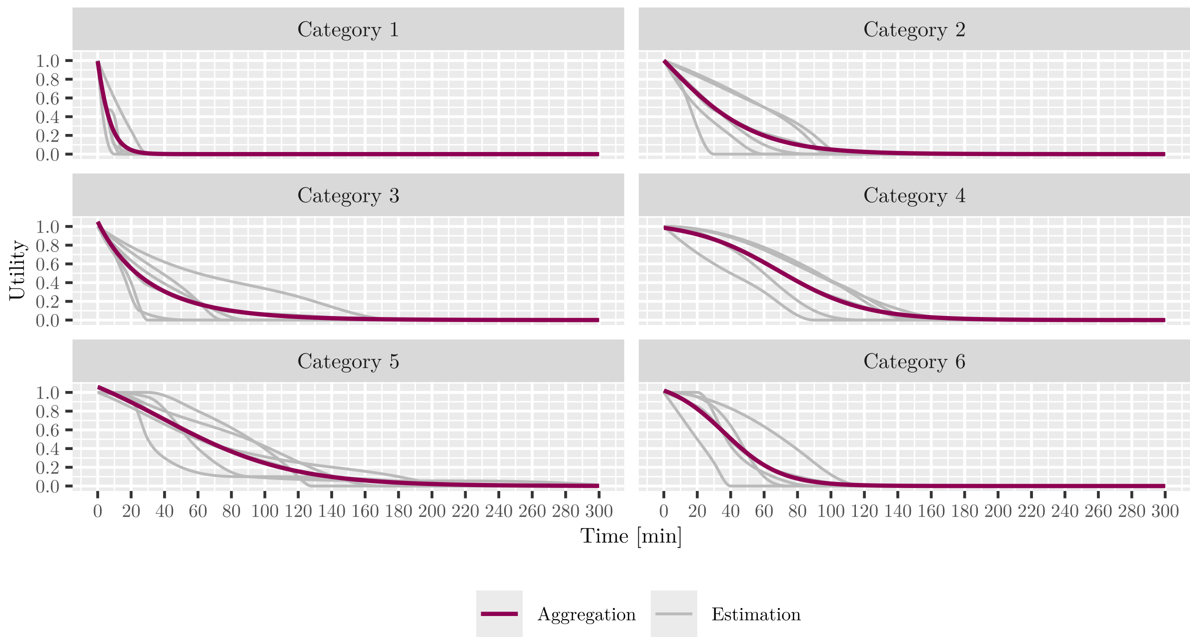


Figure 5.4: Aggregations of individual estimations to utility functions.

Fig. 5.5 shows the utility functions in direct comparison, separated by whether the relevant time from a patient perspective is the response time or the prehospital time. The utility functions exhibit distinct differences between categories. This supports the initial finding based on the patient categories that the needs of patients differ. Additionally, it shows that even though the individual estimations in Fig. 5.4 vary between the experts, the aggregated utility functions still capture differences between the categories.

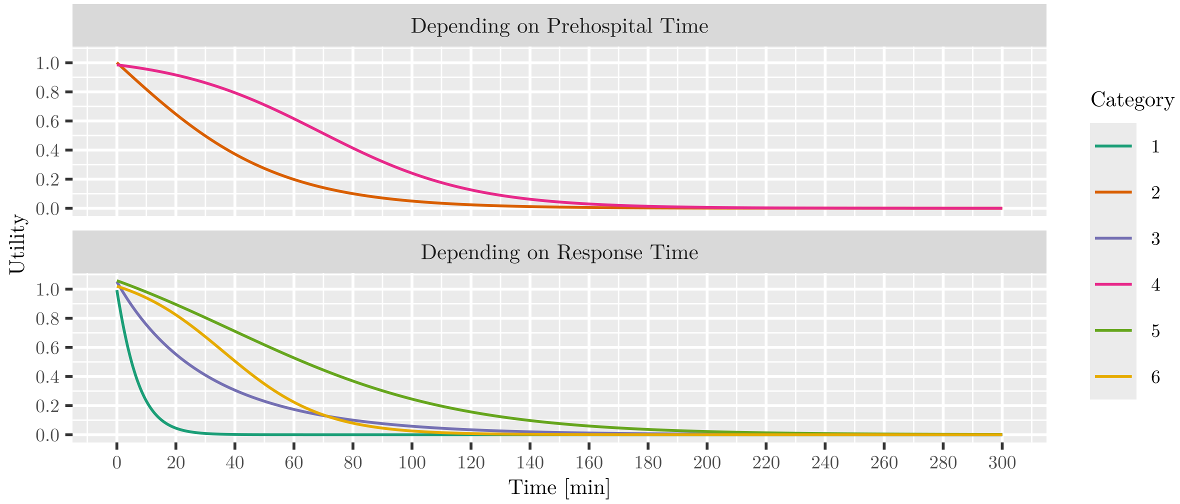


Figure 5.5: Overview of estimated utility functions.

Table 5.3: Utility functions for each patient category. Time t represents the response time for categories 1, 3, 5, and 6, and the prehospital time for categories 2 and 4.

Patient Category	Utility Function
Category 1	$\frac{1}{0.252+\exp(-0.281+0.167t)}$
Category 2	$\frac{1}{0.492+\exp(-0.679+0.037t)}$
Category 3	$\frac{1}{-0.266+\exp(0.197+0.027t)}$
Category 4	$\frac{1}{0.949+\exp(-2.722+0.039t)}$
Category 5	$\frac{1}{0.691+\exp(-1.370+0.026t)}$
Category 6	$\frac{1}{0.877+\exp(-2.271+0.059t)}$

5.3.2.2 Properties of the Utility Functions

The utility functions result from calculating the parameters a, b, and c of the basis function $f(time) = \frac{1}{a+\exp(b+c*time)}$ using non-linear regression for each patient category. $u_C(t) = \frac{1}{a_C+\exp(b_C+c_C t)}$, $C \in \{1, \ldots, 6\}$ denotes the utility function for patient category C in the following. Several observations can be made based on the formulations of the utility functions listed in Table 5.3.

Lemma 1. *Let $u_C : \mathbb{R} \to \mathbb{R}$ denote the utility function of patient category C with $C \in \{1, \ldots, 6\}$. Then it holds that $u_C(t) > 0$, $\forall C \in \{1, \ldots, 6\}$, $t \geq 0$*

Proof. Starting with $u_C(t) > 0 \Leftrightarrow a_C + \exp(b_C + c_C t) > 0$ and $\exp(x) > 0, \forall x \in \mathbb{R}$, two cases can be distinguished:

Case 1: If $a_C \geq 0$, it is obvious that $u_C(t) = \dfrac{\overbrace{1}^{>0}}{\underbrace{a_C}_{\geq 0} + \underbrace{\exp(b_C + c_C t)}_{>0}} > 0, \forall t \geq 0$. This is the case for $C \in \{1, 2, 4, 5, 6\}$ (s. Table 5.3).

Case 2: If $a_C < 0$, the following condition has to hold:

$$a_C + \exp(b_C + c_C t) > 0 \Leftrightarrow b_C + c_C t > \ln(-a_C).$$

$a_C < 0$ is the case for $C = 3$, with $a_3 = -0.266$, $b_3 = 0.197$, $c_3 = 0.027$, for which the condition holds for $t \geq 0$:

$$b_3 + c_3 t = 0.197 + \underbrace{0.027t}_{\geq 0} \geq 0.197 > \ln(0.266) = \ln(-a_3).$$

□

We can therefore conclude that the utility values of every patient category are greater than zero. This property can be relevant from a mathematical point of view, for instance, when aggregating individual utilities by multiplication as in Jagtenberg and Mason (2020b)

Lemma 2. *Let $u_C : \mathbb{R} \rightarrow \mathbb{R}$ denote the utility function of patient category C with $C \in \{1, \ldots, 6\}$. Then it holds that $u'_C(t) < 0, \forall C \in \{1, \ldots, 6\}, t \geq 0$.*

Proof. The derivative of $u_C(t) = \frac{1}{a_C + \exp(b_C + c_C t)}$ is $u'_C(t) = -c_C \dfrac{\overbrace{\exp(b_C + c_C t)}^{>0}}{\underbrace{(a_C + \exp(b_C + c_C t))^2}_{>0}}$. As $c_C > 0 \ \forall C \in \{1, \ldots, 6\}$ (s. Table 5.3), it holds that $u'_C(t) < 0, \ \forall C \in \{1, \ldots, 6\}, t \geq 0$. □

This implies that a change of the relevant time (response or prehospital time) will always lead to a change in utility for a given patient category. One drawback of the currently used RTTs is their binary nature, where a change in response times does not necessarily lead to a change in the fulfillment of a RTT. This drawback does not exist when using utility functions.

An additional observation is that the maximum utility is not 1. It is not directly ensured that $u_C(0) = 1 \ \forall C \in \{1, \ldots, 6\}$ due to the chosen basis function used in the regression to calculate the utility functions. The utility functions can be scaled by multiplication with the inverse of $u_C(0)$ if this property is desired for specific calculations. The scaling will not impact the previously shown properties, as this scaling factor is a positive number for all categories.

5.3.3 Analyzing Historical Patient Utilities in Baden-Württemberg

The expert panel conducted a mapping of the categories to medical diagnoses and keywords during the consensus process. The following sections firstly describe the diagnoses and keywords onto which the categories were mapped. The historical data described in Chapter 3 contains these diagnoses and keywords. This enables the retrospective categorization of emergency calls. Historical call volumes of the different categories, the predictive power of the categorization based on keywords, and historical utility values of patients are analyzed in the following, based on this mapping. The terms keyword-category and diagnosis-category will be used in the following for the ease of reading. The keyword-category of a call refers to the category based on the keyword, and the diagnosis-category to the category based on the diagnosis.

5.3.3.1 Mapping Categories to Diagnoses and Keywords

The Minimal Emergency Dataset MIND (German: Minimaler Notfalldatensatz) contains a defined list of diagnoses that ambulance crews and emergency physicians can document at the scene of the emergency and is authorized by the German Interdisciplinary Association for Intensive and Emergency Medicine DIVI (German: Deutsche Interdisziplinäre Vereinigung für Intensiv- und Notfallmedizin) (Deutsche Interdisziplinäre Vereinigung für Intensiv- und Notfallmedizin, n.d.). Each diagnosis on this list was mapped to a category. Some diagnoses possess a range of severities, which makes it impossible to map them to one distinct category. Additional vital parameters, such as pulse, blood pressure, or breathing frequency, were considered for those diagnoses. The values of these parameters, at which a diagnosis would switch category, were taken from the M-NACA system (Schlechtriemen et al., 2005). The M-NACA system classifies emergencies into seven categories based on vital parameters with defined threshold values for each vital parameter. One issue occurred during the mapping of historical call data to the categories. Based on the current dataset, it is not possible to distinguish between calls where no diagnosis was documented because there was no patient contact and calls where no diagnosis was documented because the patient did not show symptoms of a diagnosis contained in the MIND. The first case would lead to diagnosis-category 7, while the second case would lead to diagnosis-category 8. All such calls are categorized into category 8 for the following analyses.

The SQR-BW uses a list of defined keywords, onto which the individual keywords of each coordination center are mapped (SQR-BW, 2016). Each of these keywords was mapped to a category. Not every keyword could be mapped to one distinct category, similar to the diagnoses. The dispatched resources (with or without an emergency physician) and the use of sirens by the ambulance were taken into account for the mapping to enable the retrospective categorization.

5.3.3.2 Analyzing Patient Category Distributions

Examining the call volumes per diagnosis-category in Fig. 5.6 leads to several observations. The diagnosis-category with the highest call volume is category 4, containing patients without life-threatening symptoms and for whom the decisive treatment can only be performed in a hospital. A substantial number of calls have been classified as diagnosis-category 8. Data quality issues where no diagnosis was transmitted could also have contributed to the call volume in this category. Categories 1 and 2 contain ~ 22% of the call volume. This supports the argument that planning for EMS should not be based exclusively on patients with life-threatening symptoms.

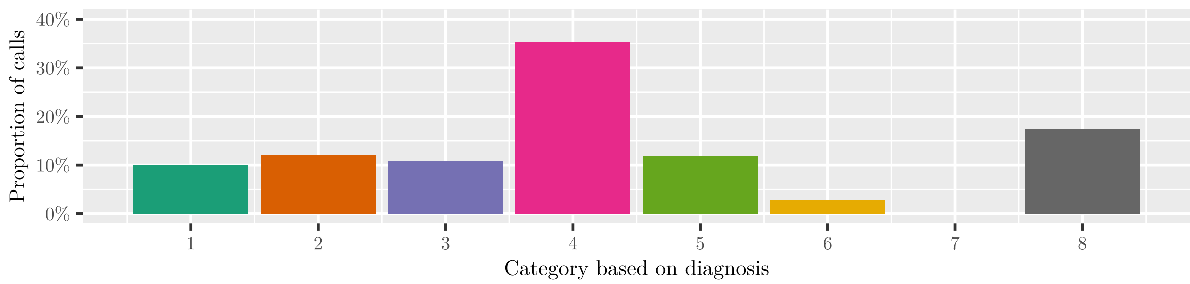


Figure 5.6: Overview of call proportions in each category based on the medical diagnosis. Based on combined data from 2021 and 2023.

Fig. 5.7 depicts the call volumes per keyword-category. The distribution of call volumes differs substantially from the call volumes per diagnosis-category. Keyword-categories 1 and 2 contain ~ 55% of calls. This indicates a tendency to overestimate the time criticality and severity of a call during the emergency call in the coordination centers. Categories 1 and 3 constitute the highest call volumes of the keyword-categories. Both categories have an emphasis on a timely arrival at the scene of the emergency. This might reflect that the main focus of coordination centers is on assigning emergency resources to emergency sites and less on guiding the whole patient pathway. The low volume in keyword-category 5 supports this observation further.

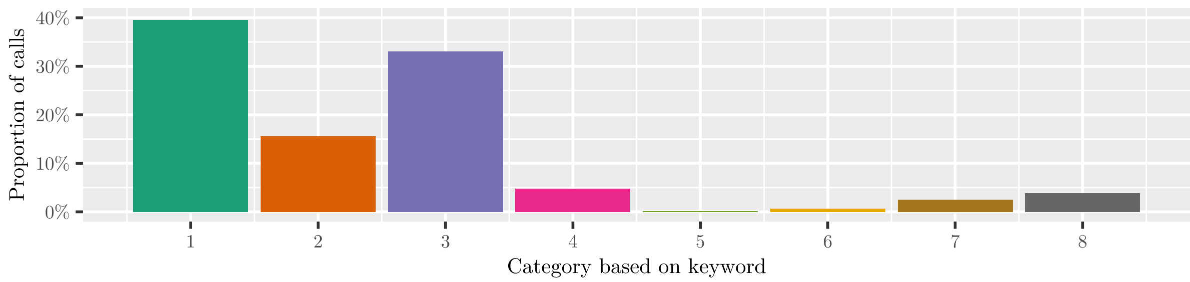


Figure 5.7: Overview of call proportions in each category based on the initial keyword. Based on combined data from 2021 and 2023.

5.3.3.3 Predicting Diagnosis-Categories with Keyword-Categories

It is important to know how well the keyword-category predicts the actual diagnosis-category, if the patient categories are to be used for operational decisions such as the dispatching decision in EMS. The diagnosis-category can be seen as the “true” category of a call and the keyword-category as a prediction of this “true” category. This perspective is closely related to classification problems encountered, for instance, in the field of machine learning. A confusion matrix can be used to analyze the performance of classification algorithms. A confusion matrix displays how often which combination of predicted and true classes occurred. Such a confusion matrix can be used to display how often which combination of keyword- and diagnosis-category occurs in the historical call data (s. Fig. 5.8).

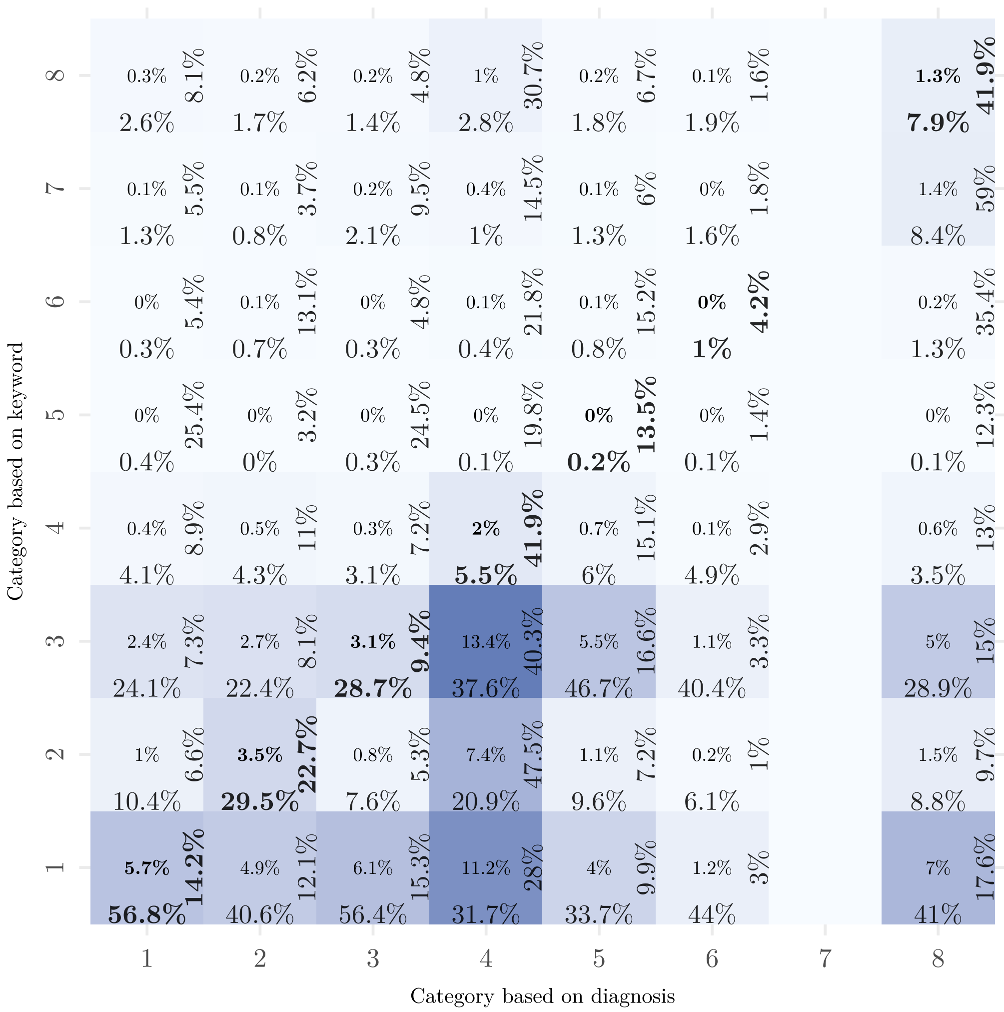


Figure 5.8: Confusion Matrix of categories based on keywords and based on diagnoses. Exemplary explanation of depicted percentages for the bottom left cell: 5.7% of all calls are of keyword-category 1 as well as diagnosis-category 1. 56.8% of the calls in diagnosis-category 1 are also in keyword-category 1. 14.2% of the calls in keyword-category 1 are also in diagnosis-category 1. Based on combined data from 2021 and 2023.

Overall, only 15.7% of calls show an identical keyword- and diagnosis category. The primary concern might be to correctly detect all life-threatening situations classified as categories 1 and 2. The confusion matrix shows that 56.8% of diagnosis-category 1 calls have been classified as keyword-category 1, and 29.5% of diagnosis-category 2 patients have been classified as keyword-category 2. To conclude, the keyword-category poorly predicts the diagnosis-category. This needs to be considered when investigating how the keyword- and diagnosis-categories can inform logistic decisions.

5.3.3.4 Analyzing Historical Utility Values of Patient Categories

The utilities resulting from the historical call data can be analyzed in order to examine whether there exists a potential to improve patient utilities. Fig. 5.9 shows the box plots of the utilities of historical calls based on the diagnosis-category. The utility value has been calculated for each historical call using the respective utility function (Table 5.3) and the historic response or prehospital time of the call response. The historic call responses led to low utilities for categories 1 and 2, medium to high utilities for categories 3 and 4, and high utilities for categories 5 and 6. Especially the low utilities for patients in categories 1 and 2 lead to the conclusion that there is potential to improve patient utilities.

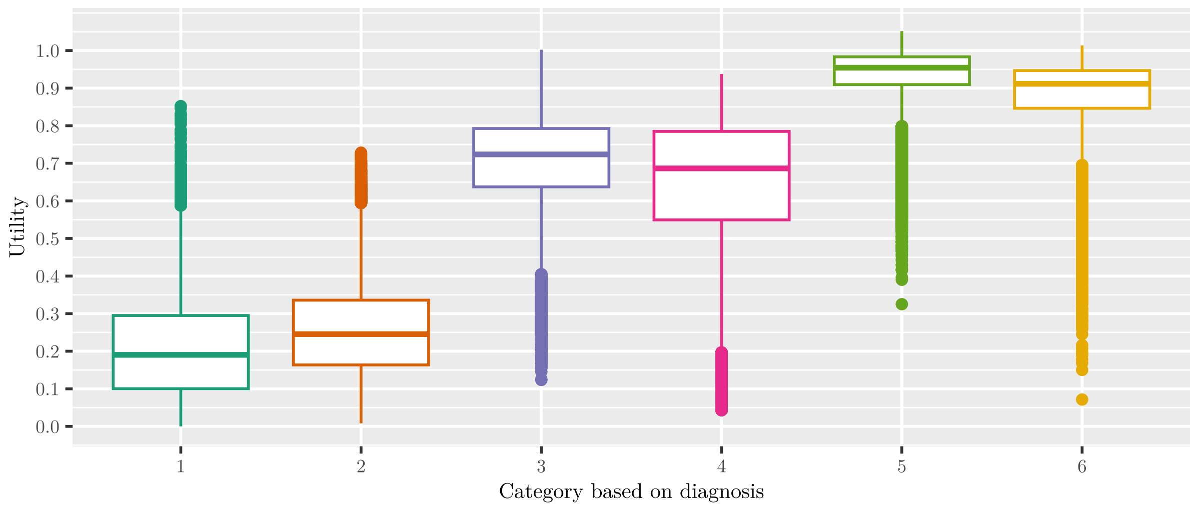


Figure 5.9: Comparison of historical utilities of patient categories. Based on combined data from 2021 and 2023.

5.4 Utility Functions for Heterogeneous Patients: Discussion

This chapter aimed to investigate whether and how the concept of survival functions can be generalized to apply to the heterogeneous patients in EMS. We conducted an expert consensus process that resulted in utility functions for patient categories, and discuss our findings and approach in the following.

Contributions

The developed patient categories capture the heterogeneity of patients encountered in EMS and demonstrate that a differentiated approach is necessary to reflect the different needs of patients in quality indicators. While there are existing categorizations of patients in EMS systems, these categorizations are based exclusively on response time targets. The developed patient categories expand the perspective to prehospital times and show that an exclusive focus on response times is not reflective of the needs of all patients. The successful mapping of diagnoses and keywords to the patient categories indicates that the categories can capture the range of emergencies.

The utility functions present a novel way to quantify the effects of time on patient outcomes for the heterogeneous patients encountered in EMS. They are a generalization of existing survival functions, which are based exclusively on patients with OHCA. While survival functions are derived directly from medical outcome data, the utility functions are based on expert judgments due to the lack of available outcome data for the entire range of emergency patients. The variations in the estimations of the utility functions within each category, on the one hand, show that different perspectives were taken into consideration in the design of the utility functions. On the other hand, the differences demonstrate that more empirical evidence is needed to support the design of utility functions.

The lack of empirical outcome data hinders the quantitative testing of the hypothesis that the utility functions do, in fact, link time to outcome quality. The response time targets currently used in EMS systems are, however, also not directly derived from medical outcome data. The utility functions, therefore, still present a more patient-centered approach to evaluating the process quality of EMS systems, as they resolve the drawback of the binary nature of response time targets.

The utilities of the patients can be seen as a novel quality measure representing the dimension of *effectiveness* in the context of the framework presented in Section 4.3. This contributes to closing the research gap identified in Chapter 4. An individual utility value can be calculated for each patient using the utility functions. The individual utilities can then be aggregated using different statistics, depending on whether the focus is on equity or efficiency. The question of which statistic should be used has not been answered by the presented work.

Practical Implications

Several practical implications can be deduced from our results. The current response time targets used in EMS systems around the world do not reflect the heterogeneous needs, and the quality of EMS systems should take the heterogeneity of patients' needs into account. The presented research, including the patient categories and utility functions, was published by Watzinger et al. (2024) as part of a research project funded by the Ministry of the Interior of Baden-Württemberg. An amendment of the Emergency Medical Services Law in Baden-Württemberg (German: Rettungsdienstgesetz (RDG), in Kraft getreten am 02.08.2024) was introduced by the Ministry of the Interior in Baden-Württemberg, which differentiates between planning targets based on response times for some patient categories and planning targets based on the prehospital time for other patient categories.

Limitations

The following limitations should be taken into account when interpreting the results. The panel size of experts was limited. Five physicians contributed to estimating the utility functions, and the resulting estimations show considerable variability between the experts. While the final utility functions were accepted by all experts, including the additional four non-physician experts, a larger panel size and a larger number of estimations could increase the validity of the results. The patient categories were developed in an exploratory manner in an open discussion. This deviates from the requirement for anonymity in a consensus process. The reason for requiring anonymity is to avoid the dominance of individual experts in the consensus process. A second group of experts was involved to independently validate the patient categories proposed by the first group to mitigate this risk. Furthermore, the mapping of emergency diagnoses to the categories demonstrates that each diagnosis can be assigned to exactly one category, and each category encompasses at least one diagnosis. This indicates that the patient categories can capture the range of emergency conditions. The main reason for developing the patient categories in an open discussion was to ensure that all participating experts had a shared understanding of the categories. This shared understanding was necessary to estimate the utility functions and map the diagnoses to categories.

The mapping of keywords is based on the keywords used by the SQR-BW to make the individual keywords used by different coordination centers comparable. These keywords were chosen in order to enable the categorization of historical call data from various coordination centers. A downside of this approach is that there can already be a loss of information in the mapping of the individual keywords of coordination centers to the standardized keywords.

The mapping of diagnoses is based on the diagnoses defined in the Minimal Emergency Dataset MIND (Deutsche Interdisziplinäre Vereinigung für Intensiv- und Notfallmedizin, n.d.) and threshold values for vital parameters from the M-NACA classification (Schlechtriemen et al., 2005). These diagnoses are used in the German EMS and facilitated the mapping of historical call data to the patient categories, but the transferability to EMS outside of Germany might be limited.

Methodological Reflections

We made several methodological choices on which we can reflect. We employed splines to interpolate point estimations from the expert panel. The splines were then aggregated using pointwise aggregation and non-linear regression. This constitutes a pragmatic approach that resulted in utility functions with similar mathematical properties to existing survival

functions, which should facilitate integrating utility functions into existing modeling approaches. A downside of this approach is that the maximum of the utility functions is not equal to one. Investigating alternatives both to interpolating points and to interpolating functions — splines in our case — could lead to utility functions with different properties.
Our expert consensus process resulted in six patient and two additional call categories. This enables a more differentiated evaluation compared to the singular response time target that was the main regulatory planning target at the time of conducting this research. Six patient categories are, however, still few categories considering the broad spectrum of emergencies. One consequence of the category sizes is seen in the variability of expert estimations within each category. An alternative approach would be to define more, but narrower, patient categories,s which could enable a more reliable estimation of utility functions for each category. These detailed categories could then still be aggregated, if limiting the number of categories is a priority. This would improvethe control of the loss of information when aggregating to broader categories.
The utilities of our patient categories depend either on the response or on the prehospital time. There are, however, emergencies that benefit from both a short response time and a short prehospital time. Integrating the possibility to consider both the response and the prehospital time for calculating utilities could strengthen the correlation between utility values and medical outcomes.

Outlook

The presented work provides several avenues for future research. The availability of medical outcome data for every emergency patient would enable basing the presented approach on empirical data instead of expert opinion, which would improve the validity of the results. Patient categories could, for instance, be developed using clustering methods based on outcome data. Similarly, the utility functions could be derived directly from outcome data using regression models.
Applying the approach to different EMS systems to assess the generalizability of the developed patient categories and utility functions presents a further possibility for future research. Similarly, the consensus mechanism could be repeated using a different and larger expert panel to investigate the robustness of the results.
The proposed utility functions present a novel way for a measure that can be computed for each call response. However, the question of which statistic should be used to aggregate individual utility values into a quality indicator remains unexplored. Future research could investigate which statistics are suitable to aggregate individual utility values into quality indicators. Additionally, only a few statistics within the equity dimension of the framework presented in Chapter 4 have been identified in the literature. Developing new statistics to quantify equity in EMS systems could enable a better quantification of the equity dimension of quality in EMS with respect to utility values, but could also be used to aggregate measures reflecting the resource perspective.

Based on the conducted mapping and data analysis, a major challenge for a practical implementation of patient categories is correctly classifying patients in the initial emergency call (s. Fig. 5.8). Investigating which questions are most suitable to quickly and accurately classify patients in the initial emergency call presents a further avenue for future research.

5.5 Utility Functions for Heterogeneous Patients: Conclusion

This chapter set out to contribute to closing the research gap in measuring the effectiveness of EMS logistics. A consensus process was designed and implemented to define patient categories and utility functions. The expert panel in the consensus process defined six patient categories and two call categories. Utility functions were estimated for the six patient categories using point estimations from the expert panel, spline interpolation with Hyman filtering to interpolate the estimations, and non-linear regression to aggregate the interpolations. The results demonstrate that the needs of patients differ significantly and that current response time targets do not fully reflect how patients are impacted by the results of logistic decisions in EMS. Utility functions are a generalization of existing survival functions and allow an evaluation of the time intervals within an emergency call response from the perspective of heterogeneous patients.

Part III

Simulating EMS System Behavior

Summary of Part III

Research Question and Objectives

Research question: How can the effects of logistic decisions on the quality of EMS be predicted?
Research objectives:

- Develop a generic concept model for EMS systems
- Implement an executable simulation model in AnyLogic
- Validate the simulation model using real-world data from Baden-Württemberg

Scope and Methodology

While the developed concept model aims to be sufficiently generic to be able to model the different resources of both emergency rescue and patient transportation operations, the implemented simulation model and the verification are tailored to ground-based emergency rescue operations with ambulances and emergency physician vehicles in Baden-Württemberg, Germany, as shown in Fig. 5.10.

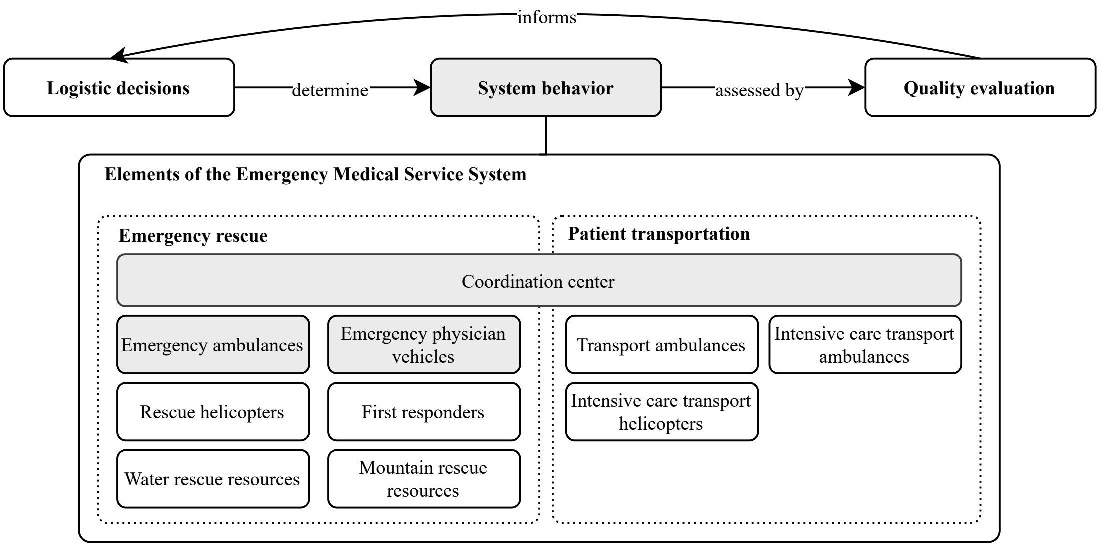


Figure 5.10: Part III investigates how the behavior of EMS systems can be modeled to investigate the effects of logistical decisions.

The concept model consists of a *state chart* for modeling resources and *process flow charts* for modeling patient pathways and handling requests in coordination centers. The simulation model combines *agent-based* and *discrete-event* modeling and is implemented in AnyLogic. We verify and validate the model using mainly *historical data validation.*

Main Contributions

- A generic concept model for EMS systems modeling emergency rescue and patient transport operations with an arbitrary number of involved resources

- Coordination concepts of task leaders and task backlogs, enabling coordination of multiple resources and integration of scheduled and ad hoc tasks
- An executable simulation model in AnyLogic combining agent-based and discrete-event modeling paradigms
- Validation of the simulation model for ground-based emergency rescue operations in Baden-Württemberg, Germany, with real-world data

Key Findings

- Simulation models combining agent-based and discrete-event paradigms are suitable for studying the effects of logistic decisions on EMS system behavior.
- Simulation models tend to be designed for a specific EMS system, hindering the transferability of models between EMS systems.
- Our simulation model closely resembles historical system behavior for emergency rescue operations in Baden-Württemberg, and additional information and data regarding the choice of transport destinations could further enhance the applicability of the model.

Chapter 6

Simulating EMS

Part II proposed a novel approach to evaluating quality in EMS. This leads to the question of how logistic decisions impact the quality of EMS. To answer this question, we require a way to study how logistic decisions impact the behavior of EMS systems.
According to Law (2015), several ways of studying a system exist. Fundamentally, a system can be studied through experiments where controlled changes are performed on the system, and the behavior of the system is observed and analyzed. Fig. 6.1 depicts the different ways to study a system.

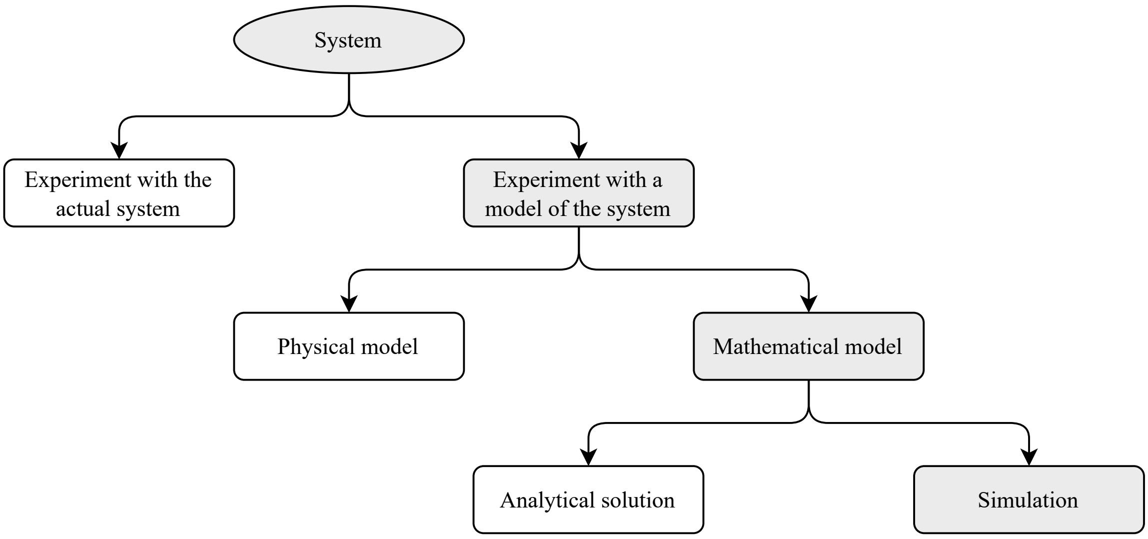


Figure 6.1: Ways to study a system (Law, 2015)

The first differentiation is whether experiments are conducted with the actual system or with a model of the system. The behavior of EMS systems directly impacts the health and well-being of patients. Experiments with the actual system should therefore be avoided if an alternative exists, and it appears suitable to conduct experiments with a model of a system. EMS additionally operate in an uncontrolled environment, which makes it difficult to perform controlled experiments in the actual system. The focus of the experiments lies on the actions and interactions of elements of EMS systems over time and not on physical properties, making a mathematical model preferable to a physical model. EMS systems are complex systems with dynamic interdependencies, for instance, regarding ambulance availabilities, and stochastic variables such as process times, call locations, or call volumes. Simulation models are suitable to capture both the stochastic and the dynamic nature of EMS systems and are a common tool in studying logistic aspects of EMS systems (Aboueljinane et al., 2013). These reasons lead us to employing simulation to predict the impact of logistic decisions on EMS system behavior.

Simulation models are typically tailored to specific EMS systems, and the development of reusable general-purpose simulation models has been identified as a challenge by Aringhieri et al. (2017). We intend to use the simulation model in this thesis to study the impact of logistic decisions using our application case of the EMS in Baden-Württemberg, Germany. The lack of ready-to-use simulation models requires us to develop and validate a custom simulation model. We additionally aim to adress the lack of reusable models by proposing a generic and adaptable modeling approach, which we subsequently tailor, implement, and validate for the EMS in Baden-Württemberg, Germany.

Section 6.1 reviews existing literature on simulation models of EMS systems to situate our work. Section 6.2 presents the concept model as well as the implementation of our simulation model. We validate the model using real-world data from the EMS in Baden-Württemberg, Germany, in Section 6.3, followed by discussing our approach and findings in Section 6.4. Lastly, Section 6.5 concludes the chapter.

6.1 Related Literature on Simulating EMS

Most simulation models focus on emergency rescues and the behavior of ambulances as the main resource. One exception is Kergosien et al. (2015), who develop a model considering both emergency rescues and patient transports, but still consider only ambulances as resources. Olave-Rojas and Nickel (2021) conclude that existing models have been designed for specific aspects of EMS systems and propose a model which includes coordination centers, ambulances and emergency physician vehicles (EPV). A limitation of the model is the assumption that a maximum of one ambulance and one EPV can attend an emergency. To compare different mathematical optimization models for the ambulance station location problem, Jánošíková et al. (2021) employ a simulation model considering ambulances as resources. One of the few open-source decision support systems for EMS systems has been published by Ridler et al. (2022). The model considers ambulances as resources, and a comparison of ambulance deployment strategies demonstrates its value. Strauss et al. (2021) model multiple resources for emergency rescues with ambulances, EPV, and multicopters. The simulation is used to compare ambulance allocations and station locations in Switzerland. To evaluate a novel optimization model for ambulance location and allocation, Grot et al. (2022) develop a simulation model considering ambulances as the only resource. Extending the scope beyond EMS, Aboueljinane and Frichi (2022) model both ambulances and emergency departments as well as observation units in their simulation. A novel direction for the application of simulation models in EMS is their use to train metamodels as demonstrated by Snyder and Smucker (2022). Their simulation model is again limited to ambulances as resources. To investigate the decision to dynamically relocate ambulances, Strauß et al. (2022) employ a simulation model considering ambulances. Two different ambulance types next to rapid response vehicles for physicians are modeled by Da Ros et al. (2024). Their decision support system aims to evaluate optimization approaches for resource location and personnel rosters.

We propose a concept model designed to model both emergency rescue and patient transport operations in the following. Additionally, we propose concepts designed to coordinate an arbitrary number of resources involved in one call response. The contribution of our approach is the capability to model both emergency rescue and patient transport operations with an arbitrary number of resources involved in one operation. We implement the concept model as an executable simulation model in AnyLogic based on the proposed concept model. The implementation and validation are tailored to emergency rescue operations in the EMS in Baden-Württemberg, Germany, and we consider ambulances and EPV as resources.

6.2 A Generic and Adaptable Simulation Modeling Approach for EMS Systems

We develop a concept model in the following Section 6.2.1 and present the implementation as an executable simulation model subsequently in Section 6.2.2. The concept model aims to be sufficiently generic and adaptable to model different EMS system designs, for instance, with separate or common resources for emergency rescue and patient transportation. We focus on the ambulance dispatching and the ambulance station location problem and study the EMS system in Baden-Württemberg, Germany, in the remainder of this thesis. The implementation of the concept model is therefore tailored to this purpose.

6.2.1 Designing the Concept Model

A concept model aims to model the system under study independently of the implementation of the simulation model, for instance, with a specific simulation software or programming language. A system is characterized by its entities and their actions and interactions. The main entities of an EMS system are its *resources*, such as ambulances or EPV, the *patients*, and *coordination centers*. We model the actions of *resources* using state charts and the actions of *patients* and *coordination centers* using process flow charts.

6.2.1.1 State chart and coordination concepts for EMS resources

Resources are a permanent part of the model and repeatedly fulfill tasks. State charts are suitable for modeling these characteristics. A state chart models the behavior of an agent by defining states and transitions between states. The state chart in Fig. 6.2 displays the states and transitions of EMS resources.

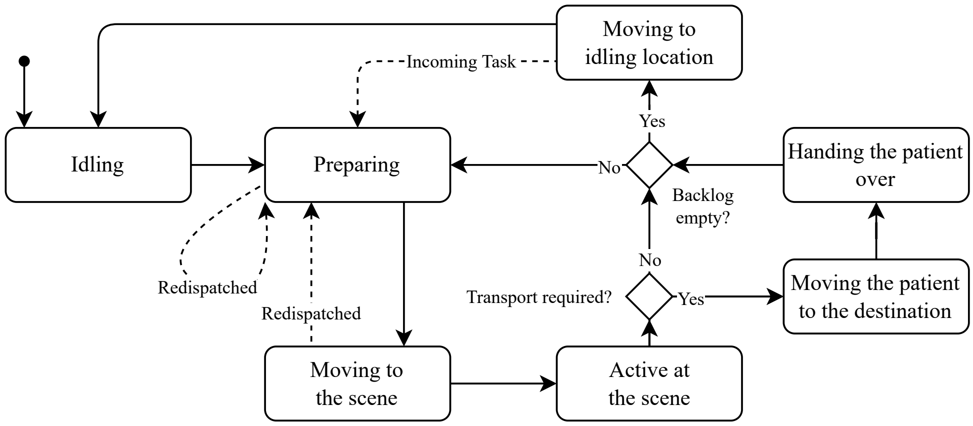


Figure 6.2: State chart of EMS resources.

While *idling*, resources are available to be dispatched to a new task. Once resources are dispatched, they start *preparing* for the task. While *preparing*, resources can be *redispatched* to a different task, for instance, to one with a higher priority. After *preparing*, resources start *moving to the scene*. While moving to the scene, they can again be *redispatched* to a different task. Once resources arrive at the scene, they are *active at the scene*. In the case of an emergency rescue, this can include first-aid measures and stabilizing the patient for transport. In the case of patient transportation, this state represents picking up the patient. Depending on the task and patient, a transport can be required. In that case, the resources performing or accompanying the transport start *moving the patient to the destination*. Once the resources arrive at the destination, they start *handing the patient over*, for instance, at an emergency department. After having handed over the patient or, if no transport is necessary, after finishing the task at the scene, two possibilities exist. If resources have already been dispatched to the next task, for instance, shortly before finishing handing over the patient or in case of scheduled patient transports, they immediately start *preparing* for the next task. If they have not already been dispatched, they start *moving to idling location*. While moving to their idling location, resources can be dispatched to new tasks, in which case they would start *preparing* again.

Two challenges emerge when developing a model architecture that is supposed to apply to different EMS system designs: How can an arbitrary number of resources of arbitrary types be coordinated during tasks? How can ad-hoc and scheduled tasks be considered in one modeling logic? To tackle these challenges, we introduce the concepts of *task leader* and *task backlog* as depicted in Fig. 6.3.

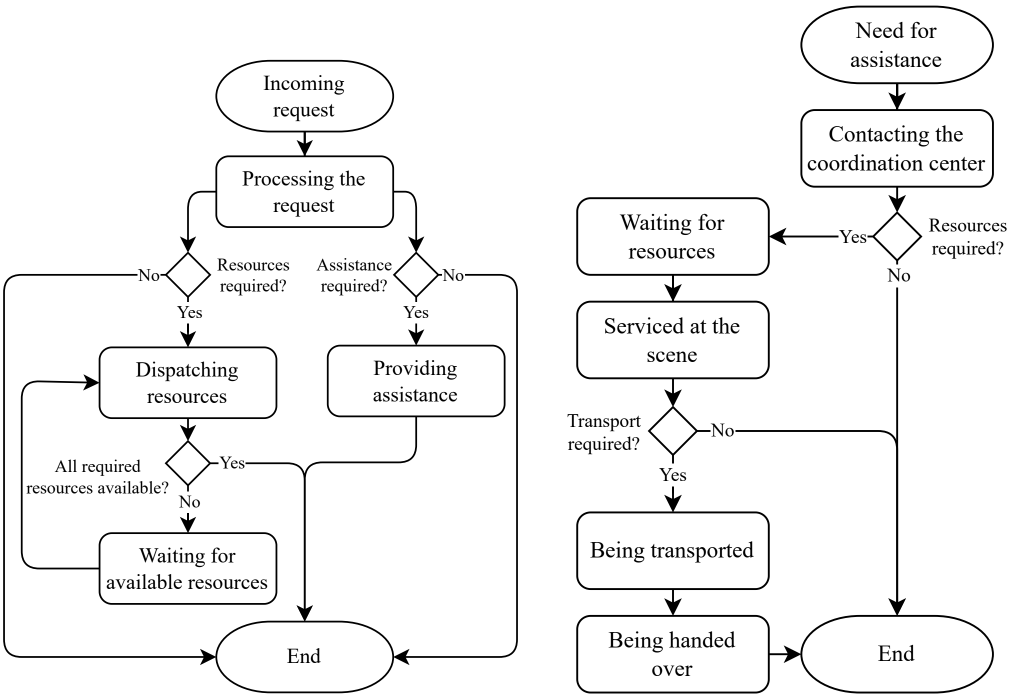
Figure 6.3: Schematic visualization of the process control concepts of task backlogs and task leaders.

To solve the first challenge, the concept of a *task leader* is introduced. For every task, exactly one EMS resource is defined as the *task leader*. This *task leader* is responsible for coordinating all involved resources in the task. This mainly includes triggering leaving the states *active at the scene* and *handing patient over* in case of transports with multiple involved resources. The role of *task leader* can change during a task, for instance, if additional resources are dispatched after the first resources arrive at the scene. This concept allows for an arbitrary number of resources to be involved and coordinated in a task. By adapting the behavior of the *task leader* to specific EMS systems, the interplay between resources can be specified without having to change the presented modeling architecture.

To solve the second challenge, the concept of *task backlogs* is introduced. Each resource contains a *task backlog*. In case of scheduled patient transports, the transports can be stored in the *task backlog*, including, for instance, information about specified time windows. In case of an emergency rescue, the concept of a *task backlog* allows dispatching ambulances while they are technically not yet idle, for instance, shortly before they finish handing over a patient at the transport destination, which does occur in practice. By defining how tasks in the *task backlog* are handled, both scheduled and ad-hoc tasks can be included without the need to adapt the model architecture itself.

6.2.1.2 Process flow charts of coordination centers and patients

The requests handled by coordination centers enter and leave the model and are processed once. Process flow charts are suitable for capturing these characteristics. Coordination centers are modeled by the process flow chart depicted in Fig. 6.4a. Coordination centers handle requests that can be initial requests, such as emergency calls via telephone, eCalls from vehicles, or inquiries for patient transports, among others. A request can also be a follow-up request, for instance, the request for additional resources, or if the coordination center is involved in deciding the transport destination for a patient (van Buuren et al., 2017). The first step following an *incoming request* is *processing the request*. This includes extracting information from the request and transferring the information into the IT system of the coordination center. After processing the request, the control center needs to decide whether EMS resources are required. If they are required, the next step is *dispatching resources*. It can occur that not all required resources are available at the time of dispatch. In this case, the request gets queued and will be *waiting for available resources* until all required resources are dispatched. After *processing the request*, further assistance by the coordination center might be necessary, for instance, instructing first-aid measures. Depending on the call-center design, there is one role in the call-center that handles the complete request, including the dispatching and assistance, or there are separate roles, where one person is responsible for the initial processing and for assisting, while a second person dispatches the required resources (van Buuren et al., 2017). Both variants can be modeled by the presented flow chart.



(a) Flow chart of coordination centers (b) Flow chart of patients

Figure 6.4: Flow charts of coordination centers and patients.

In order to collect the necessary measurements to calculate different performance metrics, the patient pathway can be modeled explicitly as presented in Fig. 6.4b. Similar to requests, patients enter and leave the model and are processed once, making process flow charts suitable. It begins with the occurrence of some perceived *need for assistance*. This need is communicated by *contacting the coordination center*, which in turn starts the process within the coordination center. If no resources are required by the patient, the pathway ends after communicating with the coordination center. If resources are required, the patient will be *waiting for resources* until they arrive. Once the resources arrive, the patient is *serviced at the scene*. This process step starts once the first resource starts being *active at the scene*. If no transport is required, the patient pathway ends after all necessary activities have been performed at the scene. If the patient requires a transport, the patient is *being transported* and then *being handed over* at the transport destination. In this case, the part of the patient pathway in which EMS is involved ends, and the patient pathway may continue in a subsequent part of the healthcare system. Explicitly modeling the patient pathway allows for collecting timestamps to calculate different performance metrics independently of, for instance, the number and types of resources involved.

In summary, the purpose of the proposed architecture is to model the range of tasks and resources present in EMS systems. Explicitly modeling the patient pathway enables the collection of necessary measures to compute different performance indicators. The model architecture is a basis for the development of executable simulation models, with which the behavior of EMS systems under different logistic decisions on the strategic, tactical, and operational levels can be studied.

6.2.2 Implementing the Concept Model for the EMS in Baden-Württemberg

We need to implement the concept model as an executable simulation model in order to conduct numerical experiments. This necessitates selecting software for the implementation. We implement the model using AnyLogic [1]. AnyLogic is a commercial simulation software tool with the functionality of combining agent-based, discrete-event, and system dynamics modeling, and we chose it for the following reasons. It is Java-based and enables implementing custom Java code, which facilitates customizing any aspect of the simulation model. At the same time, it offers GIS capability, animations, and core simulation features such as managing the simulation time, which reduces the implementation effort. In the following, we present the implementation of a state chart for EMS resources and the process flow charts for patients and coordination centers. We additionally present the GIS-environment used for modeling the EMS resource travels and animations used for validating the model. Calculating routes along the street network is the operation that requires the most computation time when running the simulation model. Routes need to be calculated mainly during the dispatching decision to identify the closest available resource. To reduce computational effort, we propose an algorithm to identify the closest available resource, which aims to reduces the number of route calculations compared to the basic AnyLogic functionality.

6.2.2.1 State chart of EMS resources

We implement the state chart modeling the behavior of EMS resources (Fig. 6.2) using the agent-based modeling capability of AnyLogic as depicted in Fig. 6.5. Firstly, we describe the parts of the state chart that are a direct implementation of Fig. 6.2, and secondly, we explain the necessary adaptations due to the technical requirements of AnyLogic.

[1] https://www.anylogic.com/

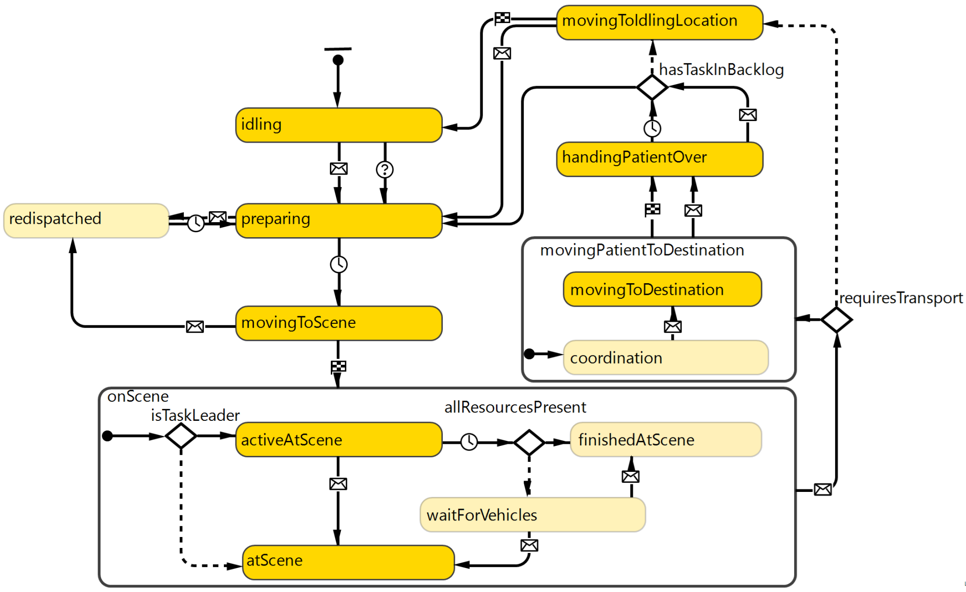


Figure 6.5: Agent-based model of EMS resources in AnyLogic.

Idling, preparing, movingToScene, movingToDestination, handingPatientOver and *movingToIdlingLocation* are direct implementations of the corresponding states in Fig. 6.2 and model the same parts of the process as described in Section 6.2.1.1. *RequiresTransport* corresponds to the *Transport required?*-decision, and *hasTaskInBacklog* to the *Backlog empty?*-decision in Fig. 6.2.

The states *redispatched, finishedAtScene, waitForVehicles* and *coordination* displayed in light yellow do not represent actual behavior of the resources but are required to trigger necessary function calls and to ensure that the resources in the simulation model are synchronized in their state transitions. *IsTaskLeader* differentiates whether a resource is the *task leader* for the current task. If so, it will transition to the state *activeAtScene* until it either finishes the necessary activities at the scene or the role of *task leader* is assigned to another resource. In this case, or if the resource was not the *task leader* at *isTaskLeader*, it will transition to *atScene* until the *task leader* triggers the transition to the next state. *AllResourcesPresent* checks whether all required resources are at the scene after the *task leader* finishes its activities, or if it is necessary to wait for further resources. This might occur, for instance, if an EPV has finished all first-aid measures, but has to wait for an ambulance to conduct the transport.

6.2.2.2 Process flow charts of patients and coordination center requests

Fig. 6.6 depicts the implementation of the process flow chart of the patient pathway (Fig. 6.4b). Patients start the process in *needForAssistance*. We combine the steps *contacting coordination center, waiting for resources* and *serviced at the scene* into *waitingForResources* to simplify the simulation model. The focus of the model lies on the EMS resources. It is therefore feasible to not simulate any patients without a need for resources, making the *resources required?*-decision in Fig. 6.4b obsolete. There is additionally no

need to differentiate between the steps *contacting the coordination center, waiting for resources* and *serviced at the scene* of the concept model, as it is still feasible to collect all measures for different performance metrics in the implemented model.
If no transport is required, the patient will flow from *exit* to *enterNoTransport* and conclude its process in *sink*. If transport is required, the process of the patient continues in the resource agent conducting the transport, flowing from *exit* to *beingTransported*. Similarly to the beginning of the process, we can collect all necessary measures without explicitly differentiating between the steps *being transported* and *being handed over* in the implementation. After being handed over, the patient flows from *exitHandover* to *enterHandover* and concludes its process in *sink*.

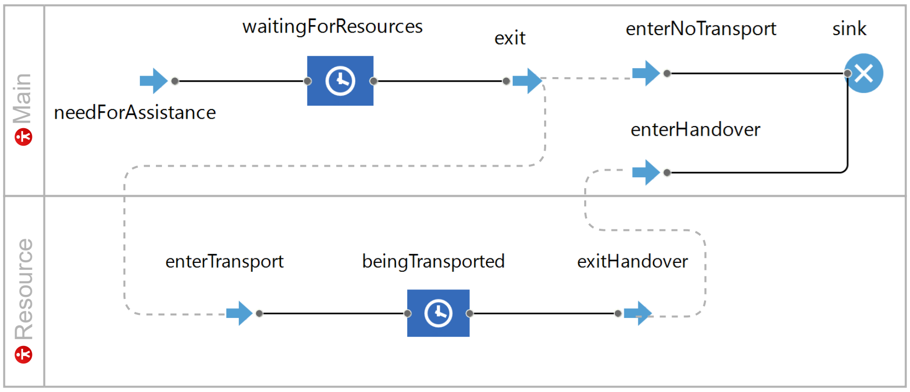


Figure 6.6: Discrete event model of patients in AnyLogic (gray elements serve explanatory purposes and are not part of the AnyLogic model).

Fig. 6.7 depicts the process flow chart for handling requests in a coordination center. Incoming requests start their process in *incomingRequest*. The first step is *processingRequest*, analogous to Fig. 6.4a. As the focus of the implemented model lies on the EMS resources, and no patients without a need for resources are simulated, the process flowchart is simplified. Resources are dispatched in *dispatchingResources*. If all required resources could be dispatched, the process concludes in *end*. If not, the request will flow to *waitingForAvailableResources* until all required resources are dispatched.

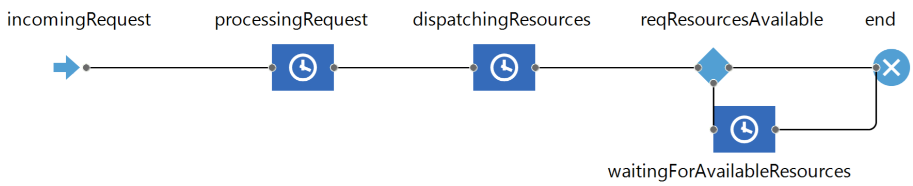


Figure 6.7: Discrete event model of coordination centers in AnyLogic.

6.2.2.3 GIS-environment and simulation dashboard

A GIS environment enables modeling realistic travel distances for EMS resources. It allows routing resources along the street network of the EMS system under study based on data obtained from OpenStreetMap (OpenStreetMap contributors, 2025). Olave-Rojas and

Nickel (2021) introduced the use of machine-learning methods to model driving speeds of EMS resources. We adopt this approach by training a decision tree model based on historical data and using the trained model in the simulation to determine the speed with which a resource travels for a given simulated task. Further information on the speed modeling is provided in Section 6.3.2. Using a GIS environment additionally aids in communicating with practitioners as well as in validating the model. For similar purposes of validation and communication, we integrate a dashboard into the simulation model, which displays performance metrics during simulation runs. For a deeper analysis of simulation results, relevant timestamps and utilization data of resources are collected during the simulation and stored in a database.

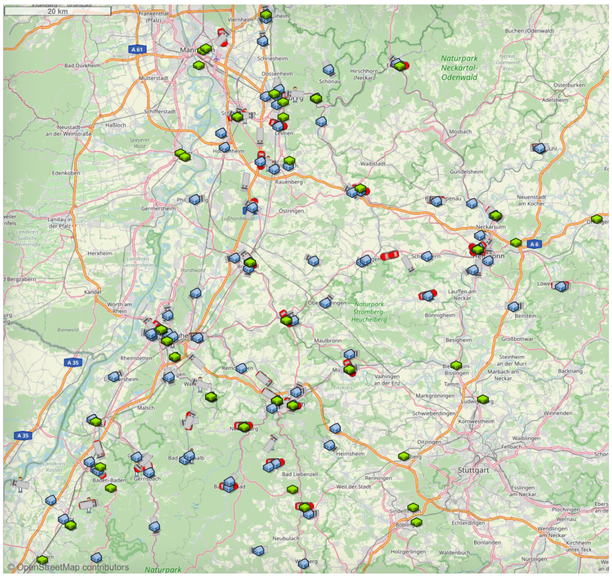


Figure 6.8: GIS-environment of the simulation model visualizing hospitals (green), stations (blue), ambulances (grey), EPV (red), and call location (white).

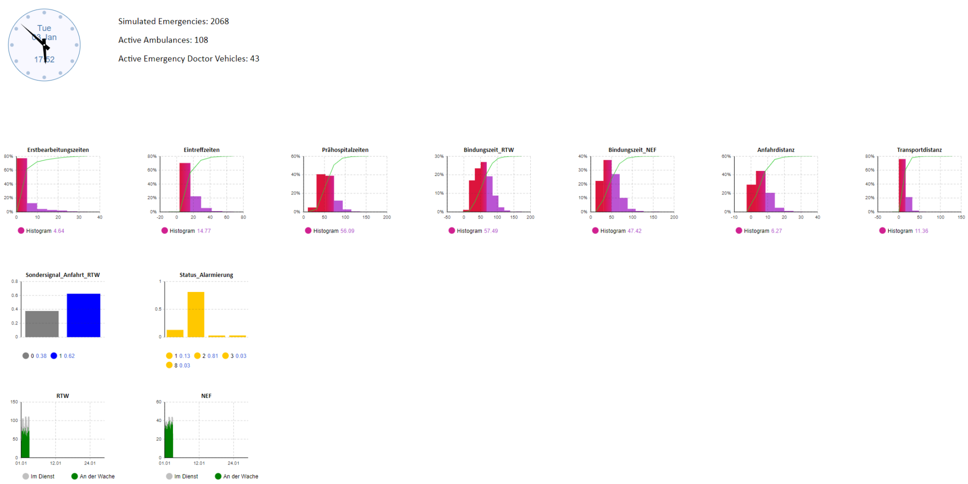


Figure 6.9: Dashboard of the simulation model displaying the distributions of intervals along the rescue chain, call response characteristics, and resource utilizations during runtime.

6.2.2.4 Identifying the closest available resource

The current dispatching strategy in EMS in Baden-Württemberg is to dispatch the closest available suitable resource to an emergency. This requires identifying the closest available resource in the simulation model to model the dispatching decision.

AnyLogic offers a built-in function to identify the closest resource by route distance along the street network. We used this function initially to identify the closest available resource for the dispatching decisions in an initial version of the simulation model. The algorithm implemented in the function is not accessible since AnyLogic is not an open-source software. The function, however, proved to consume more computation time than any other function of the simulation model during performance analysis and hindered the scalability. The assumption is that the route from every available resource to the emergency location is calculated and then used to identify the resource with the shortest route. This requires calculating the shortest path along the street network from every available resource to every call, which is computationally expensive.

We therefore design and implement a custom algorithm for identifying the closest available resource to reduce the computational effort. The algorithm builds on the following core observations:

- If several resources are at the same location, for instance, a station, we do not need to calculate the route for every resource.
- If we can calculate an upper bound for the route distance of the closest available resource, we can reduce the set of potentially closest available ambulances for which we need to calculate the route distances.

Algorithm 1 presents the algorithm that works as follows. We start with the set of available resources R^{disp} out of which one resource can be dispatched. We first reduce the cardinality of this set by selecting only one representative resource for each location at which at least one resource is located. This results in the smaller set $R^{uniqueLoc}$.

Secondly, we calculate an upper bound for the route distance of the closest available resource as follows. We calculate the Euclidean distances d_r^e from each resource r in $R^{uniqueLoc}$ to the emergency location e, and identify a resource with the shortest Euclidean distance. Next, we calculate the route distance $\hat{d}^s$ of this resource to the emergency location. This distance serves as an upper bound for the route distance of the closest available resource. The Euclidean distance will never be longer than the distance along the street network. This allows us to further subset $R^{uniqueLoc}$ to R^{feas} by eliminating all resources with an Euclidean distance above $\hat{d}^s$.

If R^{feas} contains more than one resource, we call the built-in AnyLogic function to identify the closest available resource within R^{feas}. If R^{feas} contains only one resource, we have already identified the closest available resource.

Algorithm 1 Identifying the closest available resource

Let R^{disp} be the set of resources available for dispatch.
Let $L = \emptyset$ be the set to contain the locations of the resources.
for $r \in R^{disp}$ **do**
  Let l_r be the location of resource r
  if $l_r \notin L$ **then**
    Add l_r to L
Let $R^{uniqueLoc} = \emptyset$ be the set to contain one resource per location.
for $l \in L$ **do**
  Let $R^l = \{r \in R^{disp} | l_r = l\}$
  if $|R^l| > 1$ **then**
    Let t_r^l be the time for which r has been at l.
    Add $r \in R^l : r = arg\,max_{r \in R^{disp}}\{t_r^l\}$ to $R^{uniqueLoc}$. Resolve ties randomly.
  else $R^{uniqueLoc} \leftarrow R^{uniqueLoc} \cup R^l$
for $r \in R^{uniqueLoc}$ **do**
  Calculate the Euclidean distance d_r^e to the call location e.
Let $d^{e^\star} = \min\{d_r^e | r \in R^{uniqueLoc}\}$.
Let $\hat{d}^s$ be the distance of the route to the call location on the street network for a random resource in $\{r \in R^{uniqueLoc} | d_r^e = d^{e^\star}\}$.
Let $R^{feas} = \{r \in R^{uniqueLoc} | d_r^e \leq \hat{d}^s\}$.
Identify the resource $r \in R^{feas}$ with the shortest route distance to the call location. Resolve ties randomly.

The presented implementation demonstrates the feasibility of the modeling architecture by implementing the presented state charts and process flow charts, as well as the process control concepts as an executable simulation model. The instantiation of the simulation model for a specific EMS system is performed through the parametrization of the model, which will be discussed in the following section.

6.3 Verification and Validation for the EMS in Baden-Württemberg

Before a simulation model can be used to evaluate changes to a system, it needs to be verified and validated to ensure that the model is a sufficiently accurate representation of the system under study for the intended purposes of the simulation model. The intended purpose of the simulation model in the scope of this thesis is to study the behavior of EMS resources depending on logistic decisions.

Verification is concerned with determining whether the concept model has been correctly implemented, while validation is concerned with whether the implemented simulation model is an accurate representation of the real-world system (Law, 2015, p.246f.). The syntactical verification of the implementation is supported by the debugging functionality of AnyLogic and Java, on which AnyLogic is based. The semantical verification is conducted jointly with the validation as follows.

Sargent (2013) presents an overview of methods for verification and validation, of which the following are applied. *Animation* of the model behavior during runtime as shown in Fig. 6.8 enables the visual validation that resources behave as expected. The *trace* of every emergency rescue is collected, including time stamps, involved resources, and coordinates of emergency locations, and is used to validate the model by analyzing the traces for logical consistency and unexpected behaviors. The integrated dashboard, as shown in Fig. 6.9, represents an instance of *operational graphics* and is used to observe the model behavior during runtime. *Face validation* is performed by presenting the model to EMS experts to check whether the modeling logic is in line with practical experiences. Lastly, we conduct *historical data validation* by comparing the simulation output with historical data. The basic idea of historical data validation is to compare the simulation output to the historical output of the system under study for similar inputs. This necessitates designing validation experiments, parametrizing the simulation model, and analyzing the simulation results, which we describe in the following.

6.3.1 Historical Data Validation: Experimental Setup

The validation experiments are based on the eleven EMS regions depicted in Fig. 6.10, from which historical data is available for this thesis.

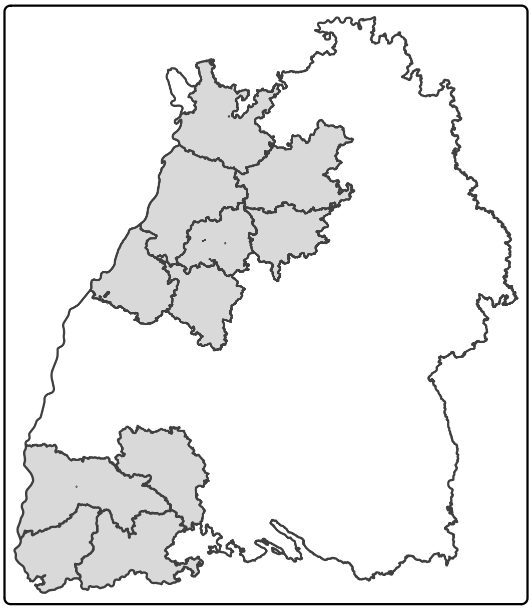


Figure 6.10: Overview of EMS regions used for historical data validation experiments.

The experiments contain all emergencies that occurred in the year 2021 and have historically been attended to by an ambulance and/or EPV stationed in one of these EMS regions. We conduct ten independent simulation runs for each EMS region, resulting in a total of about 4 million simulated complete call responses. A warm-up time of 24 hours is excluded from the analysis based on preliminary experiments.

6.3.2 Historical Data Validation: Simulation Parametrization

The input of the simulation model consists of several parameters. Historical data validation experiments aim to parametrize the simulation model as closely as possible to the historical system for the intended purposes of the simulation model. The necessary parameters of the simulation model can be grouped into:

- System configuration parameters: resource, hospital, and street network characteristics
- Emergency parameters: Dates, locations, and categories of emergencies
- Process parameters: Durations of process steps and control of process variants

The setting of these parameters is described in the following. The parametrization has largely been adopted from the master's thesis of Watzinger (2020), which was based on the simulation model developed in Olave-Rojas and Nickel (2021).

6.3.2.1 System Configuration Parameters

The system configuration parameters contain information on the *hospitals*. The set of hospitals in Baden-Württemberg is denoted as H in the following. Hospital characteristics entail the location in the form of coordinates and the facilities of hospitals for each EMS region in the set $\mathcal{R}$ of simulated EMS regions. Hospitals relevant for an EMS region $r \in \mathcal{R}$ are not necessarily located geographically within r. To identify the set $H_r^\star$ of relevant hospitals for r, we firstly calculate the proportion p_r^h of transports conducted by resources stationed in r to each hospital $h \in H$. The hospitals are added to $H_r^\star$ in descending order of p_r^h until $\sum_{h \in H_r^\star} p_r^h \geq 0.95$.

Facilities are considered to respect treatment requirements for tracer diagnoses. Tracer diagnoses in Baden-Württemberg contain the emergencies sudden cardiac arrest, ST elevation myocardial infarction, severely injured/polytrauma, severe traumatic brain injury, and sepsis. Hospitals need to offer specific facilities to be able to treat patients with a tracer diagnosis, which is considered in the hospital selection logic in the model.

The parameters of *EMS resources* contain the location of the resource stations as well as the shift schedule for each day of the week. We consider all ambulances and EPV stationed in a simulated EMS region.

Lastly, the street network needs to be defined. We use OpenStreetMap data to define the street network in the simulation model.

6.3.2.2 Emergency Parameters

The emergency parameters control when and where an emergency is generated. Each emergency additionally contains the information whether it is one of the tracer diagnoses and if it belongs to a certain call category.

One approach to controlling the call generation would be to model the spatial and temporal distribution of emergency occurrences based on the historical data. This would facilitate studying the behavior of an EMS system under different emergency occurrence scenarios. Developing such a model, however, is not trivial, and the intended purpose of the simulation model does not lie in investigating the system boundaries. We therefore use the historic timestamps of call arrivals at the coordination center, as well as the historic emergency locations and information on tracer diagnoses and categories.

6.3.2.3 Process Parameters

The process parameters determine the frequency of process variants and the duration of process steps. Process variants include decisions such as whether an EPV is dispatched, whether an ambulance uses sirens, or whether patient transport is required. We model all process variant decisions as binary and determine them through probabilities. For most process steps, we directly specify duration parameters as distribution functions. We draw one sample from these distributions for each simulated call response. For travel times, we set an average travel speed along the street network. We determine the speed for each travel using a decision tree as a regression model. Table 6.1 lists the process parameters and their conditioning variables, followed by an explanation of how they are determined.

Table 6.1: Simulation process parameters and their conditioning variables.

Parameter	Type	Conditioning variables
initial EPV dispatch	probability	category
ambulance sirens	probability	initial EPV dispatch, category
dispatch time	distribution	sirens, category
ambulance chute time	distribution	status, station type, sirens, time of day
EPV chute time	distribution	status, station type, time of day
ambulance speed	speed	status, category, sirens, distance, EMS region, hour, day of the week
EPV speed	speed	status, distance, EMS region, hour, part of the week
requested EPV dispatch	probability	category
ambulance on-scene time	distribution	category
EPV on-scene time	distribution	dispatch type (initial or requested)
ambulance transport	probability	category
EPV transport	probability	category
ambulance sirens during transport	probability	EPV transport involvement, category
transport speed	speed	category, sirens, distance, hour, part of the week, EMS region
ambulance handover time	distribution	EPV involvement
EPV handover time	distribution	-
ambulance return speed	speed	distance, hour, part of the week, EMS region
EPV return speed	speed	distance, hour, part of the week, EMS region

We determine the values of the parameters based on historical data. The conditioning variables result from the iterative verification and validation process. The probabilities equal the historical proportions of process variants. The distributions equal the empirical distribution functions of historical process times. As an example, Fig. 6.11 depicts the empirical distribution functions for the chute times of EPV when being dispatched from their station. The conditioning variables are the type of station and the time of day. The stations were clustered into slow, medium, and fast stations using k-means clustering based on their historical average chute times. This reflects the influence of different station layouts on the chute time. Additionally, the hours from 1 to 6 AM are classified as night, and the remaining hours of the day as day. During night hours, the EPV crew might rest, which can increase the chute times.

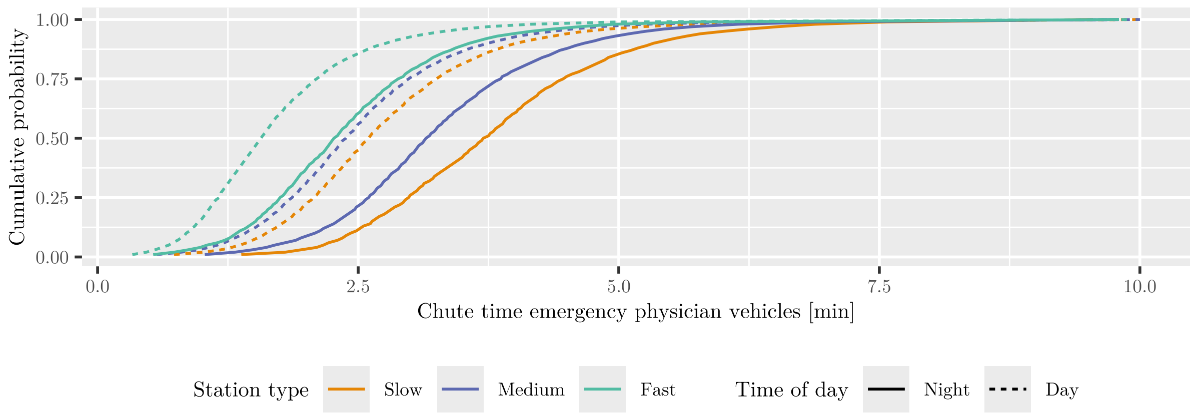


Figure 6.11: Empirical distributions of EPV chute times.

For the speed modeling, we follow the approach introduced by Olave-Rojas and Nickel (2021). The authors trained a regression model to calculate a table of travel speeds for a set of discrete feature values. We expand on this approach by training regression models based on the historic travel time data and then using the trained models to predict travel speeds during simulation runtime. To facilitate the implementation in AnyLogic, the Weka Java API was used (Frank et al., 2016), which limited the available model types for regression. A preliminary comparison of a support vector machine, a multilayer perceptron, and a decision tree led to selecting the decision tree as a regression model, specifically the REPTree class implemented in Weka. Feature engineering resulted in training three separate decision trees, one for the ambulance travels to the scene, one for EPV travels to the scene, and one for transports from the scene to a hospital. Transports are conducted either jointly by an ambulance and EPV or by a single ambulance. All models consider the features of the EMS region, the call category, the travel distance, and the hour of the day. The ambulance travel model additionally considers whether the ambulance is alarmed at or outside its station, whether sirens are used, and the day of the week. The EPV travel model additionally considers whether the EPV is alarmed at or outside its station, and the part of the week (weekday or weekend). The transport model additionally considers whether the ambulance uses sirens, and the part of the week (weekday or weekend). The historic data does not contain the timestamps of arrivals back at the station after serving a call, preventing the use of historic data to model the speeds when driving back to a station from the scene of the emergency or from a hospital. We use the transport speed model based on discussions with practitioners and validation results.
The root mean squared error is used as a performance metric. The decision trees are evaluated using 5-fold cross-validation, following Olave-Rojas and Nickel (2021).

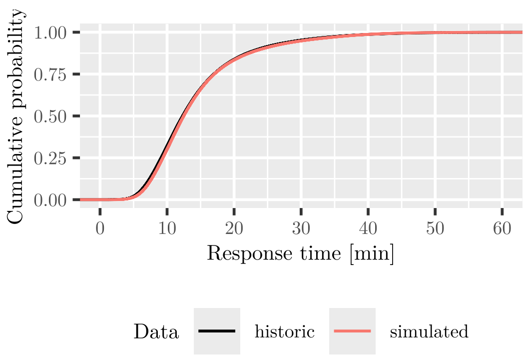
As a benchmark, we compare the evaluation results to predicting travel speeds with the historical mean in Table 6.2, showing that the decision tree improves on the prediction quality.

Table 6.2: Root mean squared errors of speed modeling.

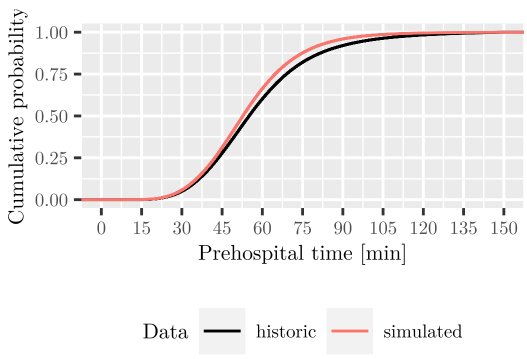
Data	Decision Tree	Mean Prediction
Travel ambulance	12.05	21.11
Travel EPV	13.96	22.17
Transport	12.21	19.36

6.3.3 Historical Data Validation: Results

The two most important time intervals from a patient's perspective are the response time and the prehospital time. Fig. 6.12 compares the empirical distribution functions of the simulated and historical response and prehospital times. Response times below 60 minutes and prehospital times below 150 minutes are shown for ease of readability. 0.03% of simulated response times, and 0.04% of simulated prehospital times exceed the threshold.



(a) Response times.



(b) Prehospital times.

Figure 6.12: Comparison of response and prehospital times in historical and simulated data.

Fig. 6.12 shows a close resemblance between historic and simulated process times. The simulated prehospital times are slightly shorter than the historic ones. The deviation stems from shorter transport distances in the simulation, which are a result of the decision to which hospital a patient is transported. In an initial version of the simulation, each patient was transported to the closest hospital, considering the necessary hospital facilities in case of a tracer diagnosis. This resulted in too short transport durations in the simulation. Therefore, the historical decisions on whether transport is required and to which hospital are followed in the simulation. The closest feasible hospital is selected if the historic destination decision is missing in the data. This reduced, but did not fully close, the gap between simulated and historic transport distances.

Similar to the displayed response and prehospital times, the empirical distributions of the remaining process steps, as well as the proportions of the process variants, were compared during validation, and no major deviations were found in the final version of the model.

6.4 Simulating EMS: Discussion

This chapter aimed to develop a simulation model to predict how logistic decisions impact the quality of EMS systems. To this end, we proposed a concept model designed to model both emergency rescue and patient transport operations with an arbitrary number of resources involved in an operation. Furthermore, we implemented and validated the concept model as an executable simulation model in AnyLogic. The implementation and validation are tailored to emergency rescue operations in Baden-Württemberg, Germany, and conducted using real-world data.

Contributions
Existing simulation models are tailored to specific EMS systems, and the lack of transferable simulation models has been identified in existing literature reviews. While our implemented model and the validation are tailored to emergency rescue operations in Baden-Württemberg, Germany, due to the available data and the scope of this thesis, the proposed concept model is designed to be adaptable to different EMS system designs.

Practical Implications
The developed simulation model can enable decision makers to evaluate logistic decisions before they are implemented. Additionally, simulation can support in understanding system boundaries, for instance, in terms of call volumes, and the impact of external changes, for instance, in the hospital landscape. The explicit modeling of the patient pathway in the simulation model allows capturing timestamps along the rescue chain. This enables assessing the simulated EMS system based on different quality indicators. Assessing the effects of changes ahead of their implementation and understanding the impacts of changes in external factors is especially valuable in the context of EMS where logistic decisions can impact the health and well-being of patients.

Limitations
The following limiting factors need to be taken into account when interpreting the results. While we argued why the proposed concept model should be transferable to different EMS systems and modeling scopes, we implemented and validated the simulation model only for emergency rescue operations in Baden-Württemberg, Germany. The claims that the concept model is indeed capable of modeling both emergency rescue and patient transport operations and arbitrary numbers of involved resources could therefore not be investigated through computational experiments. The results of the validation experiments

can therefore not be generalized beyond the studied scope. Limited data regarding hospital facilities and treatment requirements for the majority of emergencies hindered modeling the hospital choice. Initial experiments demonstrated that the hospital choice is not only based on the distance to the emergency location. While following the historic hospital choices in the simulation enabled validating the model, it hinders applying the simulation model to study, for instance, the impacts of changes in the hospital landscape.

Methodological Reflections

We made several methodological choices that influence the results. The validation experiments use the historic call dates and locations. Alternatively, calls could be generated randomly based on a spatio-temporal distribution. This would have necessitated modeling the spatio-temporal distribution of calls, but would facilitate experimenting with, for instance, different call volumes.

The conditioning variables for the process parameters were identified based on an analysis of historical data and discussions with EMS experts. We did, however, not systematically investigate how considering each conditioning variable impacts the results of the validation experiments. The close resemblance of simulated and historical process times indicates that we did not overlook conditioning variables with a major effect on the results, but it might have been interesting to investigate which conditioning variables are necessary to validate the model.

Outlook

The presented simulation model presents various avenues for further research. Modeling the travel speeds using regression models offers potential for further development. A broader comparison of regression models independently of the implementation in Java could lead to identifying further suitable model classes for predicting travel speeds. Another area for improvement is the feature engineering. It could be investigated how additional features, such as the street types along the route, impact the prediction quality. Hyperparameter tuning could additionally improve the model performance.

The simulation model has been validated for ground-based emergency rescue operations in the EMS in Baden-Württemberg. The model could be expanded by additionally modeling EMS helicopters. Likewise, patient transport operations could be included to investigate, for instance, sharing resources between emergency rescues and patient transportation.

Applying the concept models to EMS systems outside Baden-Württemberg could shed light on whether the concept model is sufficiently generic to be able to model different EMS systems.

The simulation model has been validated based on historical data. Investigating how scenarios for the future development of the spatio-temporal distribution of calls can be generated would provide a valuable input to the simulation model when using the simulation model in a planning process for EMS in practice.

6.5 Simulating EMS: Conclusion

This chapter addressed the question of how to predict the impact of logistic decisions on the quality of EMS. We identify simulation as a suitable and commonly employed method for predicting how EMS systems operate, and how logistic decisions impact operations. Hybrid simulation combining agent-based and discrete-event modeling paradigms appears especially suitable to model EMS operations. Existing simulation models of EMS systems often suffer from a lack of transferability as they tend to be designed for a specific EMS system. We propose a generic and adaptable concept model to address this limitation. Implementing the concept model in AnyLogic and validating it for emergency rescue operations using real-world data results in a tool for analyzing the impact of logistic decisions on the EMS in Baden-Württemberg, and we use the simulation model in the computational experiments of the subsequent chapters.

Part IV

Dispatching Ambulances for Heterogeneous Patients

Summary of Part IV

Research Question and Objectives

Research Question: How can the ambulance dispatching decision consider heterogeneous patients?
Research Objectives:

- Review and classify ambulance dispatching strategies
- Identify suitable dispatching strategies for the EMS in Baden-Württemberg and the patient categories developed in Part II
- Evaluate and compare suitable dispatching strategies

Scope and Methodology

We focus on how we can initially dispatch ambulances to non-prioritized calls. The dispatching decision is an operational decision in EMS, as depicted in Fig. 6.13

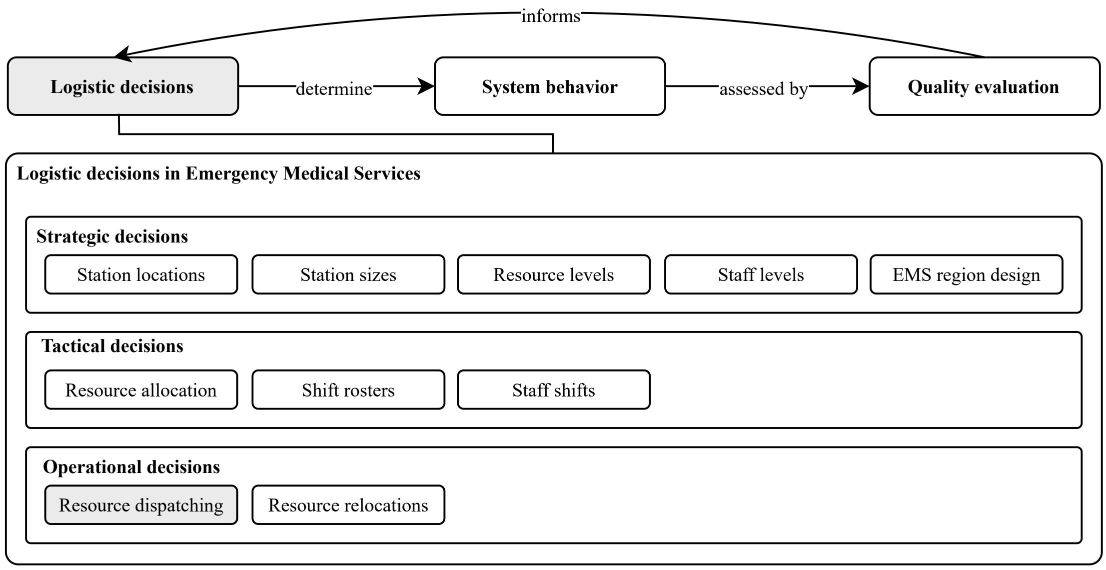


Figure 6.13: Part IV investigates the dispatching decision for heterogeneous patients.

We conduct a *scoping review* to identify ambulance dispatching strategies in the literature, and to derive relevant attributes for selecting suitable dispatching strategies for heterogeneous patients in the EMS in Baden-Württemberg. We select eight dispatching strategies, partly adapt them, and compare them using the *simulation* model developed in Part III.

Main Contributions

- Scoping review of ambulance dispatching strategies
- Selection of strategies suitable for heterogeneous patients in the EMS in Baden-Württemberg
- Comparison of suitable strategies in simulation experiments with real-world data

Key Findings

- Driving times to prioritized calls can be reduced significantly by not always dispatching the closest available ambulance to non-prioritized calls.
- The impact of not always dispatching the closest available ambulance depends on the region under study.
- Not always dispatching the closest-available ambulance significantly increases driving times to non-prioritized patients in a larger magnitude than the decrease in driving times to prioritized patients on our experiments.
- If actually prioritized calls are classified as non-prioritized in the dispatching decision too often, driving times to actually prioritized calls can increase beyond always dispatching the closest available ambulance.

Chapter 7

Dispatching Strategies for EMS Resources: A Scoping Review

We developed a novel approach to evaluating the quality of EMS in Part II, and a tool to evaluate the effects of logistic decisions on the quality of EMS in Part III. This enables us to investigate how logistic decisions can impact the quality of EMS systems.

There are several logistic decisions in EMS systems on the operational, tactical, and strategic level, and we focus on the operational dispatching decision in this part of the thesis. The dispatching decision refers to the decision of how EMS resources are dispatched to calls. A common dispatching strategy in practice is the closest-available dispatching strategy according to the reviews of Aringhieri et al. (2017) and Bélanger et al. (2019). The closest available dispatching strategy always dispatches the suitable resource that can reach the call location in the shortest amount of time.

Carter et al. (1972) demonstrated in an early study that this myopic strategy is not always optimal, even without differentiating between call types. In a more recent study, Bandara et al. (2014) differentiate between Priority 1 and Priority 2 calls, and show that the performance for Priority 1 calls can be improved when resources are not dispatched by the closest-available strategy for Priority 2 calls. This indicates that the knowledge about patient categories and their utilities from Part II can be used in the dispatching decision to improve utilities, especially for critical patients.

To investigate how the patient categories and utilities can be considered in the dispatching decisions, we should first examine which existing dispatching strategies can be adapted to patient categories and utility functions. Secondly, to investigate dispatching strategies using our simulation model of the EMS in Baden-Württemberg, we need to identify which dispatching strategies are suitable for the design of the EMS in Baden-Württemberg, Germany. Several reviews of logistic decisions in EMS include dispatching strategies. Reuter-Oppermann et al. (2017) present a review of logistic problems in EMS, including a section on dispatching strategies. Aringhieri et al. (2017) and Bélanger et al. (2019) both conducted reviews of logistic problems and modeling approaches in EMS, including the dispatching decision. To the best of our knowledge, Neira-Rodado et al. (2022) conducted the only systematic literature review of logistic decisions in EMS, focusing on dynamically locating ambulances and following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines by Moher et al. (2009).

None of these reviews, however, focus explicitly on dispatching strategies, and it is unclear how comprehensive the included dispatching strategies are. Additionally, no classification of dispatching strategies is proposed, based on which a selection of suitable strategies could be conducted. To the best of our knowledge, there exists no dedicated review of dispatching strategies for EMS resources that would allow us to systematically identify suitable dispatching strategies.

To address this gap, we conduct a scoping review of dispatching strategies for EMS resources to answer the following questions:

- Which dispatching strategies for routine EMS operations exist?
- Through which attributes can the suitability of dispatching strategies for a specific EMS system be assessed?

The remainder of this chapter is structured as follows: We describe the review methodology in Section 7.1. Section 7.2 presents the identified dispatching strategies as well as their attributes. We discuss our approach and findings in Section 7.3 and conclude the chapter in Section 7.4.

7.1 Review Methodology

We conduct the scoping review following the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR)-guidelines, as described by Tricco et al. (2018). The PRISMA-ScR-statement provides guidelines for scoping reviews and is an extension of the PRISMA guidelines followed by Neira-Rodado et al. (2022). In the following, we present the information sources, the search strategy, and the selection process of the review. We use the term record to refer to database entries and the term report to refer to full-texts in the following.

The information sources are the bibliographic databases Web of Science, Scopus, and PubMed. Web of Science and Scopus are both large multidisciplinary bibliographic databases, while PubMed focuses on biomedical literature.

We construct our search string based on an exploratory search phase for suitable reports. The search string consists of a combination of a domain-specific and a method/decision-specific part. The search string is presented in Table 7.1. Web of Science and Scopus were last queried on 21.08.2023, and PubMed on 20.10.2023.

Part	Search string
Domain	ambulance* OR ems OR “emergency medical service*” OR “emergency resource*” OR “emergency vehicle*”
Method/Decision	dispatch* OR allocation OR operation* OR management OR “response strateg*” OR mdp OR heuristic* OR “online optimization” OR “approximate dynamic programming” OR “hypercube queuing” OR “markov decision process”

Table 7.1: Search strings for the scoping review of dispatching strategies

The eligibility criteria for the inclusion of reports are as follows:

- written in the English or German language
- published in a peer-reviewed journal or peer-reviewed conference proceedings
- presentation of an original dispatching strategy
- applicable to daily, routine EMS operations, and not only for disaster situations

The first criterion ensures that the reviewer can analyze the original publication. The second criterion aims to include only high-quality reports in the review. The third criterion refers to whether a novel, original dispatching strategy is proposed in a report or whether existing strategies are analyzed. The last criterion ensures that the dispatching strategy applies to daily EMS operations. Disaster situations in this context describe situations in which resources need to be dispatched once to a set of multiple known patients, while resources are dispatched repeatedly over time in routine EMS operations.
To identify suitable reports, we first query the databases with the search string and extract all identified records. We then screen the records based on their titles. We seek to retrieve the full reports for the eligible records and — if retrievable — assess them further by reading the abstracts and — in case of abstract ambiguity — full texts. Lastly, we conduct a backward and forward search based on the remaining reports.
The review aims to gain an overview of dispatching strategies and to identify attributes of dispatching strategies that make a dispatching strategy more or less suitable for a specific EMS system. We therefore analyze all included reports in full, identify attributes that differentiate the dispatching strategies, and extract the attribute values for each dispatching strategy.

7.2 Dispatching Strategies and Attributes

The following sections present the results of the selection process and the extraction of attributes of dispatching strategies. Section 7.2.1 provides the evolution of reports and records throughout the selection process as well as an overview of the included reports and their attributes. Section 7.2.2 analyzes the distributions of attributes and discusses their relevance for identifying dispatching strategies for a specific EMS system.

7.2.1 Overview of Dispatching Strategies

Fig. 7.1 details the evolution of the number of records and reports throughout the selection process. We exclude the majority of records based on screening the titles. Assessing the reports for eligibility based on screening abstracts and full texts narrows the selection down further. The resulting reports are augmented by six further reports identified through backward and forward search, resulting in a total of 43 reports included in the review.

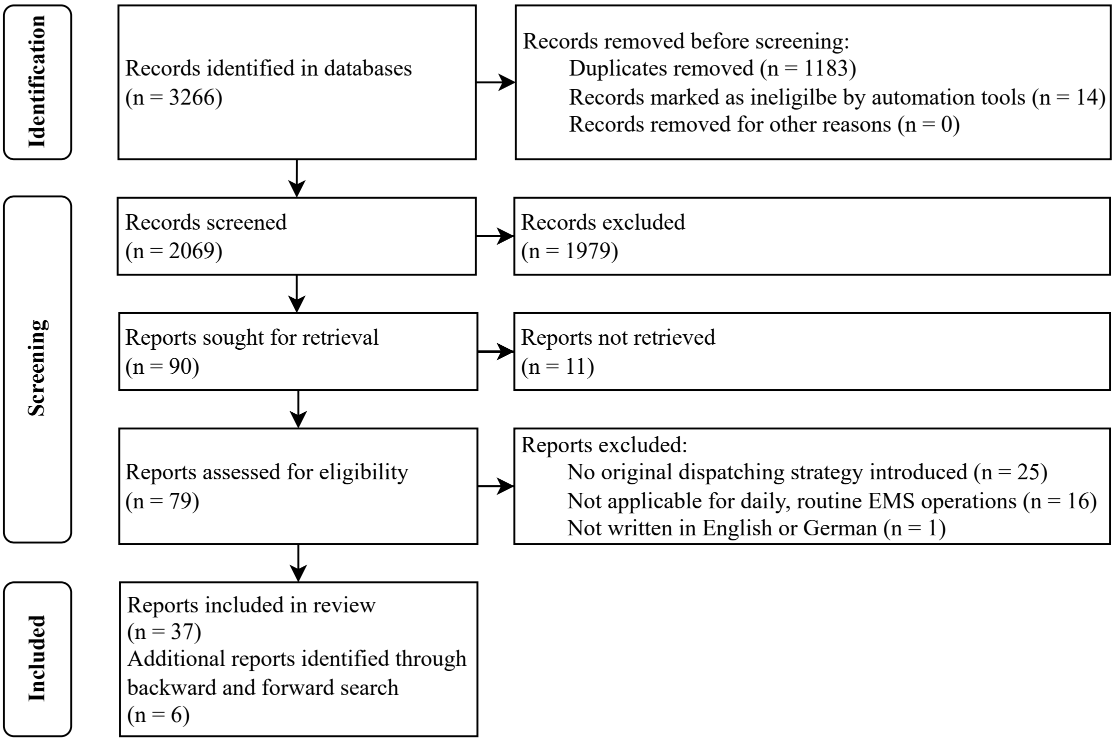


Figure 7.1: Selection process.

Table 7.2 lists the included reports with their attribute values. We briefly introduce the attributes and will explain them and their values in more detail in Section 7.2.2. The attribute *optimization approach* refers to whether the report employs an offline or online approach. Offline approaches execute algorithms once and result in a static output, which is then used during EMS operations for dispatching. Online approaches execute algorithms during EMS operations at the time of dispatch and result in a suggestion for which resource to dispatch. The attribute *resource types* refers to the number of considered resource types. Systems with multiple resource types are classified with the attribute *resource system characteristics* since multiple resource types are considered in different ways. The reports further vary in the number of *call types*, as well as how call types are differentiated, which we capture with the attribute *call type differentiator*. Some reports consider the dispatching decision integrated with either an operational relocation decision or the strategic location decision, which we capture with the attribute *decision integration*. Lastly, the reports differ in the *decision scope*, which refers to the set of actions considered as part of the dispatching decision.

Table 7.2: Overview of reports and their attributes included in the review of dispatching strategies.

Report	**Optimization approach**	**Resource types**	**Resource system characteristics**	**Call types**	**Call type differentiators**	**Decision integration**	**Decision scope**
Carter et al. (1972)	Offline	1	-	1	-	Not integrated	Ambulance choice
Jarvis (1981)	Offline	1	-	flexible	Performance metrics	Not integrated	Ambulance choice
Ahituv and Berman (1987)	Offline	1	-	1	-	Not integrated	Ambulance choice
Haghani et al. (2004)	Online	1	-	5	Performance metrics	Not integrated	Ambulance choice
Yang et al. (2005)	Online	3	Multiple response; strict compatibility	5	Vehicle requirements; performance metrics	With relocation	Ambulance choice
T. Andersson and Värbrand (2007)	Online	1	-	3	Weight of performance metrics	Not integrated	Ambulance choice
Lopez et al. (2008)	Online	2	Single response; preference	1	-	Not integrated	Ambulance choice
Iannoni et al. (2009)	Offline	1	-	2	Vehicle requirements	Not integrated	Ambulance choice
S. Lee (2011)	Online	1	-	1	-	Not integrated	Ambulance choice
Chanta et al. (2012)	Offline	1	-	1	-	With location	Ambulance choice
Ibri et al. (2012)	Online	1	-	flexible	Weight of performance metrics	With relocation	Ambulance choice
Bandara et al. (2012)	Offline	1	-	2	Weight of performance metrics	Not integrated	Ambulance choice
Schmid (2012)	Offline	1	-	1	-	With relocation	Ambulance choice
Ibri et al. (2013)	Online	1	-	2	Weight of performance metrics	With relocation	Ambulance choice; dispatch or wait
Mclay and Mayorga (2013a)	Offline	1	-	2	Weight of performance metrics	Not integrated	Ambulance choice
Mclay and Mayorga (2013b)	Offline	1	-	2	Weight of performance metrics	Not integrated	Ambulance choice

Continued on next page

Continued from previous page

Report	**Optimization approach**	**Resource types**	**Resource system characteristics**	**Call types**	**Call type differentiators**	**Decision integration**	**Decision scope**
Mayorga et al. (2013)	Offline	1	-	2	Weight of performance metrics	Not integrated	Ambulance choice
Meinzer and Storandt (2014)	Online	1	-	1	-	Not integrated	Ambulance choice
Billhardt et al. (2014)	Online	1	-	1	-	Not integrated	Ambulance choice
Bandara et al. (2014)	Offline	1	-	2	Weight of performance metrics	Not integrated	Ambulance choice
Sudtachat et al. (2014)	Offline	2	Multiple response; hierarchical compatibility	3	Vehicle requirements; weight of performance metrics	Not integrated	Ambulance choice
Toro-Díaz et al. (2015)	Offline	1	-	1	-	With location	Ambulance choice
H. Li and Chai (2015)	Online	1	-	3	Performance metrics; weight of performance metrics	Not integrated	Ambulance choice
Ansari et al. (2017)	Offline	1	-	2	Weight of performance metrics	With location	Ambulance choice
S. Lee (2017)	Online	1	-	1	-	Not integrated	Ambulance choice
Jagtenberg, Bhulai, and van der Mei (2017)	Online	1	-	1	-	Not integrated	Ambulance choice
Jagtenberg, Bhulai, and van der Mei (2017)	Offline	1	-	1	-	Not integrated	Ambulance choice
Yoon and Albert (2018)	Offline	1	-	2	Weight of performance metrics	With location	Ambulance choice; dispatch or wait
Amorim et al. (2018)	Online	1	-	2	Performance metrics	Not integrated	Ambulance choice
Nasrollahzadeh et al. (2018)	Online	1	-	2	Weight of performance metrics	Not integrated	Ambulance choice; dispatch or wait

Continued on next page

Continued from previous page

Report	**Optimization approach**	**Resource types**	**Resource system characteristics**	**Call types**	**Call type differentiators**	**Decision integration**	**Decision scope**
Aringhieri et al. (2018)	Online	1	-	3	Performance metrics	Not integrated	Ambulance choice; dispatch or wait
Mukhopadhyay et al. (2019)	Online	1	-	1	-	Not integrated	Ambulance choice
Enayati et al. (2019)	Offline	1	-	2	Weight of performance metrics	With location	Ambulance choice
Park and Lee (2019)	Offline	2	Single response; preference	2	Performance metrics	With relocation	Ambulance choice
Yoon and Albert (2020)	Offline	2	Multiple response; hierarchical compatibility	3	Vehicle requirements; performance metrics	Not integrated	Ambulance type(s)
Bélanger et al. (2020)	Offline	1	-	1	-	With location	Ambulance choice
K. Liu et al. (2020)	Offline	1	-	1	-	Not integrated	Ambulance choice
Roa et al. (2020)	Online	2	Single response; preference	2	Performance metrics	Not integrated	Ambulance choice
Yoon and Albert (2021)	Offline	2	Single response; preference	3	Vehicle requirements	Not integrated	Ambulance type(s)
Yoon and Albert (2021)	Online	2	Single response; preference	3	Vehicle requirements	Not integrated	Ambulance type(s)
Theeuwes et al. (2021)	Online	1	-	2	Performance metrics	Not integrated	Ambulance choice
Albert (2023)	Offline	1	-	2	Weight of performance metrics	Not integrated	Ambulance choice; dispatch or wait
MacLachlan et al. (2022)	Offline	3	Multiple response; strict compatibility	flexible	Vehicle requirements	Not integrated	Ambulance choice
MacLachlan et al. (2023)	Offline	flexible	Multiple response; strict compatibility	flexible	Vehicle requirements; weight of performance metrics	Not integrated	Ambulance choice; dispatch or wait; call choice

7.2.2 Attributes of Dispatching Strategies

In the following, we describe the examined attributes of dispatching strategies, explain the attribute values, and reason how these attributes influence the suitability of dispatching strategies for a specific EMS system.

Optimization Approach

The *optimization approach* refers to the fundamental modeling approach of the dispatching decision. Offline approaches solve a model once and result in either a priority list or a policy table. Priority lists are based on aggregated call locations, such as zip codes or a grid structure. A priority list contains an ordered list of resource stations for each aggregated call location. Resources are then dispatched from the most preferred station at which a resource is available at the time of dispatch. Policy tables define one preferred station for each call location for each possible system state. The system state describes at which stations resources are available at the time of dispatch.

Online approaches, on the other hand, solve a model or execute an algorithm at the time of dispatch based on the current system state. This allows, for instance, considering the exact location of each resource, even if it is not located at its station.

Which approach is suitable for an EMS system depends on the technical capabilities. If the exact locations of resources are accessible in the coordination center and the necessary computing resources are available to run online approaches, they can be employed. Otherwise, offline approaches appear more suitable. Fig. 7.2 depicts the proportions of approaches in the considered reports. The majority of reports present offline approaches, but there is still a considerable number of online approaches available.

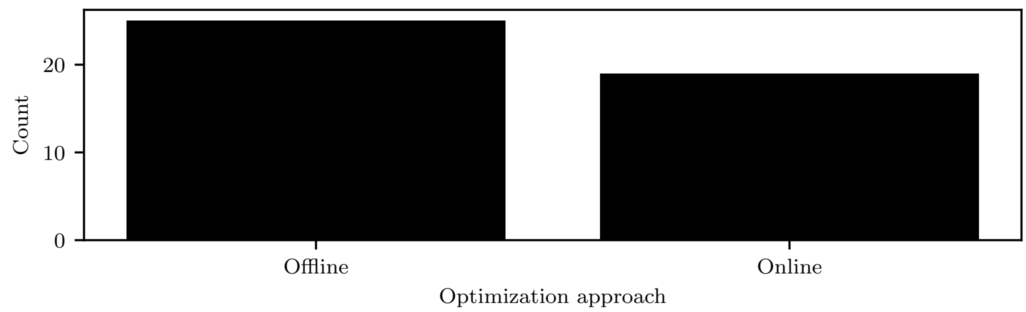


Figure 7.2: Distribution of optimization approaches

Resource Types

The design of EMS systems differs significantly, and one aspect of the design of EMS systems is the *resource types*. Consequently, the dispatching strategies presented in the selected reports differ in the number of considered resource types, as shown in Fig. 7.3. Most approaches assume a single resource type. There are, however, modeling approaches that consider more than one resource type.

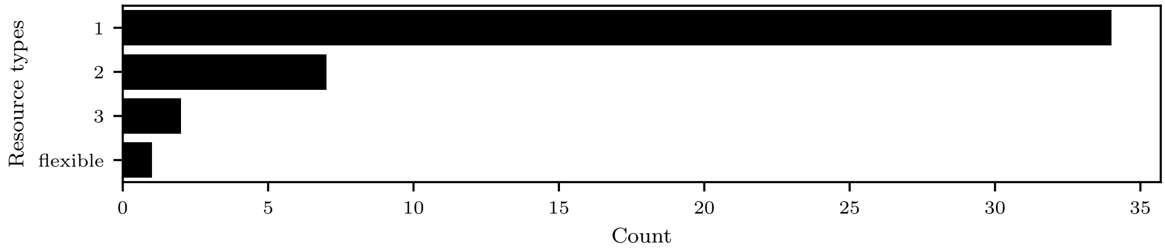


Figure 7.3: Distribution of the considered number of resource types

Resource System Characteristics

When more than one resource type is considered, we further classify the studied EMS systems as depicted in Fig. 7.4. We differentiate whether a *single resource responds* to a call or whether *multiple resources respond*, usually of different resource types. If a single resource is dispatched to a call, preference information for the resource types, sometimes in combination with the call type, is known and considered in the dispatching decision. Systems where multiple resources of different types can be dispatched can be further divided into systems with *hierarchical compatibility* or *strict compatibility*. In systems with *hierarchical compatibility*, resources with tiered qualifications and equipment exist. Both identified reports that consider hierarchical compatibilities also exhibit multiple call types. Higher-tiered resources must respond to critical calls, while lower-tiered resources or higher-tiered resources can serve less critical calls. *Strict compatibility* refers to systems where each call or call type has strict requirements for the number and types of resources necessary to serve the call.

The number of considered resources and the resource system characteristics should match the design of the EMS system under study. Otherwise, it needs to be investigated whether the dispatching strategy can be adapted to the EMS system characteristics.

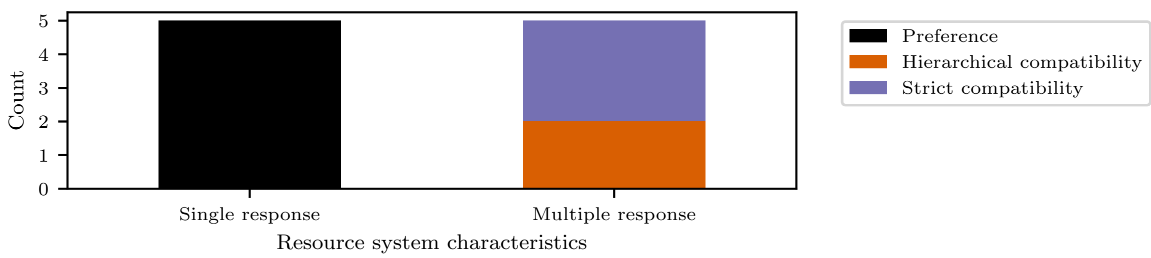


Figure 7.4: Distribution of characteristics of systems with multiple resources

Call Types

Next to the resource types, EMS systems additionally differ in whether and how they differentiate between *call types*. Fig. 7.5 displays the distribution of the number of different call types in the reports. Some approaches consider only a single call type, but most dispatching strategies consider multiple call types.

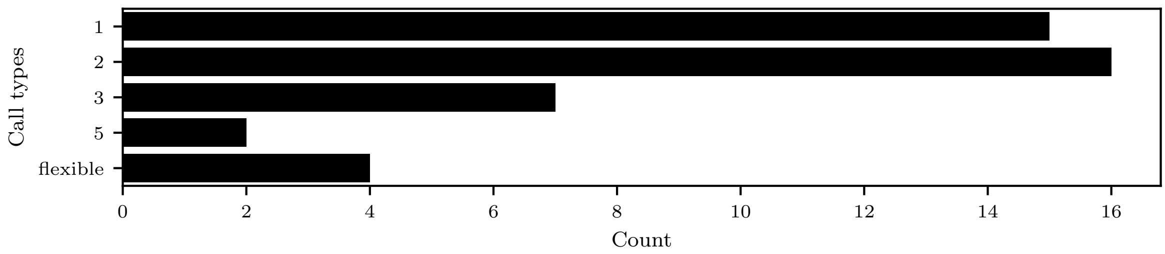


Figure 7.5: Distribution of the considered number of call types

Call Type Differentiators

Even if reports consider the same number of call types, they differ in how call types are differentiated. We present the distribution of *call type differentiators* in Fig. 7.6. For this review, the differentiator does not refer to medical aspects but to how call types are considered in the dispatching strategy. The most common differentiator is the *weight of performance metrics*. In this case, each call type is assigned a different weight that is used in calculating an aggregate performance metric on which the dispatching decision is based. The purpose of the weights is to prioritize more critical call types. Next to the weight of performance metrics, call types are also differentiated by the *performance metrics* themselves. For instance, a survival probability is considered for life-threatening calls, and time-based metrics for less urgent calls. If multiple resource types are present in an EMS system, a call type can also have specific *vehicle requirements*. Vehicle requirements can take the form of strict requirements for a certain resource type or preferences that impact performance metrics.

Whether a dispatching strategy differentiates between call types should match whether the EMS system under study differentiates between call types.

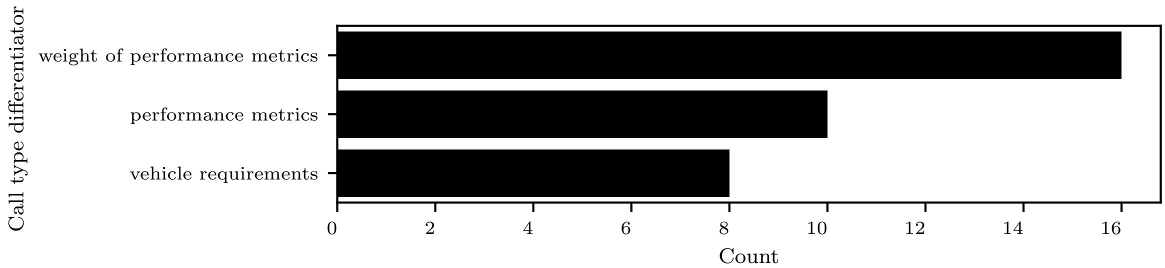


Figure 7.6: Distribution of call type differentiators

Decision Integration

The attribute of the *decision integration* refers to whether dispatching is modeled in isolation or integrated with further decisions. Most approaches model the dispatching decision *not integrated*, as shown in Fig. 7.7. Integrated approaches consider either the strategic station location or the operational relocation decision.

All dispatching strategies modeled integrated *with the location* decision are offline approaches resulting in priority lists. The motivation for integration stems from the influence of the dispatching decision on the proportions of demand serviced by each station. These proportions subsequently influence how many resources are required at a station.

If dispatching is considered integrated *with the relocation* decision, the dispatching strategies are either online approaches or offline approaches resulting in policy tables. Relocating ambulances supports keeping EMS systems prepared for future incoming emergencies.

Whether to consider integrating the dispatching decision depends on the setting of the study and the EMS system characteristics. The location decision has much larger financial implications compared to the dispatching decision. Decision makers might therefore be interested in investigating whether EMS quality can be improved by changing the dispatching strategy in isolation. Considering the relocation decision is only sensible in systems in which resource relocations are conducted. While relocating resources might improve EMS quality, it also has practical implications, for instance, regarding labor laws or organizational challenges, if different resource stations are operated by different providers.

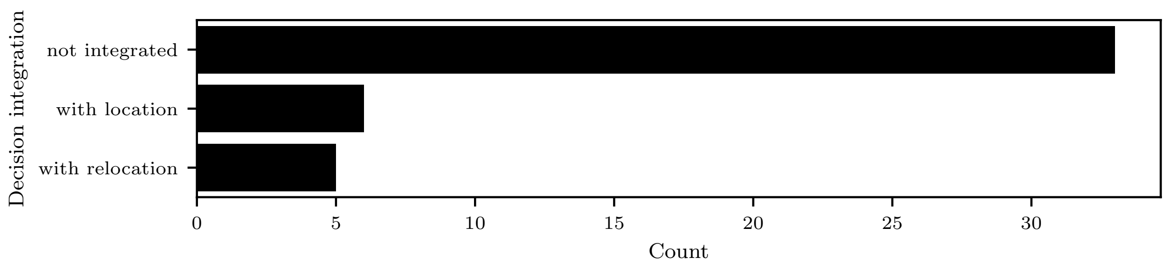


Figure 7.7: Distribution of decision integrations

Decision scope

The reports study the dispatching decision with different *decision scopes*. The decision scope refers to which actions the decision entails. As we can see in Fig. 7.8, the most commonly considered action is the *resource choice*. The resource choice determines which specific resource of a given type is dispatched to a call. Some reports additionally consider the action of not dispatching a resource immediately but instead queuing the call until the system is in a certain state, denoted by *dispatch or wait*. EMS systems can operate with multiple resource types, where more than one resource type can be compatible with a call. In these systems, the considered action can be the *resource type* to be dispatched. Most approaches to modeling the dispatching decision assume that a dispatcher has to assign one incoming call to one or multiple resources out of a set of available resources. If calls are actively queued, or if the EMS system is overloaded, the situation can reverse, and one resource has to be assigned to one out of a set of unserved emergencies, leading to the action of the *call choice*.

The scope that should be considered in the dispatching decision depends on the system under study. Considering the *dispatch or wait* action requires the regulatory option to not immediately respond to an incoming call. Choosing the *resource type* is only required in systems with multiple resource types. The *call choice* appears less relevant in EMS systems, which do not actively queue emergencies and which are not expected to regularly operate under overload conditions where no resources are available when a call occurs.

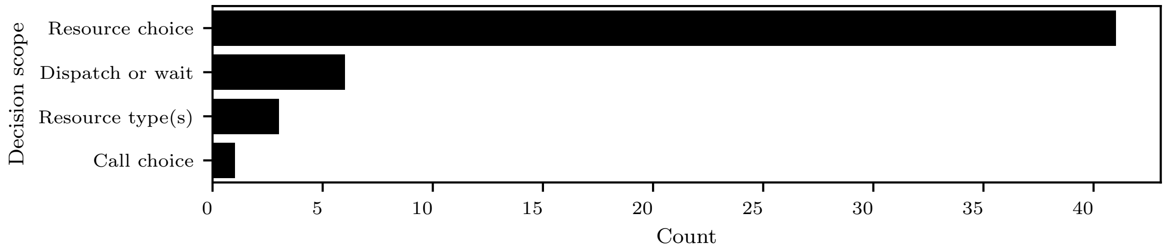


Figure 7.8: Distribution of decision scopes

7.3 Reviewing Dispatching Strategies: Discussion

The aims of this chapter were, firstly, to gain an overview of existing dispatching strategies, and, secondly, to identify attributes of dispatching strategies on which a selection of a dispatching strategy for an EMS system can be based. We conducted a scoping review following the PRISMA-ScR guidelines to achieve these aims, and discuss our approach and results in the following.

Contributions

To the best of our knowledge, we conducted the first scoping review of ambulance dispatching strategies. While existing reviews of logistic decisions in EMS systems encompass dispatching strategies, they have either not been conducted systematically or have not focused on dispatching strategies. We additionally propose a set of attributes based on which dispatching strategies can be classified. These attributes are the optimization approach, the number of resource types, the resource system characteristics, the number of call types, the call type differentiators, the decision integration, and the decision scope.

Practical Implications

The systematic review enables decision makers to narrow down dispatching strategies suitable for the specific design of their EMS system. The proposed attributes can furthermore help to classify and compare novel dispatching strategies.

Limitations

The following limitations should be considered when interpreting the results of the systematic review. Several reports were not accessible, either because the record was ineligible by the employed reference management software Zotero, or because the full reports could not be retrieved. The choice of the databases Web of Science, Scopus, and PubMed excludes any reports from the review that are not listed in one of these databases. Lastly, the review has been conducted by a single reviewer. Conducting the selection process with multiple reviewers would have reduced the risk of bias or misclassifications of records and reports.

Methodological Reflections

The search string could have been refined further, especially through examining the usage of abbreviations such as "ems" or "mdp" in other research streams. The abbreviation "ems", for instance, also stands for "Energy Management System" and "Environmental Energy System", which led to a multitude of records being excluded in the record screening phase. Six out of the 43 included reports were identified through forward and backward search. This indicates that the search term might have been refined further initially to capture more eligible reports. Lastly, we followed the PRISMA-ScR guidelines, which are primarily designed for studying health interventions. They have been applied in a similar review of logistic decisions in EMS, and we found no reason to adapt them for our review, but other guidelines for scoping reviews could still have been examined and compared.

Outlook

Identifying a suitable dispatching strategy for a specific EMS system is challenging for two main reasons, which both present avenues for further research. Firstly, the characteristics of the EMS systems for which dispatching strategies have been designed differ in multiple aspects, such as call types and considered resources, but no classification scheme for EMS systems was found in the literature. This necessitates examining a report in detail to investigate whether the proposed dispatching strategy is suitable for a specific EMS system. Developing a classification scheme for EMS systems from a logistic point of view could help to more easily identify suitable dispatching strategies. Secondly, even if dispatching strategies are designed for EMS systems with equal attribute values, we found no systematic comparison to alternative dispatching strategies in the reports. One possible reason for this is the lack of available benchmark instances, which would enable a more systematic comparison of novel dispatching strategies.

7.4 Reviewing Dispatching Strategies: Conclusion

We conducted a scoping review of dispatching strategies for EMS resources to identify existing approaches and their attributes. The review identified 43 reports presenting original dispatching strategies for routine EMS operations.

Our review reveals differences in dispatching strategies across six key attributes: optimization approach, resource types, resource system characteristics, call types, call type differentiators, decision integration, and decision scope.

We addressed a gap in the literature by conducting, to the best of our knowledge, the first scoping review of dispatching strategies, resulting in a comprehensive overview of dispatching strategies and identifying their key attributes. The identified attributes enable systematic classification of dispatching strategies and support decision makers and researchers in selecting approaches suitable for their specific EMS system characteristics. For our investigation of the EMS in Baden-Württemberg, we can now systematically identify dispatching strategies that match the system characteristics: single resource type, multiple call types differentiated, and technical capabilities supporting both offline and online approaches. The review provides the foundation for selecting and adapting suitable dispatching strategies to incorporate patient categories and utility functions in the following chapter.

Chapter 8

Comparing Dispatching Strategies for Heterogeneous Patients

The dispatching decision is an operational decision in EMS and allocates ambulances to calls. The most common dispatching strategy in practice is the *Closest Available* strategy, which always dispatches the ambulance to an incoming call that is expected to reach the call location quickest. Fig. 8.1 illustrates an example in which this strategy can lead to undesired situations. According to the *Closest Available* strategy, an ambulance from location I is dispatched to call X. When call Y occurs, while the ambulance from location I is still servicing call X, an ambulance from location II is dispatched. If we assume that call X is less urgent compared to call Y, it would have been preferable to dispatch an ambulance from location II to call X to reduce the response time to call Y.

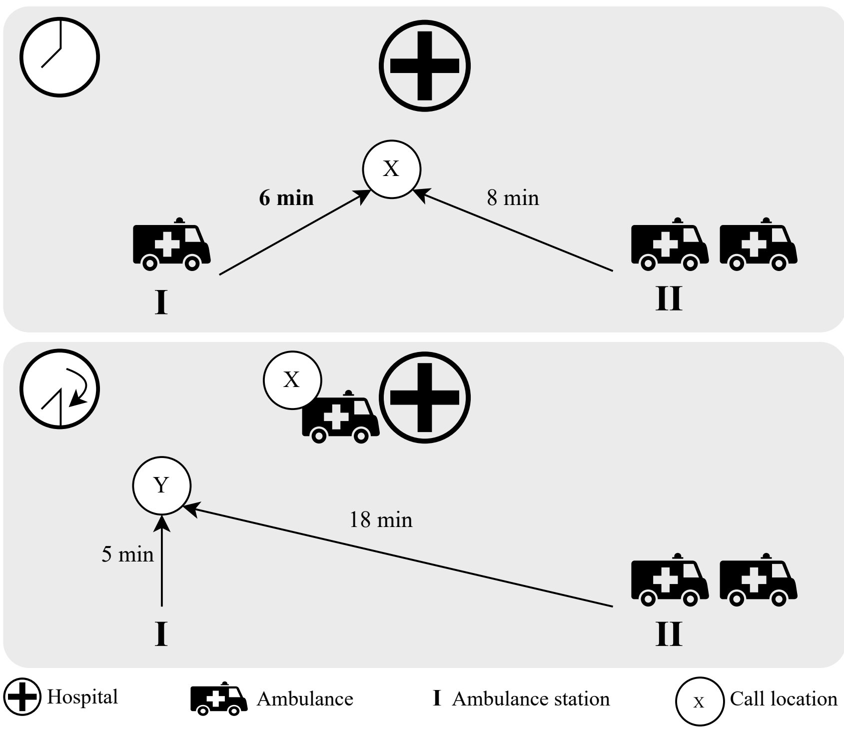


Figure 8.1: Exemplary illustration of the potential downside of the closest available dispatching strategy.

The example demonstrates that not dispatching the closest available ambulance, and therefore increasing driving times to some calls, can reduce driving times to other calls. The patient categories and utility functions developed in Chapter 5 indicate that we can distinguish calls by urgency. Let us initially distinguish between prioritized calls and non-prioritized calls in the following. We additionally assume to always dispatch the closest available ambulance to prioritized calls. This leads to the following questions that we address in this chapter:

- Which dispatching strategies are suitable for dispatching ambulances to non-prioritized calls in the EMS in Baden-Württemberg?
- How can the dispatching decision influence the driving times to and utilities of patients?

To answer these research questions, we firstly select suitable dispatching strategies based on the scoping review conducted in Chapter 7, and adapt several strategies to improve the comparability of strategies or to fit the patient categories and utility functions developed in Chapter 5. Secondly, we implement the selected dispatching strategies in the simulation model presented in Chapter 6 and parameterize the strategies based on historical data from the EMS in Baden-Württemberg. Lastly, we analyze the simulation results to assess the impact of the strategies on utilities and driving times. We focus our comparison on the initial dispatching decision when a call occurs and do not reevaluate dispatching decisions when ambulances become available after servicing a call.

Section 8.1 describes how we select dispatching strategies. We detail our selection process in Section 8.1.1, and introduce the selected dispatching strategies, including performed adaptations in Section 8.1.2. Section 8.2 specifies the experimental setup, including how we parameterize the dispatching strategies in Section 8.2.1, and how we set up the simulation experiments in Section 8.2.2. We analyze the results of the simulation experiments in Section 8.3. Section 8.3.1 presents the results of an initial comparison of dispatching strategies, and Section 8.3.2 follows with a more extensive analysis of effective strategies. We discuss our approach and findings in Section 8.4, and conclude the chapter in Section 8.5.

8.1 Selecting Dispatching Strategies

This section aims to select suitable dispatching strategies to decide which ambulance to dispatch to a non-prioritized call. We first describe the selection process based on the scoping review presented in Chapter 7 and then introduce the selected strategies and conducted adaptations.

8.1.1 Selection Process

We select dispatching strategies based on the scoping review presented in Chapter 7. We filter strategies based on attribute values that we deem fitting for the EMS in Baden-Württemberg and for investigating if utilities for prioritized calls can be improved by not necessarily dispatching the closest available ambulance to non-prioritized calls.

We select strategies with an online *optimization approach*. The location of each ambulance is available to the dispatchers in the EMS in Baden-Württemberg, making online approaches feasible. We limit strategies to consider a single *resource type* or multiple resource types with the *vehicle system characteristic* of single response. The second system design can easily be adapted to our purposes of studying ambulances by reducing the set of ambulance types to one. We do not limit strategies based on the number of *call types*, but consider only strategies with a *call type differentiator* of performance metrics or weight of performance metrics. The focus of this chapter lies exclusively on the dispatching decision, and we select only strategies where the dispatching decision is not integrated with other decisions, captured by the attribute *decision integration*. Lastly, we limit the comparison to strategies with a *decision scope* of the ambulance choice. Table 8.1 lists the described attribute values leading to a preliminary selection of strategies.

Table 8.1: Attribute value filters for selecting dispatching strategies for a comparison

Attribute	Value
Optimization approach	Online
Resource types	All
Resource system characteristics	- OR single response
Call types	All
Call type differentiator	- OR performance metrics OR weight of performance metrics
Decision integration	Not integrated
Decision Scope	Ambulance choice

Applying these filters results in a total of 15 dispatching strategies. We exclude eight of these strategies from the comparison for the following three reasons and include one additional dispatching strategy. Table 8.2 provides an overview of the included and excluded strategies and the reasons for exclusion or inclusion.

Four strategies focus on redispatching ambulances once the system status changes, for instance, when an ambulance is dispatched or when a new call occurs. The scope of the comparison in this chapter is the initial dispatching decision for non-prioritized calls.

Three further strategies require generating input parameters that are outside the scope of this thesis. Lopez et al. (2008) proposes an auction mechanism in a multi-agent system that centers around the trust in whether a specific ambulance crew can reach an emergency within the estimated driving times. The mechanism assumes that the experience level of ambulance crews influences driving times. Considering characteristics of ambulance crews is outside the scope of this thesis. Mukhopadhyay et al. (2019) proposes a real-time semi-Markov decision process approximation. For each dispatching decision, a state-action tree is computed to assess the effect of the decision on future states of the EMS system. Monte Carlo Tree Search is then applied to evaluate the impact on future system states. The state-action tree is based on a real-time incident prediction model as well as an environmental factor prediction model that contains both a traffic prediction model and a route finding algorithm. While there is no methodological reason to exclude this strategy, the necessary implementation and adaptation effort is outside the scope of this thesis. Lastly, Nasrollahzadeh et al. (2018) developed a strategy based on approximate dynamic programming (ADP). The authors report a runtime of two days of central processing unit time on an Intel Core i7 3.4 GHz processor with 16 GB of random access memory for an instance with 17 ambulances for one iteration of the ADP algorithm. The ADP algorithm requires multiple iterations and needs to be computed for each instance. The underlying objective of the proposed strategy is response time-based, while we focus on patient utilities. The ADP approach employs basis functions that are specifically designed to provide a good estimation of how a decision impacts the objective. We therefore would likely need to extensively adapt the approach, which is outside the scope of this thesis. The strategy additionally considers only ambulances at their station for dispatching, while ambulances returning to their station are considered for dispatching as well in the EMS in Baden-Württemberg.

Table 8.2: Result of the selection process and reasons for exclusion or inclusion.

Strategy	Reason for exclusion
T. Andersson & Värbrand (2007)	-
H. Li & Chai (2015)	-
S. Lee (2017)	-
Jagtenberg, Bhulai, & van der Mei (2017)	-
Amorim et al. (2018)	-
Aringhieri et al. (2018)	-
Roa et al. (2020)	-
Haghani et al. (2004)	Focus on redispatching
Meinzer & Storandt (2014)	Focus on redispatching
Billhardt et al. (2014)	Focus on redispatching
Theeuwes et al. (2021)	Focus on redispatching
Lopez et al. (2008)	Implementation outside thesis scope
Nasrollahzadeh et al. (2018)	Implementation outside thesis scope
Mukhopadhyay et al. (2019)	Implementation outside thesis scope
S. Lee (2011)	Author published an improved version of the same approach
Strategy	**Reason for inclusion**
Bandara et al. (2014)	Offline heuristic can be computed Online

One further strategy is excluded from the comparison. S. Lee (2011) published an initial preparedness-based dispatching strategy. S. Lee (2017) improved on this strategy but follows the same basic approach to taking the dispatching decision. We therefore only include the latter dispatching strategy in the comparison.
Bandara et al. (2014) introduced an algorithm to compute priority lists in an offline approach. We found, however, that it is computationally feasible to rerun the algorithm online for each dispatching decision with promising results and therefore include the strategy in the comparison.

8.1.2 Strategies and Adaptations

The following section introduces the eight selected dispatching strategies. Several common concepts and considerations apply to all strategies. We describe these concepts and considerations in the following section before we present the individual strategies.
Our approach to the dispatching decision is to always dispatch the closest available ambulance to prioritized calls, but not necessarily dispatch the closest available ambulance to non-prioritized calls. The dispatching strategies are therefore applied to select the ambulance to dispatch for non-prioritized calls. We refer to the call for which the dispatching decision is taken as the "call at dispatch".
All strategies are based on selecting an ambulance from a set of ambulances that are considered for dispatching to the call at dispatch. This set can, for instance, be defined based on response time targets. If the set is empty, we always dispatch the closest available ambulance. If no ambulance in the entire EMS system is available for dispatch, the call is queued and will wait for the next ambulance to become available.
The dispatching strategies base their computations on an aggregation of possible future call locations to demand nodes. A demand node is a discrete location in the region under study and allows computing, for instance, expected driving times from ambulance stations to future calls.
Several strategies employ the concept of coverage. Coverage is related to response time targets. A call or demand node is covered if an ambulance is expected to be able to reach it within the response time target.
Ambulances are not necessarily located at their station at the time of dispatch, but can, for instance, be returning from a hospital or call scene back to their station. Some strategies consider the expected driving time of ambulances to demand nodes in their computations. Let us consider two options to compute these driving times if an ambulance is not located at its station when the dispatching decision is taken. The first option is to compute the driving times from the current location to each demand node. We will refer to this location as the location at dispatch in the following. The second option is to use the driving time from the station of the ambulance to the demand nodes. Both options can misjudge the actual driving time at the point in time when the next call occurs, for which

the ambulance is considered for dispatching. If the next call occurs shortly after the current dispatching decision, the error is smaller when basing the driving times on the location at dispatch. If, however, sufficient time passes for the ambulance to be closer to its station than to the location at dispatch, the error is smaller when basing the driving times on the station locations. Basing the driving times on ambulance stations further allows precomputing the driving times, which makes the implementation and simulation experiments computationally more tractable. A quantitative comparison of the errors of both options would only be possible by implementing and comparing both variants, since the dispatching decision influences the locations of ambulances. This is outside the scope of this thesis, and we use the station locations to compute driving times from ambulances to demand nodes.

A further consideration is our naming scheme for the strategies. Some strategies, such as the Dynamic Maximum Expected Coverage Location Problem (DMEXCLP) by Jagtenberg, Bhulai, and van der Mei (2017), were published with a distinguishable name, but this is not the case for each strategy. We therefore use the last name of the first author of the publication of each strategy to refer to the dispatching strategy.

We unify the notations from the original publication to facilitate comparing the strategies. For each strategy, we present the central objective, the approach of the strategy, and the formal description. We adapt several strategies to make the strategies more comparable. If we adapt a strategy, we describe the adaptations and their reasons after the formal descriptions.

8.1.2.1 *Andersson*

T. Andersson and Värbrand (2007) developed the *Andersson* strategy, which aims to keep ambulances close to areas with a high expected call volume. The strategy is based on a preparedness index as a quantitative measure.

The strategy calculates a preparedness index P_i for each demand node $i \in V$ that is affected by the dispatching decision. The objective of the strategy is to maximize the lowest preparedness of demand nodes that are affected by the dispatching decision. The authors consider the L_i closest ambulances to a demand node i for the preparedness calculations. Each of these ambulances is assigned an index l from 1 to L_i in non-decreasing order of driving times to i. Indexing the ambulances based on driving times is necessary because the closer an ambulance is located to the demand node i, the more it contributes to its preparedness P_i. The contribution factor γ^l reflects this in the calculation. Next to the driving times, the demand proportion d_i is considered to give more weight to demand nodes with a high expected call occurrence. Table 8.3 provides the overview of the notation followed by the definitions of the preparedness index P_i in Eq. (8.1.1) and the formal description of the strategy in Algorithm 2.

Notation	Description
A_{idle}^{+}	Set of ambulances considered for dispatching to the call at dispatch
V	Set of demand nodes
L_i	Number of ambulances contributing to preparedness of zone i
d_i	Weight of demand node i
τ_{li}	Driving time from l to i
γ^l	Contribution factor of the l-th closest ambulance

Table 8.3: Notation for *Andersson*

The preparedness index P_i is calculated as follows:

$$P_i = \frac{1}{d_i} \sum_{l=1}^{L_i} \frac{\gamma^l}{\tau_{li}} \tag{8.1.1}$$

with:

$$\tau_{1i} \leq \tau_{2i} \leq \ldots \leq \tau_{L_i i} \tag{8.1.2}$$

$$\gamma^1 > \gamma^2 > \ldots > \gamma^{L_i} \tag{8.1.3}$$

Algorithm 2 Dispatching Strategy *Andersson*

1: Let $x^\star$ notate the ambulance that is dispatched and let $P_{min} = 0$.
2: **for** $x \in A_{idle}^{+}$ **do**
3: Remove ambulance x from the lists of ambulances contributing to the preparedness, and recalculate the preparedness, P_k, according to Eq. (8.1.1) in all demand nodes that are affected by this action, denoted by $\tilde{V}$.
4: **if** $\min_{k \in \tilde{V}} P_k > P_{min}$ **then**
5: $P_{min} = \min_{k \in \tilde{V}} P_k, x^\star = x$
6: Dispatch $x^\star$

T. Andersson and Värbrand (2007) consider only the L_i ambulances for dispatching, which are closest to the demand node at which the call at dispatch occurs. We consider all $|A_{idle}^{+}|$ ambulances within the respective threshold of the category of the call at dispatch to maintain consistency with the other dispatching strategies in the comparison.

8.1.2.2 *Lee*

The objective of the *Lee* strategy, proposed by S. Lee (2017), is to maximize a trade-off ratio between how well prepared the set of ambulances remains after dispatching a specific ambulance and how long the driving time of that ambulance is to the call at dispatch.
S. Lee (2017) define a preparedness measure P_A for a set of ambulances A. The basis for the preparedness measure is the driving time from the closest ambulance $x \in A$ to each demand node $i \in V$. The weighted average shortest driving time to the demand nodes is then calculated using the expected call rates λ_i of each demand node as weights. The

larger this weighted sum, the less prepared the set of ambulances is assumed to be. The reciprocal of the weighted sum is defined as the preparedness measure P_A to reflect the intuitive notion that a higher numeric value of the preparedness index corresponds with a better preparedness of the set of ambulances (s. Eq. (8.1.4)). To consider the driving time to the call at dispatch, the ratio of the preparedness index P_A and the function $1 + \tau_{x\hat{\imath}}$ of the driving time for ambulance x to reach the location $\hat{\imath}$ of the call at dispatch is computed. The calibration factor α allows controlling the relative weights given to the preparedness after dispatching and the driving time to the call at dispatch. Table 8.4 lists the notation for the subsequent formal description of the *Lee* strategy in Algorithm 3.

Notation	Description
A^+_{idle}	Set of ambulances considered for dispatching to the call at dispatch
A_{idle}	Set of ambulances idle at or moving back to their station
V	Set of demand nodes
S	Set of ambulance stations
λ_i	Call rate at demand node i
τ_{xi}	Driving time from x to i
α	Calibration parameter
$s(x)$	Station of ambulance x

Table 8.4: Notation for *Lee*

The preparedness for an ambulance fleet A is defined as:

$$P_A = \frac{1}{\sum\limits_{i \in V} \lambda_i (1 + \min_{k \in A} \tau_{s(k)i})} \tag{8.1.4}$$

Algorithm 3 Dispatching Strategy *Lee*

1: Let i be the location of the call at dispatch
2: Let A_{idle} be the set of idle ambulances
3: Let $P_{A_{idle}}$ be the preparedness level of A_{idle}
4: Let $\alpha > 0$ be a calibration parameter to tune, how much emphasis is put on the preparedness in relation to the driving time
5: **for** $x \in A^+_{idle}$ **do**
6: Calculate the preparedness level $P_{A_{idle} \setminus \{x\}}$ if x would be dispatched.
7: Dispatch $x^\star = arg\,max_{x \in A^+_{idle}} \frac{P^\alpha_{A \setminus \{x\}}}{1 + \tau_{xi}}$

8.1.2.3 *Roa*

Roa et al. (2020) propose the *Roa* strategy that aims to maximize the availability of ambulances within the response time thresholds of all categories.

The dispatching decision is based on a preparedness index of the EMS system. The preparedness index of the system is calculated as the sum of preparedness values P_i of the demand nodes $i \in V$. The preparedness values P_i explicitly consider all categories $\varepsilon \in E$, and not only prioritized patients. Calculating the utilization of ambulances caused by serving calls of category ε at demand node i is at the core of the preparedness calculation. The utilization is derived from modeling each demand node as a (M/M/n) queuing system with exponential inter-arrival and service times, and n servers. This allows calculating the utilization based on the arrival rate $\lambda_{i\varepsilon}$ and service rate μ_ε of calls of category ε at demand node i, as well as the number n^c_ε of ambulances within the response time threshold of call category ε. The utilizations resulting from serving the call categories are aggregated by calculating the weighted sum $\sum\limits_{\varepsilon \in V} \alpha_{\varepsilon i} \left(\lambda_{i\varepsilon} / n^c_\varepsilon \mu_\varepsilon\right)$, where $\alpha_{\varepsilon i}$ is the proportion of calls of category ε at demand node i. Roa et al. (2020) do not explicitly argue why the weighting factor $\alpha_{\varepsilon i}$ is necessary, considering that the proportions of call volumes are implicitly contained in the call rates $\lambda_{i\varepsilon}$. Subtracting this utilization measure from 1 results in an availability measure of ambulances at the demand node i. A lower bound of 0 is set for the preparedness P_i of a demand node i since this availability measure can be negative depending on the call rates, service rates, and number of covering ambulances. The preparedness index for the system is calculated at the time of dispatching, before choosing an ambulance to dispatch. The strategy then computes by how much the preparedness index would deteriorate when dispatching an ambulance $x \in A^+_{idle}$ and dispatches the ambulance that deteriorates the preparedness index the least. Table 8.5 provides the notation for the following formal description of the strategy in Algorithm 4.

Notation	Description
A^+_{idle}	Set of ambulances considered for dispatching to the call at dispatch
A	Set of ambulances
V	Set of demand nodes
E	Set of call categories
$d_{i\varepsilon}$	Weight of demand node i for call category ε
λ	Call rate
μ	Service rate
T_ε	Driving time threshold for category ε
n^c_ε	Number of idle ambulances which are able to reach $i \in V$ within T_ε

Table 8.5: Notation for *Roa*

The weight of call category ε is calculated with Eq. (8.1.5).

$$\alpha_{\varepsilon i} = \frac{d_{i\varepsilon}}{\sum\limits_{\varepsilon' \in E} d_{i\varepsilon'}} \quad \forall \varepsilon \in E, i \in V \tag{8.1.5}$$

The preparedness value of a demand node i is calculated with Eq. (8.1.6).

$$P_i = \max \left\{ 0; 1 - \sum_{\varepsilon \in V} \alpha_{\varepsilon i} \left(\frac{\lambda_{i\varepsilon}}{n_\varepsilon^c \mu_\varepsilon} \right) \right\} \quad \forall i \in V \tag{8.1.6}$$

Algorithm 4 Dispatching Strategy *Roa*

$TotalInitialPreparedness \leftarrow$ Compute total initial preparedness of demand nodes $i \in I$ as $\sum\limits_{i \in I} P_i$
for all $x \in A_{idle}^{+}$ **do**
  choose x as candidate ambulance to be dispatched to i
  $NewTotalPreparedness \leftarrow$ Compute new total preparedness of demand nodes i as $\sum\limits_{i} P_i$ based on $A \backslash \{x\}$
  $MarginalImpact_x \leftarrow TotalInitialPreparedness - NewTotalPreparedness$
Dispatch the ambulance $x \in A_{idle}^{+}$ with the lowest $MarginalImpact_x$

We adapted the strategy from the original publication as follows. Roa et al. (2020) studied an EMS system with two types of ambulances, where it is preferred to dispatch the better-equipped ambulance to prioritized calls and no differentiation between ambulances is considered for non-prioritized calls. We omit considering ambulance types for prioritized calls and adopt the part of the dispatching strategy that selects an ambulance to dispatch to non-prioritized calls.

8.1.2.4 *Jagtenberg*

The intuition behind the dispatching strategy *Jagtenberg*, introduced by Jagtenberg, Bhulai, and van der Mei (2017), is to maximize the remaining expected coverage of prioritized calls after dispatching an ambulance to a non-prioritized call.

The strategy computes the marginal contribution of each ambulance $x \in A_{idle}^{+}$ to the expected coverage of prioritized calls. The marginal contribution of an ambulance x is equal to the weighted sum of marginal contributions to each demand node $i \in V$ which can be reached by x within the response time threshold $T_{\mathcal{A}}$ of prioritized calls. The marginal contribution to a demand node is the probability that x is the only available ambulance within $T_{\mathcal{A}}$. This probability depends on the global busy fraction q of ambulances. The probability of x being the only available ambulance is the joint probability of x not being busy and all other covering ambulances being busy. Let us denote the number of ambulances in the set A_{idle} of idle ambulances that can reach x within $T_{\mathcal{A}}$ as $n^c(i, A_{idle})$.

We then can calculate the marginal contribution to a demand node i as

$$(1-q)q^{n^c(i,A_{idle})-1}$$

. The marginal contribution of ambulance x can now be calculated as the sum of the marginal contributions to all demand nodes within a driving time of at most $T_{\mathcal{A}}$ weighted by the proportion of prioritized calls d_i expected to occur at i. Table 8.6 lists the notation for the formal description of the dispatching strategy *Jagtenberg*.

Notation	Description
A^+_{idle}	Set of ambulances considered for dispatching to the call at dispatch
A^-_{idle}	Set of infeasible ambulances. Defined as ambulances in a dispatchable state that are not able to reach the call at dispatch within its response time threshold
A_{idle}	Set of ambulances idle at or moving back to their station
V	Set of demand nodes
$\mathcal{A}$	Prioritized category to which always the closest idle ambulance is dispatched
$T_{\mathcal{A}}$	Driving time threshold for category $\mathcal{A}$
q	Busy fraction
d_i	Weight of demand node i
$n^c(i, A_{idle})$	Number of idle ambulances able to reach $i \in V$ within $T_{\mathcal{A}}$
τ_{is}	Driving time from i to s
$s(x)$	Station of ambulance x

Table 8.6: Notation for *Jagtenberg*

We can formally describe the dispatching strategy *Jagtenberg* as follows: If $A^+_{idle} \neq \emptyset$, select the ambulance to dispatch according to Eq. (8.1.7), otherwise dispatch the closest available ambulance.

$$arg\,min_{x \in A^+_{idle}} \sum_{i \in V} d_i(1-q)q^{n^c(i,A_{idle})-1} \cdot \mathbb{1}_{\tau_{is(x)} \leq T_{\mathcal{A}}} \tag{8.1.7}$$

We slightly adapt the strategy for the following experiments from its original publication as follows: If $A^+_{idle} = \emptyset$, the closest available ambulance is dispatched instead of dispatching the ambulance according to Eq. (8.1.8). The reasons are firstly, that the threshold determining A^+_{idle} is chosen as being the largest acceptable response time for the specific dispatching category. If this threshold cannot be kept, the time by which it is exceeded should be kept minimal, which is achieved by dispatching the closest available ambulance. Secondly, dispatching according to Eq. (8.1.8) could have led to exceedingly long travel times due to the size of the simulated regions.

$$arg\,min_{x \in A^-_{idle}} \sum_{i \in V} d_i(1-q)q^{n^c(i,A_{idle})-1} \cdot \mathbb{1}_{\tau_{is(x)} \leq T_{\mathcal{A}}} \tag{8.1.8}$$

8.1.2.5 *Amorim*

Amorim et al. (2018) proposed the *Amorim* strategy. Our adaptation of the strategy intends to maximize the overall utility of both the call at dispatch and expected future calls.

The dispatching decision considers both the utility resulting from dispatching an ambulance $x \in A^{+}_{idle}$ to the call at dispatch and the expected utility in the system without ambulance x. We calculate the utility for the call at dispatch based on the utility functions developed in Chapter 5. The expected utility in the system, called system utility status $\mathcal{S}$, builds on calculating ambulance availabilities with a global busy fraction q. The probability that an ambulance x responds to a call at demand node $i \in V$ given that it is the k-th closest ambulance to i is calculated as the probability q^{k-1} that all closer ambulances are busy multiplied by the probability $1-q$ that x is available. The expected utility that ambulance x provides to demand node i for call category ε is therefore the product of the provided utility $u\left(\varepsilon, s(x)\right)$ and the probability that ambulance x responds to a call at i. The sum of the expected utilities provided to each demand node i for category ε weighted by the proportion $d_{i\varepsilon}$ of calls of category ε occurring at i equates to the system utility status $\mathcal{S}$, formally defined in Eq. (8.1.9). Table 8.7 defines the notation for the *Amorim* strategy and Algorithm 5 formally describes the strategy. Subsequently, we detail our adaptations from the strategy as it was proposed by Amorim et al. (2018).

Notation	Description
A_{idle}	Set of ambulances idle at or moving back to their station
A^{+}_{idle}	Set of ambulances considered for dispatching to the call at dispatch
V	Set of demand nodes
E	Set of call categories
x_s	Ambulance stationed at station s
$s(x)$	Station of ambulance x
τ_{si}	Driving time from s to i
$d_{i\varepsilon}$	Weight of demand node i
z_{kis}	1, if s is the k^{th} nearest station to i, 0 otherwise
$u\left(\varepsilon, \tau\right)$	Utility for category ε for driving time t
q	Busy fraction

Table 8.7: Notation for *Amorim*

Calculation of the System Utility Status $\mathcal{S}$:

$$\mathcal{S} = \sum_{i \in V} \sum_{x \in A_{idle}} \sum_{k \in \{1,\dots,|A_{idle}|\}} \sum_{\varepsilon \in E} d_{i\varepsilon} u\left(\varepsilon, \tau_{s(x)i}\right) (1-q)\, q^{k-1} z_{kis(x)} \tag{8.1.9}$$

Algorithm 5 Dispatching strategy *Amorim*

1: **for** $x \in A^{+}_{idle}$: **do**
2: calculate the reduced System Utility Status $\mathcal{S}'$ with Eq. (8.1.9) where A_{idle} is replaced by $A_{idle} \setminus \{x\}$
3: calculate $Q_x = u\left(\varepsilon, \tau_{xi}\right) + \mathcal{S}'$
4: Dispatch the ambulance with the highest Q_x

We performed the following adaptations from the strategy as it was published by Amorim et al. (2018), partly to make the strategy more comparable to the other strategies in our comparison and partly due to formal ambiguities in the publication.

Amorim et al. (2018) use survival functions as functions of the response time in the form $\left(e^{c^{\varepsilon} \cdot r}\right)^{-1}$, where e is the Euler constant, c^{ε} is a parameter defining the shape of the survival function for category $\varepsilon \in E$, and r is the response time. We exchange the survival functions with the utility functions developed in Chapter 5 and accordingly rename $\mathcal{S}$ from System Survival Status to System Utility Status. Using utility functions aims at making the dispatching strategy more comparable to the other strategies in the comparison, which are partly based on response time thresholds derived from the utility functions.

Amorim et al. (2018) model distinct periods and define response times and call proportions dependent on the period. We omit modeling periods to ensure consistency with the other strategies in the comparison, which do not model periods.

Lastly, we base the availability of ambulances on a global busy fraction in line with similar approaches such as Jagtenberg, Bhulai, and van der Mei (2017). We assume that the publication of Amorim et al. (2018) intends to use station-specific busy fractions, but there is ambiguity in the formal description, which hinders adopting the station-specific busy fractions as we reason in the following. We use the notation from the original publication in the following paragraph to stay consistent with the original publication and argue our perception of ambiguity.

The System Survival Status at period s is defined as

$$S^p = \sum_i \sum_s \sum_k \sum_{\varepsilon} d_i^{\varepsilon s} (e^{c^{\varepsilon} \times r^p_{is}})^{-1} \times (1 - q_s) \times q^{k-1} \times z_{kis}$$

.

The first ambiguity is due to the index s. $d_i^{\varepsilon s}$ is defined as the predicted number of calls of category ε during period s. At the same time, z_{kis} is defined as 1, if s is the k-th nearest station to demand node i. It is therefore unclear whether s refers to a station or a period. This is problematic since the busy fraction q_s in the calculation of the System Survival Status depends on the subscript s, and it is not defined what s refers to in the context of the busy fraction. Furthermore, a calculation for q is provided as

$$q = \frac{\sum_{\varepsilon} d_i^{\varepsilon p}}{a_s}$$

. We assume that the subscript s has been omitted from q by accident, and the calculation most likely refers to calculating the busy fraction for station s. A definition for a_j is provided as the number of idle vehicles allocated at station s. The subscript j is not defined, and we assume it should be a_s. Since $d_i^{\varepsilon s}$ is defined as the predicted number of calls of category ε during period s, we assume that p refers to the period in the calculation of q. There is, however, no explanation for which value p should be set to. A further open question is how the parameter i is to be set in the calculation of q, for which no explanation is provided. Lastly, the definition implies that the busy fraction is computed as the ratio of the number of calls and the number of ambulances. Conceptually, this is only intuitive if the number of calls refers to the number of calls occurring during the average service time required by an ambulance to serve a call. Otherwise, the busy fraction would change depending on the length of the period, which appears counterintuitive.
Further ambiguities arise in the calculation of q^{k-1} which is presented as

$$q^{k-1} = \prod_{j=1}^{k-1} q_j$$

. If we follow our previous assumption that the subscript of q refers to a station, this would imply that q^{k-1} is calculated as the product of the busy fractions of stations with indices $1, \ldots, k-1$. If the busy fraction is intended to vary between stations, this would imply that the value of q^{k-1} depends on how indices are assigned to stations, which seems counterintuitive. If j does not refer to the index of a station, no explanation is provided as to what it refers to in the calculation of q^{k-1}. The intention most likely is to calculate the probability that all $k-1$ stations which are closer to the demand node i than station s are busy. This would, however, necessitate making the calculation dependent on both the station s and demand node i, which is not the case in the publication.
Due to the presented ambiguities in the publication, we use a global busy fraction q in Eq. (8.1.9) in line with the concept of expected coverage employed by Jagtenberg, Bhulai, and van der Mei (2017).

8.1.2.6 *Li*

H. Li and Chai (2015) propose dispatching the ambulance $x \in A^+_{idle}$ with the lowest utilization. The authors, however, do not specify how the utilization is calculated. We interpret the utilization to be the average utilization of an ambulance under the *Closest Available* strategy. The intuition behind this interpretation is to dispatch the ambulance which is least likely to be the closest ambulance to an incoming prioritized call. The more often an ambulance is the closest to incoming calls, the more often it would be dispatched under the *Closest Available* strategy, and the higher its utilization would consequently be. The dispatching decision influences the utilizations of ambulances. Another possibility would therefore be to use the utilization of some defined time interval preceding the dispatching decision. This would likely be more effective in balancing the utilizations of ambulances, but could be less effective in reducing driving times to prioritized calls.

We formally describe the dispatching strategy in Eq. (8.1.10) and provide the notation in Table 8.8.

Notation	Description
A^{+}_{idle}	Set of ambulances considered for dispatching to the call at dispatch
q_x	Utilization of ambulance x

Table 8.8: Notation for *Li*

Dispatch according to:

$$arg\,min_{x \in A^{+}_{idle}} \{q_x\} \tag{8.1.10}$$

8.1.2.7 *Bandara*

Bandara et al. (2014) introduced what we denote as the *Bandara* strategy. The fundamental idea of the strategy is to dispatch the ambulance that is the least likely to be the closest ambulance to the next occurring prioritized call.

First, we identify for each ambulance $x \in A^{+}_{idle}$ to which demand node $i \in I$ it is the closest, second-closest,..., $|A^{+}_{idle}|$th closest ambulance. This allows calculating the total proportion of expected calls to which the ambulance is the k-th closest one by considering the proportions d_i of calls at each demand node. The ambulances are then sorted in increasing order of the proportion of calls to which they are the closest. Ties are broken by considering the proportion of demand to which the ambulances are the second-closest ones. Or, more generally: Ties while sorting ambulances by the proportion of calls to which they are the k-th closest one are broken by considering the proportion of calls to which they are the $k+1$-th closest ambulance. The ambulance at the first position of this sorted list of ambulances is dispatched.

The notation is listed in Table 8.9, and the algorithm is formally described in Algorithm 6.

Notation	Description
A^{+}_{idle}	Set of ambulances considered for dispatching to the call at dispatch
V	Set of demand nodes
τ_{is}	Driving time from i to s
$s(x)$	Station of ambulance x
d_i	Weight of demand node i

Table 8.9: Notation for *Bandara*

Algorithm 6 Dispatching Strategy *Bandara*

Let $m = |V|$ be the total number of demand nodes and $n = |A^{+}_{idle}|$ be the total number of feasible ambulances.
Step 1:
let $x_1, x_2, x_3, \ldots, x_k, x_n$ be a permutation of $(1, 2, 3, \ldots, n)$ and $\tau_{is(x_k)}$ be the travel time from the station of ambulance x_k to demand node i
for all i, where $i = 1, 2, 3, \ldots, m$ **do**
    rank travel times $\tau_{is(x_k)}$ as $\tau_{is(x_1)} \leq \tau_{is(x_2)} \leq \tau_{is(x_3)} \leq \ldots \leq \tau_{is(x_k)} \ldots \leq \tau_{is(x_n)}$
    Let $g_{ij} \in G$ be an $(m \times n)$ matrix, where the ith row of G denotes $x_1, x_2, x_3, \ldots, x_k, x_n$
Step 2:
Consider the row vector d where d_i denotes the proportion of calls from demand node i, $i = 1, 2, 3, \ldots, m$
Step 3:
Let $b_{kj} \in B$ be the heuristic proportion matrix of size $(n \times n)$
for $j = 1, 2, 3, \ldots, n$ **do**
    for $k = 1, 2, 3, \ldots, n$ **do**
        $sum = 0$;
        for $i = 1, 2, 3, \ldots, m$ **do**
            if $g_{ij} == k$ **then**
                $sum = sum + d_i$
        $b_{kj} = sum$
Step 4:
Let $seq = 1, 2, 3, \ldots, n$ be a column vector and $B' = adjoin(B, seq)$
Do priority sort on B' in sequential order to permute the last column of B'. The last column of B' gives the dispatching order of ambulances.

While Algorithm 6 is intended to compute a static priority list of ambulances for each demand node, it is computationally feasible to recalculate the priority order for each dispatching decision based on the available ambulances at the time of dispatch. In the implemented dispatching strategy, n therefore represents the number of ambulances that can reach the emergency within its threshold, which means $n = |A^{+}_{idle}|$. If $n = 0$, the closest available ambulance is dispatched.

8.1.2.8 *Aringhieri*

Aringhieri et al. (2018) introduced the *Aringhieri* dispatching strategy. The aim of the strategy is to dispatch an ambulance from the station with the largest oversupply (or smallest undersupply) of ambulances.

To estimate the over- or undersupply of a station, an expected number A^e_s of ambulances at station $s \in S$ and the number A^a_s of actually idle ambulances of station s is determined. The strategy then selects the station with the largest difference between A^a_s and A^e_s as the station to dispatch from. The basis for calculating the expected number of ambulances A^e_s at a station is the proportion of calls to which the station is the closest one. Multiplying this proportion by the total number $|A|$ of ambulances in the set A of ambulances results in the expected number of ambulances A_s^e. Table 8.10 lists the notation used for the following formal description of the *Aringhieri* strategy.

Notation	Description
A	Set of ambulances
A_{idle}^{+}	Set of ambulances considered for dispatching to the call at dispatch
S	Set of ambulance stations
S_{idle}^{+}	Set of stations of ambulances which can be dispatched and which can reach the call at dispatch within its response time threshold
V	Set of demand nodes
V_s^c	Set of demand nodes, for which station $s \in S$ is the closest station, with $V_s^c \subset V$
d_i	Weight of demand node i
W_s	$\sum_{i \in V_s^c} d_i$: Sum of the weights of the nodes in V_s^c

Table 8.10: Notation for *Aringhieri*

If $A_{idle}^{+} = \emptyset$, the closest available ambulance is dispatched; otherwise, the station is selected as follows. If multiple ambulances are available at the selected station, the ambulance with the smallest index is dispatched.

$$arg\,max_{s \in S_{idle}^{+} : A_s^a > 0} \{A_s^a - A_s^e\} \tag{8.1.11}$$

with:

- A_s^a: number of idle ambulances of station s at the moment of the decision.
- A_s^e: estimated ambulances at station $s \in S$, calculated as

$$A_s^e = |A| \frac{W_s}{\sum_{i \in V} d_i} \tag{8.1.12}$$

8.2 Setting up the Experiments

The following section describes the experimental setup for the simulation-based comparison of the dispatching strategies. We present how we set the parameters of the dispatching strategies in Section 8.2.1 and describe the design of the simulation experiments in Section 8.2.2.

8.2.1 Dispatching Strategy Parameters

Implementing dispatching strategies requires setting their parameters. In the following, we firstly describe parameters shared by all strategies, followed by strategy-specific parameters. To conclude, Table 8.13 provides a comprehensive overview of how the parameters are set for the simulation experiments.

8.2.1.1 From Patient to Dispatching Categories

This chapter aims to investigate how the dispatching decision can impact the utilities of patients based on the patient categories and utility functions developed in Chapter 5.

We first define which patient categories are treated as prioritized in the dispatching decision. The patient categories 1 and 2 are defined as life-threatening situations, and we define these patient categories as prioritized. Accordingly, patient categories 3–6 are treated as non-prioritized in the dispatching decision.

We always dispatch the closest available ambulance to prioritized calls and therefore do not need to distinguish between patient categories 1 and 2. We therefore define dispatching category A to refer to prioritized calls in the dispatching decision.

The first step in dispatching an ambulance to a non-prioritized call is defining the set of ambulances out of which an ambulance is selected. We assume that this set should be defined such that the resulting driving times to non-prioritized calls still lead to acceptable utilities. This allows deriving time thresholds from the utility functions, given a utility value that is deemed acceptable. We assume that a utility value of 0.5 is deemed acceptable. The experts who participated in the expert consensus process to define the utility functions provided point estimations for utility values between 0 and 1. A utility value of 0.5 can therefore be interpreted as neither a high nor a low utility, leading to our choice as an acceptable threshold. The utility functions for patient categories 3, 5, and 6 are functions of the response time, and we can calculate the response time resulting in a utility value of 0.5. Table 8.11 lists the largest integer response times that result in a utility of at least 0.5 for the categories 3, 5, and 6. The utility function for patient category 4 depends on the prehospital time. We cannot directly derive a response time threshold from a prehospital time threshold, because the maximum response time with which a prehospital time threshold can be kept depends on both the on-scene time and the transport time to a hospital, which in turn depends on the locations of calls and hospitals. One possibility would be to derive demand-node-specific response time thresholds based on the locations of demand nodes and their closest hospital. The medical severity of patients in category 4 was deemed similar to the severity of patients in category 3, with the differentiator being whether patients can be treated at the scene or in the hospital. To facilitate the comparison of dispatching strategies, we therefore adopt the response time threshold of patient category 3 for patient category 4.

The presented assumptions result in the set E of four dispatching categories listed in Table 8.11. We name the categories alphabetically in increasing order of their response time thresholds. Dispatching category A represents the prioritized calls for whom we always dispatch the closest available ambulance, and dispatching categories B–D represent the non-prioritized calls for whom we dispatch ambulances according to one of the presented strategies. They differ in their response time thresholds that define the set of ambulances out of which one is selected for dispatch.

Patient category	Dispatching category	Response time threshold [min]
1	A	-
2		
3	B	23
4		
5	D	63
6	C	40

Table 8.11: Dispatching categories and response time thresholds

Next, we need to assign the dispatching categories to historical calls for the simulation experiments. This can be done using either the keyword-categories or the diagnosis-categories of historical calls, as described in Section 5.3.3. From a practical point of view, the keyword-category appears more suitable, since only the keyword is available at the time of dispatch. The experts in the consensus process, however, agreed that the keywords to which the categories are mapped were not designed with logistic categories in mind and, as such, do not represent how well the categories could be detected in a system designed based on dispatching categories. We therefore assign the dispatching categories to historical calls based on their diagnosis category.

8.2.1.2 Demand Nodes

The dispatching strategies base their computations on a graph representation, in which a set V of demand nodes represents possible emergency locations and a set S of supply nodes represents ambulance stations. Edge-weights in the graph store the travel times τ_{si} between supply nodes $s \in S$ and demand nodes $i \in V$. This necessitates defining the locations of the demand nodes. To determine the locations of demand nodes, administrative zones such as zip code areas have been used (Jagtenberg, Bhulai, & van der Mei, 2017). An alternative approach is to generate a grid structure and use the centroids of each grid cell as a demand node (Amorim et al., 2018). A downside of both approaches is that they only consider demand nodes within the geographic boundary of a region. In practice, however, cross-boundary use of resources occurs. Nair and Miller-Hooks (2009) propose the use of clustering algorithms. This approach is able to consider the locations of emergencies independent of geographic boundaries. For this case study, the demand node locations are calculated with the k-means clustering algorithm. Fig. 8.2 displays the resulting locations of demand nodes and historic call locations for an exemplary EMS region.

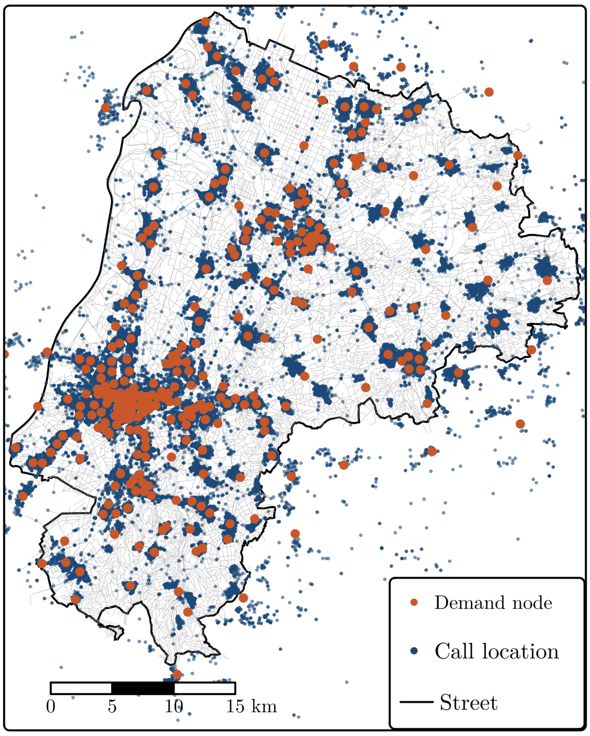


Figure 8.2: Exemplary demand nodes based on k-means clustering of historic call locations.

Using the k-means clustering algorithm results in a higher demand node density in regions with a high historic call volume and a lower demand node density in regions with a low demand node density. This implies that the representation of call locations is more accurate in regions with a high call occurrence and avoids demand nodes in regions with no or very few call occurrences, which would increase the computational effort. The number of clusters k in the k-means algorithm is equivalent to an average area of 4 km^2 per demand node based on the area of the EMS region. Considering call locations outside the administrative boundaries of the EMS regions will slightly increase the area per demand node. The demand node density is in line with existing case studies of the compared strategies, as listed in Table 8.12. The proportions d_i of demand and call rates λ_i are based on assigning calls to the closest demand node by Euclidean distance.

Publication	km^2 per demand node
T. Andersson & Värbrand (2007)	5
Roa et al. (2020)	14
Jagtenberg, Bhulai, & van der Mei (2017)	7
Amorim et al. (2018)	0.25
H. Li & Chai (2015)	10
Bandara et al. (2014)	5.5
Aringhieri et al. (2018)	3.5
This thesis	4

Table 8.12: Comparison of demand node density across studies providing information about the demand node density.

8.2.1.3 Further Common Parameters

All strategies in our comparison share the following additional parameters. The set A of ambulances and the locations of ambulance stations are both based on the historic station locations and shift schedules of ambulances in each EMS region. An ambulance $x \in A$ is considered dispatchable and part of the set A_{idle} of idle ambulances if it is in an active shift and located either at its station or driving back to its station from either a hospital or a call location. We additionally consider ambulances already dispatched and either driving to or preparing to drive to a call of dispatch categories $\varepsilon \in \{B, C, D\}$ for dispatch to a call in the dispatching category $\varepsilon = A$, as this occurs in practice. If an ambulance is redispatched from a call in dispatch categories B–D to a call in dispatching category A, a new ambulance needs to be dispatched to the call of dispatch categories B–D. This decision is taken by the respective dispatching strategy, but the response time threshold T_ε used to define the set of ambulances for dispatch is reduced by the already passed response time to avoid exceedingly long response times. Lastly, we calculate the driving times τ_{si} from stations $s \in S$ to demand nodes $i \in V$ with the speed model described in Section 6.3.2.3.

8.2.1.4 Strategy-Specific Parameters

The following parameters are only required for specific strategies. The strategies *Jagtenberg* and *Amorim* require a *busy fraction* q for their computations. We set one global busy fraction for ambulances stationed in an EMS region. It is set to the average utilization of ambulances in an EMS region when simulating the *Closest Available* strategy.

The strategy *Roa* uses a *service rate* μ. The hourly service rate is set to the reciprocal of the average service time when simulating the *Closest Available* strategy.

The concept of expected coverage of prioritized calls is at the center of the *Jagtenberg* strategy. This requires setting a response time threshold T_A for prioritized calls in dispatching category A. The current regulations for EMS in Baden-Württemberg contain response time thresholds for distinct patient categories and require a response time of 12 minutes for patients in life-threatening situations. Deriving the response time threshold from the utility function of patient category 1, analogous to dispatch categories B–D, would result in a threshold of 4 minutes. Such a low threshold would likely exclude most demand nodes from the coverage calculations, as only those demand nodes are considered that have at least one ambulance within their coverage threshold. We therefore consider a threshold of 12 minutes.

A *contribution factor* γ^l for the l-th closest ambulance to the call at dispatch is part of the computations in the *Andersson* strategy. The authors T. Andersson and Värbrand (2007) propose to set $\gamma^l = 1/2^{l-1}$, and we follow this proposal. The strategy additionally requires setting the number of ambulances contributing to the preparedness of a demand node. We follow the proposal of the authors to consider the closest seven ambulances for the preparedness calculations.

The *Lee* strategy requires a *calibration parameter* α. S. Lee (2017) varies the parameter from 0 to 2 in steps of 0.1. Carvalho et al. (2020) adopt the *Lee* strategy in a case-study and set the value of α to 1. An extensive sensitivity analysis for the impact of the value of *Lee* is outside the scope of this thesis, and we follow Carvalho et al. (2020) and set α to 1. Table 8.13 provides a concluding overview of the parameter settings.

Parameter	Value	Required for
A	Historical allocations and shift schedules	all
A_{idle}	All ambulances in an active state and either at or returning to their station	all
S	Historic locations of stations	all
V	k-means clustering based on historic call locations	all
E	A, B, C, D	*Roa*, *Amorim*
τ_{si}	Calculated with ambulance speed model developed in Chapter 6	all
q	Average utilization of ambulances in each EMS region based on simulating the *Closest Available* strategy	*Jagtenberg*, *Amorim*
d_i	Historic proportion of demand to which demand node i is closest by Euclidean distance	*Jagtenberg*, *Bandara*, *Aringhieri*, *Roa*, *Amorim*, *Andersson*
λ_i	Historic hourly call rates of calls to which demand node i is closest by Euclidean distance	*Lee*
μ	Average service rates based simulating the *Closest Available* strategy	*Roa*
$T_{\mathcal{A}}$	Based on 12 minutes response time target	*Jagtenberg*
T_ε	Derived from the utility function of category $\varepsilon \in E \backslash \{A\}$ and 12 minutes for $\varepsilon = A$	*Roa*
L_i	7	*Andersson*
γ^l	$\frac{1}{2^{l-1}}$	*Andersson*
α	1	*Lee*

Table 8.13: Overview of parameter settings for comparing dispatching strategies

8.2.2 Simulation Setup

We conduct two simulation experiments. The first experiment is an initial comparison of all dispatching strategies. Three strategies prove effective in reducing driving times to prioritized calls. We compare these three strategies in a second experiment for an extended analysis. Each simulation experiment consists of several simulation runs.

A simulation run is defined by its temporal and spatial extent. The temporal extent for the initial comparison is 01.01.2021–31.12.2021, and we conduct simulation runs for the regions KA and WT (s. Fig. 8.3). The temporal extent for the extended analysis is 01.01.2023–31.12.2023, and we conduct simulation runs for the regions North and South. The region North is located in the north of Baden-Württemberg and contains the EMS region KA and all neighboring EMS regions. The region South is located in the south of

Baden-Württemberg and contains the EMS region WT and all neighboring EMS regions. Table 8.14 lists the number of calls in each region during the spatial extents, as well as the call density.

Table 8.14: Description of instance characteristics for comparison of dispatching strategies

Region	Calls	Calls / km^2	Year
KA	64.797	52	2021
WT	15.745	14	2021
North	313.672	47	2023
South	113.119	25	2023

A simulation run simulates all calls that have historically been attended by an ambulance stationed in the simulated region. The calls are simulated at their historic date and time of occurrence as well as location. We employ the variance reduction technique of common random numbers as described by Law (2015, p.588 ff.). For each call response, we set the process times, such as pre-travel delays or on-scene times, equal to the historic process times, except for the driving times. In case of missing data, the process times have been imputed by drawing from the empirical distributions of existing process times. Only the driving times therefore differ between the simulation runs due to the dispatching strategy employed in the simulation run.

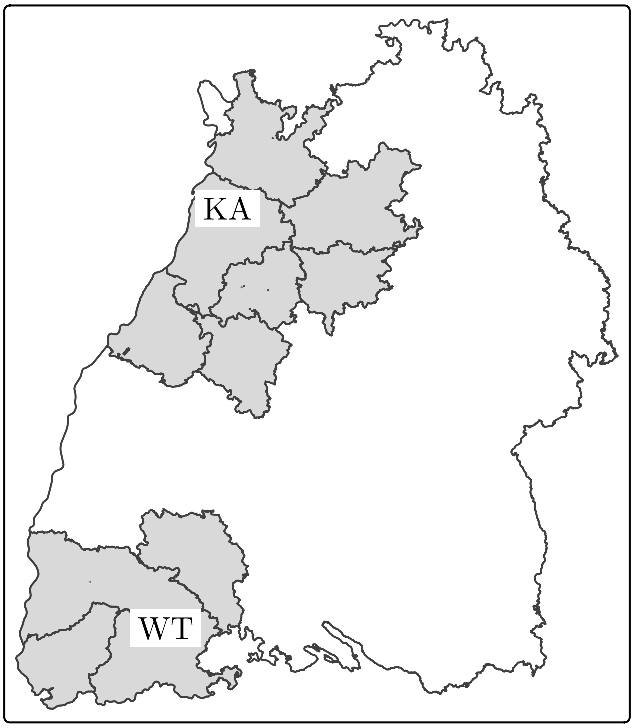


Figure 8.3: EMS regions in Baden-Württemberg on which the simulation experiments are based. The initial comparison is based on two singular EMS regions, and the subsequent analysis is based on the aggregated regions in the north and the south.

8.3 Analyzing the Results

This section presents the results of the simulation experiments comparing the selected dispatching strategies against the *Closest Available* strategy. The analysis of the results aims to investigate how the dispatching strategies influence patient utilities. The median and the 5th percentile of patient utilities are the main indicators of the analysis. The median captures the central tendency, and the 5th percentile quantifies the effects on the "worst-off" patients. We calculate the indicators based on the samples of simulated calls and assess the 95% confidence intervals to determine the statistical significance of differences. The utilities are functions of response or prehospital times. The dispatching decision immediately influences only the driving times of ambulances. To gain further insights into the effects on EMS operations, we additionally calculate the median and 95th percentile of driving times, including the 95% confidence intervals.

We compare all dispatching strategies in initial experiments to examine the impact on utilities of the prioritized calls in dispatching category A in Section 8.3.1. None of the strategies significantly improves patient utilities, but three strategies significantly reduce the median and 95th percentile of driving times in one of the studied regions.

Section 8.3.2 explores the results of extended experiments with these three strategies. Next to analyzing utilities and driving times for dispatching category A calls in Section 8.3.2.1, we also evaluate the effects on utilities and driving times of dispatching categories B–D in Section 8.3.2.2. Lastly, we investigate how sensitive the results are to the accuracy of categorizing calls at the time of dispatch in Section 8.3.2.3.

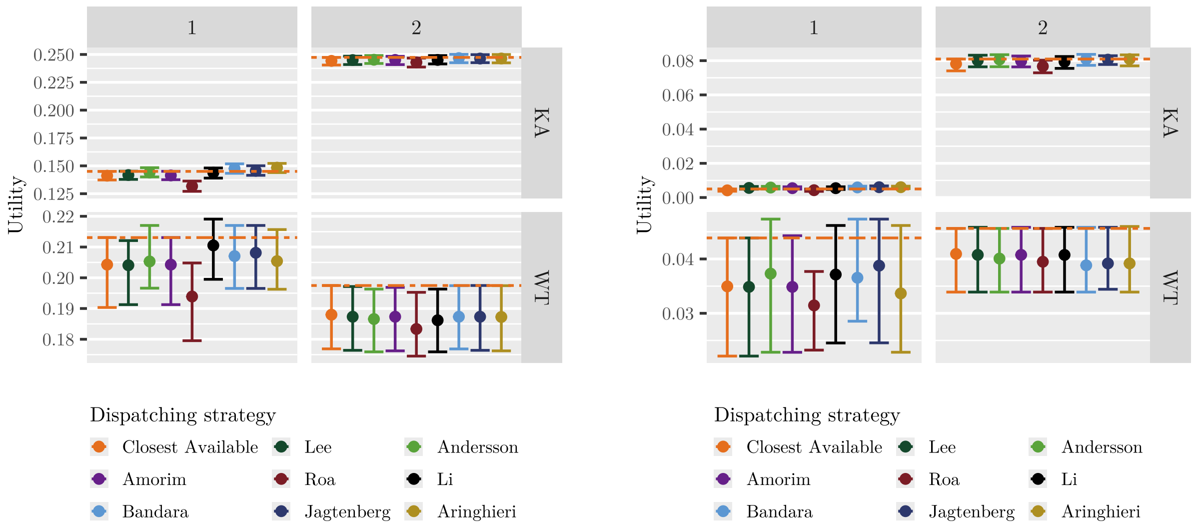
8.3.1 Initial Comparison of Strategies

The following section presents the results of the initial comparison of dispatching strategies for the two regions of KA and WT based on simulating calls in 2021.

Analyzing Utilities

Fig. 8.4 displays the median and 5th percentile of utilities for patient categories 1 and 2 in dispatching category A with 95% confidence intervals for the studied regions. The numeric results can be found for the medians in Table B.1 and the 5th percentiles in Table B.2.

None of the strategies significantly increases the patient utilities in any of the indicators compared to the *Closest Available* strategy. Several factors can limit the impact of the dispatching decision on patient utilities. The slope of the utility functions is low for low utility values, and even the median utilities are below 0.25 for all strategies. As a consequence, the reduction in response or prehospital times has to be sufficiently large to result in a significant change in utilities. Furthermore, the dispatching strategies are only applied to ambulances and not to EPV. Reducing ambulance response times, therefore, only has an effect if the ambulance arrives before the EPV. Simulating the dispatching strategies in isolated EMS regions might additionally limit the effects of dispatching strategies compared to simulating larger regions. Allowing longer response time thresholds for non-prioritized calls could lead to larger differences in larger regions compared to the *Closest Available* strategy. Lastly, the purpose of dispatching strategies is to avoid situations where no ambulance is available at the closest station to a prioritized call. The locations and allocations of ambulances can, therefore, limit the potential for improving patient utilities through the dispatching decision. The initial results do not indicate that patient utilities can be improved through the dispatching decision, but experiments in larger regions might lead to different results.



(a) Median (b) 5th percentile

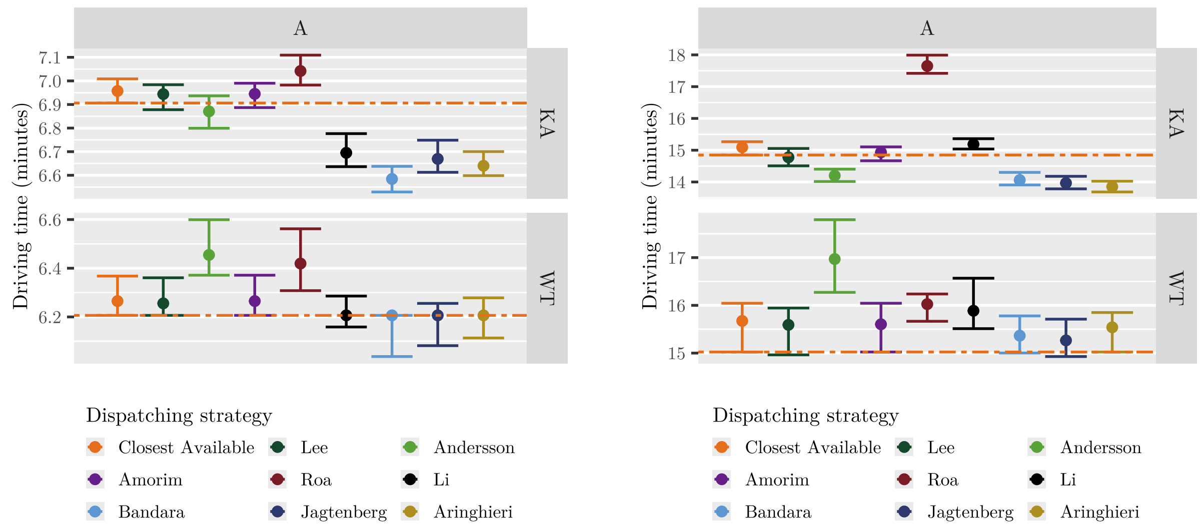
Figure 8.4: Utilities with 95% confidence intervals of simulation experiments for 2021. Different y-axis scales are used to improve legibility.

Another observation is that the *Roa* strategy significantly decreases the median utilities for category 1 patients in the region KA. The *Roa* strategy bases the dispatching decision on the aggregated impact on all categories (s. Eq. (8.1.6)) weighted by the proportions of calls in each category. As the prioritized calls make up the minority of the call volume, the impact of dispatching an ambulance on prioritized calls is weighted less than the impact on non-prioritized calls. This might lead to dispatching decisions resulting in a reduced utility for prioritized calls, and the results indicate that the *Roa* strategy is not suitable for the characteristics of the system under study in this thesis.

The lowest change in utilities is observed for the strategies *Lee* and *Amorim*. The *Lee* strategy contains a parameter with which the relative importance of the impact of the dispatching decision on the preparedness for future calls in comparison to the driving time to the current call can be controlled. We set the parameter value following an implementation of the *Lee* strategy by Carvalho et al. (2020), but performing a sensitivity analysis for the impact of this parameter on the utilities might lead to different insights. This is, however, outside the scope of this thesis. The *Amorim* strategy maximizes the sum of the utility for the current call and the expected average utility of future calls. Choosing the closest available ambulance will lead to the highest utility for the current call. If the drop in utility by choosing a non-closest ambulance is higher than the increase in the average expected utility for future calls, the closest available ambulance will be chosen. Both the *Lee* and *Amorim* strategies, therefore, likely often dispatch the closest available ambulance, leading to similar utilities to the *Closest Available* strategy.

Analyzing Driving Times

Fig. 8.5 displays the median and 95th percentile of driving times for dispatching category A calls with 95% confidence intervals. Table B.3 and Table B.4 list the numeric results.



(a) Median (b) 95th percentile

Figure 8.5: Driving times with 95% confidence intervals of simulation experiments for 2021. Different y-axis scales are used to improve legibility.

The strategies *Bandara*, *Jagtenberg*, and *Aringhieri* significantly reduce both the median and 95th percentile of driving times in the region KA compared to the *Closest Available* strategy. *Bandara* results in the largest reduction in the median with 23 seconds, which equates to a relative reduction of 5.5%. The largest reduction in the 95th percentile is caused by the *Aringhieri* strategy with 74 seconds, equating to a relative reduction of 8.2%. These significant reductions in driving times imply that the strategies dispatch ambulances to non-prioritized calls in such a way that closer ambulances are available for prioritized calls compared to the *Closest Available* strategy. While these reductions did not prove to be sufficient to significantly improve patient utilities, deviating from the closest available dispatching strategy can significantly reduce driving times to prioritized calls.

In contrast to the region KA, no significant reduction in any of the indicators can be observed in the region WT. WT is a more rural region compared to the region KA with a lower call density and a lower spatial ambulance density. A lower spatial density of ambulances leads to fewer ambulances being available within the response time thresholds.

For instance, there are an average of 10.3 ambulances available in the region KA within the response time thresholds for the dispatching categories B–D, while only an average of 5.3 ambulances are available in the region WT in the simulation experiment for the strategy *Bandara*. Fewer available ambulances within the response time thresholds provide fewer alternatives to dispatching the closest available ambulance, which may limit the impact of not using the *Closest Available* strategy. Preliminary experiments with the *Closest Available* strategy additionally showed that the average ambulance utilization in the region WT is lower than in the region KA. A lower utilization implies that ambulances are available at their station more often. The more often ambulances are available at their station when using the *Closest Available* strategy, the less potential there is to improve ambulance availability for prioritized calls by using a dispatching strategy other than the *Closest Available*. The difference in results between the regions KA and WT indicates that the effects of using a dispatching strategy other than the *Closest Available* strategy strongly depend on the characteristics of the region in which they are employed.

The largest difference in results between the regions KA and WT exists for the *Andersson* strategy, which significantly decreases the 95th percentile of driving times in the region KA, but significantly increases it in the region WT. The dispatching decision in the *Andersson* strategy depends only on the demand node with the lowest preparedness after dispatching. This might lead to a larger sensitivity of the strategy to characteristics of the regions compared to the other dispatching strategies.

The median of driving times for the *Bandara* strategy in the region WT coincides with the upper limit of its 95% confidence interval. This is caused by a subset of call responses with equal driving times. A sufficiently large subset of calls with equal driving times leads to the same values for the median and its upper limit of the confidence interval, as both are based on the driving times of calls in specific locations of the set of calls ordered by the driving time. The calls in the subset with equal driving times all occurred at the same location at a medical center in the region and are all served from the same ambulance station in the simulation, resulting in equal driving times.

The strategies *Bandara*, *Jagtenberg*, and *Aringhieri* all significantly decrease driving times in the observed indicators for the region KA, but they do not differ significantly from each other in any of the indicators. There is so far no indication as to which of these strategies is most effective for the EMS in Baden-Württemberg.

8.3.2 Analysis of Effective Strategies

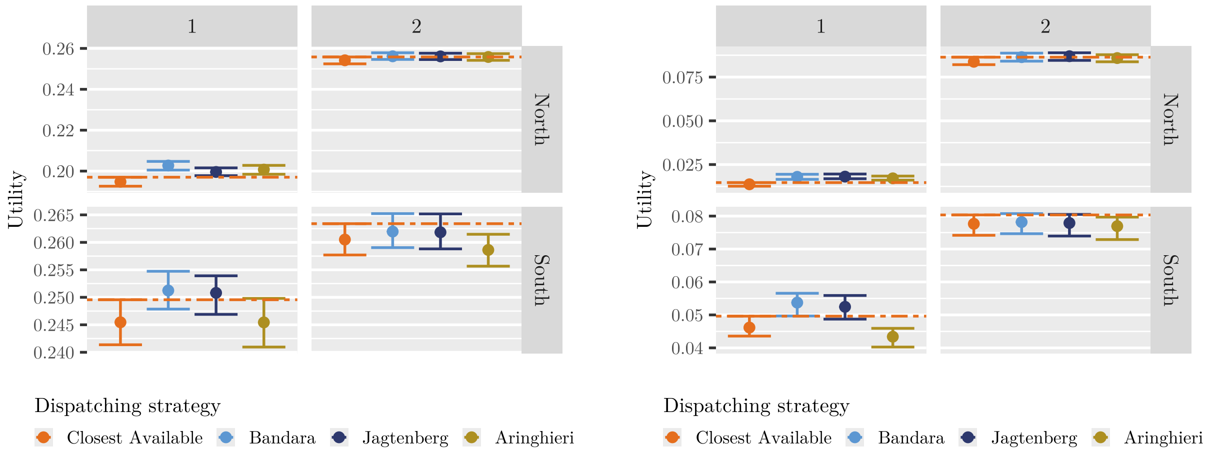
This section describes the results of comparing the dispatching strategies *Bandara*, *Jagtenberg*, and *Aringhieri* in the regions North and South based on simulating call responses in 2023. We first explore the effects on prioritized calls in Section 8.3.2.1. Secondly, we examine the impact on non-prioritized calls in Section 8.3.2.2. Lastly, we investigate the sensitivity to the categorization accuracy in Section 8.3.2.3.

8.3.2.1 Impact on Prioritized Calls

The following paragraphs analyze the impacts on the utilities of and driving times to prioritized calls.

Analyzing Utilities

Fig. 8.6 depicts the median and 5th percentile of utilities of patients in patient categories 1 and 2 for the studied regions North and South based on simulated calls in 2023. The associated numeric results can be found in Tables B.5 and B.6.



(a) Median (b) 5th percentile

Figure 8.6: Utilities with 95% confidence intervals of simulation experiments for 2023 for patient categories 1–2. Different y-axis scales are used to improve legibility.

All strategies significantly increase the median and the 5th percentile of utilities for category 1 patients in the region North. The *Bandara* strategy leads to the highest increase of 0.008 in the median compared to the *Closest Available* strategy, which is a relative increase of 3.9%. The 5th percentile is increased equally by the *Bandara* and *Jagtenberg* strategies by 0.004, equating to a relative increase of 22.2%. Simulating the strategies in larger regions increases the effects of not dispatching by the *Closest Available* strategy. This is supported by the observation of an average of 22.6 available ambulances within the response time thresholds for dispatching categories B–D for the *Bandara* strategy in region North (compared to 10.3 in KA in the initial comparison), and 11.7 in the region South (compared to 5.3 in WT in the initial comparison). The increase in sample size additionally leads to narrower confidence intervals, which aids in achieving statistically significant differences. The largest absolute increase in the median of utilities in the initial comparison was 0.007, and 0.002 in the 5th percentile. Employing the dispatching strategies in larger regions can therefore increase their impact.

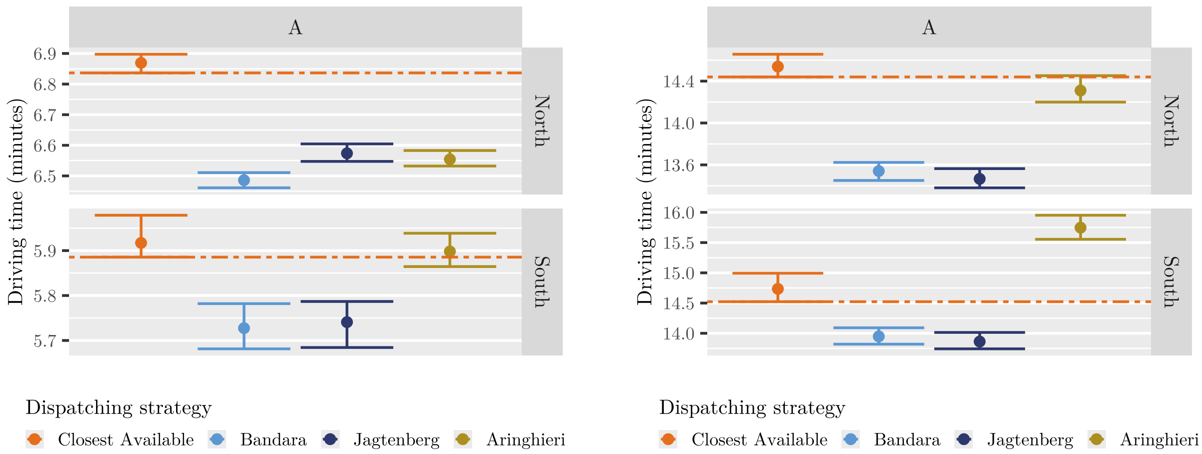
The utilities do not differ significantly in any of the indicators for the region South. The region South is a more rural region with a lower call and ambulance density. This leads to fewer ambulances being available within the response time thresholds of the dispatching categories B–D. Fewer available ambulances likely leads to dispatching the closest available ambulance more often, which results in smaller differences to the *Closest Available* strategy. This strengthens the insight from the initial comparison that the impact of dispatching strategies depends on the characteristics of the region in which they are employed.

In contrast to category 1 patients, no significant change in utilities for category 2 patients is observed in any of the indicators. The utilities for category 2 patients are a function of prehospital times. The dispatching decision impacts only the driving times, which constitute a relatively small part of the prehospital time. The impact of reduced driving times appears not to be sufficient to reduce the prehospital times in a magnitude that significantly increases utilities. Additional logistic decisions should therefore be examined if the aim is to improve the utilities for category 2 patients.

Even with the significant improvements in utilities for category 1 patients in the region North, the absolute utility values are low in the median, with all values below 0.265 and extremely low in the 5th percentile, with all values below 0.08. One likely reason is that the impact of the dispatching strategies on driving times is limited by the location and allocation of ambulances. Secondly, the utilities do not only depend on driving times, which further limits the potential impact of the dispatching decision. This indicates that not only the dispatching decision should be examined when aiming to improve utilities.

Analyzing Driving Times

After analyzing the impact on utilities, we assess the simulated driving times in the following. Fig. 8.7 depicts the median and 95th percentile of driving times to calls in dispatching category A. The numeric results can be found in Tables B.7 and B.8.



(a) Median (b) 95th percentile

Figure 8.7: Driving times with 95% confidence intervals of simulation experiments for 2023. Different y-axis scales are used to improve legibility.

Both strategies *Bandara* and *Jagtenberg* significantly reduce the driving times in all indicators and both regions. The largest decrease compared to the *Closest Available* strategy is 23 seconds in absolute, and 5.5% in relative terms, caused by the *Bandara* strategy in the region North. The 95th percentile is decreased most strongly by *Jagtenberg* in the region North with an absolute decrease of 64 seconds and a relative decrease of 7%. Both strategies succeed in dispatching ambulances to non-prioritized calls in such a way that closer ambulances are available more often for prioritized calls. Deviating from the *Closest Available* can therefore reduce the driving times to prioritized calls.

The *Aringhieri* strategy significantly decreases the median in the region North, leads to no significant differences in the median in the region South and the 95th percentile in the region North, and significantly increases the 95th percentile in the region South. This indicates that the effects of the *Aringhieri* strategy depend more strongly on the characteristics of the region than the effects of the *Bandara* and *Jagtenberg* strategies.

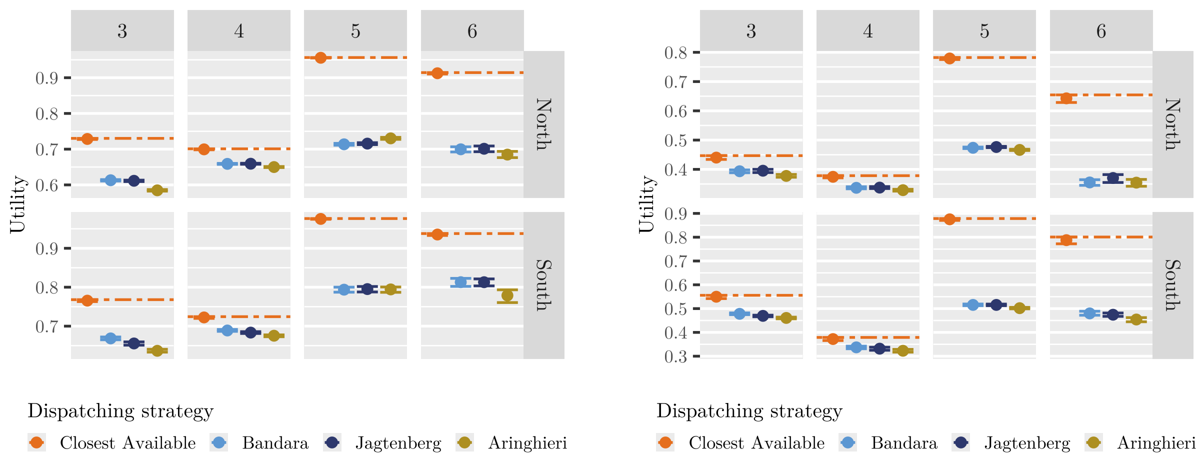
The *Bandara* strategy dominates the *Jagtenberg* and the *Aringhieri* strategies. It leads to significantly shorter driving times in the median in the region North, and it does not lead to significantly longer driving times in any of the other indicators. The *Bandara* strategy most effectively reduces driving times to prioritized calls in our experiments.

8.3.2.2 Impact on Non-Prioritized Calls

Our results demonstrate that there is potential to improve utilities for prioritized calls by not always dispatching the closest available ambulance to non-prioritized calls. We analyze the effects on the utilities and driving times of non-prioritized calls in the following.

Analyzing Utilities

Fig. 8.8 depicts the utilities of patients in patient categories 3–6. The numeric results are listed in Tables B.9 and B.10.



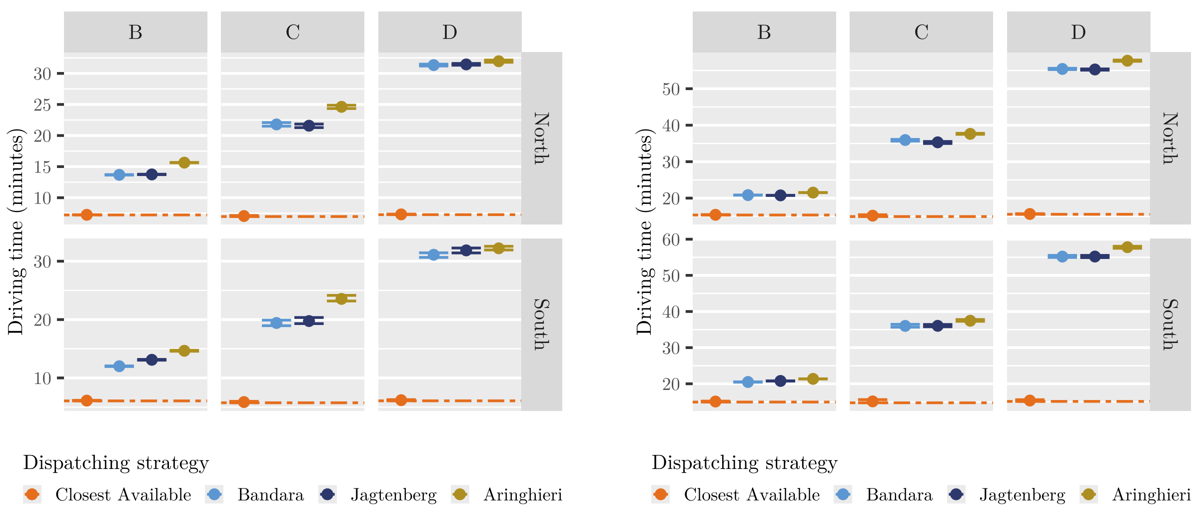
(a) Median (b) 5th percentile

Figure 8.8: Utilities with 95% confidence intervals of simulation experiments for 2023 for patient categories 3–6. Different y-axis scales are used to improve legibility.

The utilities decrease significantly for all indicators, in both regions, and for all dispatching strategies compared to the *Closest Available* strategy. The utilities decrease more strongly with increasing driving time thresholds. The strategies appear to exploit the allowed driving time thresholds for non-prioritized calls to reduce driving times for prioritized calls. Decision-makers, therefore, need to decide on a trade-off between improving utilities for prioritized calls by reducing utilities for non-prioritized calls when implementing one of the compared dispatching strategies.
The response time thresholds are derived from the utility functions based on a utility threshold of 0.5. The median utilities stay above 0.5 for all categories in both regions, but the 5th percentiles are below 0.5 except for category 5 in the region South. Utilities are functions of either response or prehospital times. The driving time thresholds are derived from response time thresholds based on percentiles of the historic distributions of the non-driving time components of the response times. The 5th percentile of utilities dropping below 0.5 indicates that a more conservative percentile might be beneficial.
The variance in utilities between categories is lower for the *Bandara*, *Jagtenberg*, and *Aringhieri* strategies compared to the *Closest Available* strategy. The response time targets are derived from the utility functions with equal utility thresholds. This appears to be leading to driving times that result in more similar utilities. Currently, the interpretation of utility values across categories is not comparable. If, however, utility functions could be designed based on interpretable and comparable outcome data, closer utility values would indicate a more equitable quality across the patient categories.

Analyzing Driving Times

Fig. 8.9 displays the median and 95th percentile of driving times for the dispatching categories B–D. The numerical results can be found in Tables B.11 and B.12.



(a) Median

(b) 95th percentile

Figure 8.9: Driving times with 95% confidence intervals of simulation experiments for 2023 for dispatching categories B–D. Different y-axis scales are used to improve legibility.

As expected, the driving times increase significantly in all indicators, and the increase is larger for categories with a larger response time threshold. The 95th percentiles of driving times are close to the response time thresholds for each category. The dispatching strategies, therefore, appear to fully exploit the larger response time thresholds in order to reduce driving times to prioritized calls. The driving times to non-prioritized calls are considered only in the form of response time thresholds by the strategies, which implies that the response time thresholds should be set with consideration.

The *Aringhieri* strategy leads to significantly longer driving times compared to the strategies *Bandara* and *Jagtenberg*. The only exception is the median for dispatching category D in the region South where there is no significant difference between the strategies *Jagtenberg* and *Aringhieri*. The strategy *Aringhieri* therefore leads to the largest increase in driving times for non-prioritized calls, and at the same time does not lead to the largest decrease in driving times for prioritized calls in our experiments.

8.3.2.3 Influence of Misclassifications

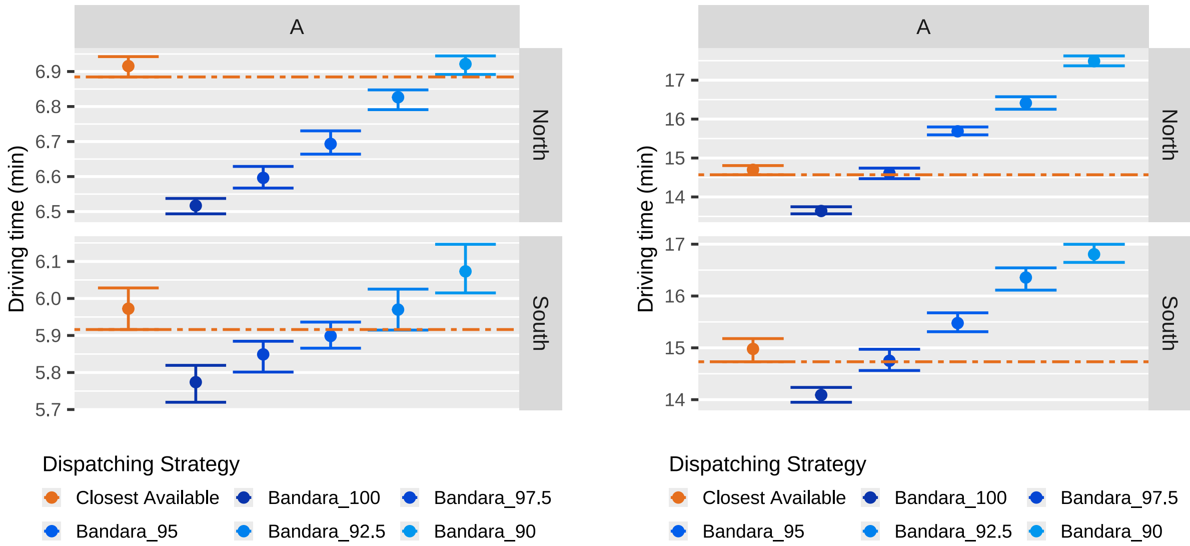
The results presented so far imply the assumption that the dispatching category of a call can be detected with perfect accuracy during the dispatching decision. This allows evaluating the results based on the same call categorization that is used for taking the dispatching decision. The assumption is in line with the literature on dispatching strategies, but is unlikely to be valid in practice. The dispatcher can base the call category assumption only on the information supplied during the emergency call, which might not necessarily be complete or accurate. Furthermore, the condition of patients can change over time, making it questionable whether the retrospective utilities should be calculated based on the category that was assumed at the time of dispatching.

Two types of misclassifications can occur. A call can be categorized as more urgent than it actually is, for instance, if a call is classified as dispatching category A, while it actually is of dispatching category C. This can lead to dispatching a closer ambulance than necessary, which can lead to longer driving times for actual dispatching category A calls. The more often this type of misclassification occurs, the smaller the reduction in driving times to prioritized calls will become. In the extreme case, every call is categorized as the most urgent category, which would equate to the *Closest Available* strategy. Investigating this type of error would require additional simulation experiments with varying proportions of prioritized calls, which is outside the scope of this thesis.

The second type of misclassification occurs if a call is categorized to be less urgent than it actually is, for instance, if a call is classified as dispatching category C, while it actually is of dispatching category A. This can lead to dispatching an ambulance with longer driving times than the closest available ambulance. If this type of misclassification occurs too often, the driving times to prioritized calls could become longer than when dispatching according to the *Closest Available* strategy. This type of misclassification is closely related to the concept of recall in classification problems in the area of machine learning, and we refer to it as categorization recall.

In the following, we assess how sensitive the *Bandara* strategy is to the categorization recall. We chose the *Bandara* strategy as an exemplary strategy, as it dominated the compared dispatching strategies in reducing driving times to prioritized calls. To assess the sensitivity, we randomly change the category of calls with dispatching categories B–D to category A. We change varying numbers of calls to quantify how actual dispatching category A driving times are impacted if only a proportion X of dispatching category A calls have been classified as such at the time of dispatch.

Fig. 8.10 displays the median and 95th percentile of simulated driving times for the regions North and South based on simulating calls in 2023. The proportion of correctly identified dispatching category A calls is reduced in steps of 2.5 percentage points. For instance, *Bandara_95* refers to the results where 95% of actual dispatching category A calls have been identified as dispatching category A calls at the time of dispatch, while the remaining 5% have been classified as one of the categories B–D.



(a) Median (b) 95th percentile

Figure 8.10: Driving times with 95% confidence intervals of simulation experiments for 2023 for decreasing categorization accuracies. Different y-axis scales are used to improve legibility.

The driving times to actual dispatching category A calls increase with decreasing proportions of correctly identified dispatching category A calls in all indicators. The relative increase in the median between a 100% and a 90% categorization recall is 6.1% in North, and 5.2% in South. The relative increase in the 95th percentile is 28.2% in the region North, and 19.3% in the region South. Decreasing the categorization recall leads to not dispatching the closest available ambulance to dispatching category A calls with increasing frequency, which increases the driving times to actual dispatching category A calls. The longer the response time thresholds for non-prioritized calls, the longer the potential driving times to misclassified dispatching category A calls become. While longer response time thresholds for non-prioritized calls can potentially reduce driving times for prioritized calls, assuming a 100% categorization recall, they likely contribute to increasing the driving times for prioritized calls if the categorization recall is below 100%.

The median of the driving times does not significantly increase compared to the *Closest Available* strategy in both regions for any of the evaluated categorization recalls, but the 95th percentile increases significantly for a categorization recall of 95% and below. The 95th percentile is impacted more strongly as the misclassified dispatching category A calls tend to have longer driving times than the correctly classified calls, which impacts the tail of the distribution of driving times more than the central tendency. The sensitivity of the impact on actual dispatching category A patients to the dispatching recall indicates that information about the categorization recall is crucial when implementing a dispatching strategy other than the *Closest Available* strategy.

8.4 Comparing Dispatching Strategies: Discussion

This chapter aimed to investigate how the utility of patients can be influenced through the dispatching decision. Of particular interest was whether the utilities of prioritized calls can be improved by not always dispatching the closest available ambulance to non-prioritized calls. To this end, we compared several dispatching strategies using simulation.

Contributions

We identified eight dispatching strategies suitable for the EMS in Baden-Württemberg and patient categories and utility functions developed in Chapter 5. The selection of strategies is based on attributes that we extracted in the scoping review of dispatching strategies in Chapter 7.

The dispatching strategies are formally defined with differing notations in their publications. We harmonized the notations and adapted several strategies to facilitate comparing the strategies based on patient categories and utility functions.

The simulation model developed in Chapter 6 serves to compare the dispatching strategies for real-world instances. We implemented the eight dispatching strategies in the simulation model, derived the necessary parameters for the strategies from historical data, and simulated the dispatching strategies for several real-world instances in Baden-Württemberg.

The parameters of the dispatching strategies include the response time targets for non-prioritized calls. Response time targets for prioritized calls have been based on medical evidence concerning survival probabilities of patients with an out-of-hospital sudden cardiac arrest. For non-prioritized calls, however, we could not find any medically grounded approach to set response time targets. The utility functions developed in Chapter 5 are derived from medical expertise, and we use the utility functions to derive response time targets for non-prioritized calls. This constitutes, to the best of our knowledge, the first approach of deriving response time targets for non-prioritized calls directly from medical estimations, even though the utility functions are not evidence-based.

Comparing eight dispatching strategies on the same set of instances constitutes, to the best of our knowledge, the most extensive comparison of dispatching strategies. The comparison resulted in identifying two strategies, proposed by Bandara et al. (2014) and Jagtenberg, Bhulai, and van der Mei (2017) respectively, that significantly reduced driving times to prioritized calls in both the median and 95th percentile in the three largest out of the four simulated instances.

The sensitivity to misclassifying patients during the dispatching decision has not been analyzed for any of the compared strategies in their publications. We investigated the sensitivity to misclassifications for the most effective strategy in our experiments and found the reduction in driving times to be highly sensitive to correctly identifying prioritized calls, especially in the 95th percentile.

Practical Implications

Our findings provide multiple practical implications for decision makers. The impact of deviating from the closest available dispatching strategy differed significantly between the investigated instances. The potential improvements of not dispatching the closest available ambulance, therefore, seem to depend on the characteristics of the region.

We found significant increases in utilities and significant decreases in driving times to prioritized calls when not dispatching the closest available ambulance. The magnitude of the improvements, however, was not found to be substantial, and the dispatching decision in isolation appears not to be sufficient to substantially improve EMS operations. At the same time, the dispatching strategies significantly increased driving times to non-prioritized calls in a much larger magnitude, and the trade-off between reducing driving times to prioritized calls and increasing driving times to non-prioritized calls should be taken into consideration.

The sensitivity of the results to correctly classifying patients implies that information about how well dispatchers can distinguish between categories is crucial when deviating from the closest available dispatching strategy.

Limitations

Several limitations should be considered when interpreting our results. We compared the dispatching strategies only on instances within Baden-Württemberg, and it is therefore unclear how well the findings translate to different EMS systems.

The initial comparison of strategies is based only on two singular EMS regions in Baden-Württemberg. Comparing all strategies in additional, and larger regions could have strengthened the results.

The parameters for the strategies are calculated based on the same calls used in the simulation experiments. This chapter aimed to investigate the potential of improving utilities through the dispatching decision, but our findings likely overestimate the impact compared to simulating the dispatching strategies for unseen call data.
We did not perform sensitivity analyses beyond investigating the sensitivity of the *Bandara* strategy to the classification recall. Our results can therefore only be interpreted in the context of our parameter choices, such as demand node definitions or response time targets. Lastly, the *Lee* strategy contains a calibration parameter to control the weight of the driving time to the call at dispatch compared to the impact of the dispatching decision on the preparedness of the EMS system for future calls. We set the value of this parameter in accordance with a previous implementation of the strategy in the literature. Performing a sensitivity analysis might have yielded different insights regarding the *Lee* strategy.

Methodological Reflections

We made several methodological choices in this chapter and reflect on them in the following. We limit the strategies for our comparison to online approaches. The *Bandara* strategy, proposed by Bandara et al. (2014), is an offline approach, but it is computationally feasible to run the heuristic online. The strategy proved to be the most effective strategy in our experiments, and systematically investigating whether further offline approaches could be employed in an online setting might have led to including additional strategies.
The *Roa* and *Amorim* strategies explicitly consider the impact of the dispatching decision on all dispatching categories and not only on prioritized calls. Neither strategy significantly reduced driving times to prioritized calls, and the *Roa* strategy significantly increased driving times to prioritized calls in one indicator. Adapting the strategies to consider only the impacts on prioritized calls might have improved their effects on prioritized calls.
Defining the locations of demand nodes is a central choice in implementing dispatching strategies, as it influences several parameters such as driving times and demand node weights. We use k-means clustering of historic call locations and use the Euclidean distance as a distance indicator in the algorithm. Another option would have been to use distances or travel times along the road network as a distance indicator in the k-means clustering algorithm. Next to tuning the k-means clustering algorithm, other clustering algorithms such as hierarchical clustering have been proposed (Olivos & Caceres, 2022). A comparison of approaches to defining demand node locations could therefore have been valuable.
We set the parameters of busy fractions and service rates based on simulating the *Closest Available* strategy. Our results showed that the driving times to non-prioritized calls increase significantly when not always dispatching the closest available ambulance to non-prioritized calls. This increases the service times and likely increases the utilization of ambulances. The fundamental issue is that the dispatching decision influences parameters that serve as an input to taking the dispatching decision. One approach to handling this interdependency would have been to conduct repeated simulation runs and update the

parameters after each run until some convergence criterion is met. Another possibility would be to initialize the parameters with some value and update them continuously, for instance, by recalculating them for each dispatching decision based on a defined time period preceding the dispatching decision.

The initial comparison of dispatching strategies is based on simulating two EMS regions. The extended analysis of effective strategies is based on simulating two larger study regions that consist of multiple EMS regions. The results indicate that the effect of deviating from the *Closest Available* strategy increases when employing them in larger regions. Comparing all strategies by simulating larger regions could have strengthened the initial comparison. We employed the variance reduction technique of common random numbers and conducted a single iteration of each simulation run. Conducting several repetitions with varying streams of random numbers, based, for instance, on empirical distribution functions of process times, could have allowed for a more extensive statistical analysis of the simulation results and further strengthened the results.

An interesting extension would be a sensitivity analysis for several parameters. Varying the call rates would allow investigating whether the effects of deviating from the *Closest Available* depend on the utilization of ambulances. We derived the response time thresholds for non-prioritized calls from the utility functions by setting a utility threshold of 0.5. This led to significantly reduced driving times for prioritized calls, but increased the driving times to non-prioritized calls significantly. Varying the response time thresholds for non-prioritized calls could shed light on the correlation between them and the reduction of driving times to prioritized calls. Reducing the driving times to non-prioritized calls would additionally likely reduce the sensitivity of the results to misclassifications.

Our analysis of simulation results focuses on the impact of dispatching strategies on driving times and utilities, especially for prioritized calls. Simulation, however, would allow further interesting analyses. The simulation could easily be extended to capture how often no ambulance is available at the closest station to a prioritized call under the *Closest Available* strategy. This could support in identifying either spatial or temporal patterns for when it might be especially beneficial to deviate from the *Closest Available* strategy. The analysis of the results, furthermore, could have been extended by analyzing measures of variance to investigate the impact of the dispatching decision on the quality dimension of fairness identified in Chapter 4. Lastly, we evaluated two percentiles of driving times and utilities, and calculated confidence intervals to assess the statistical significance of differences. Statistical tests, t-tests for instance, designed to test the hypothesis that samples differ in a specific indicator, could have enhanced the interpretation of the simulation results.

Outlook

We identify several avenues for further research based on comparing dispatching strategies. We see a promising opportunity in using short-term forecasting models to generate parameter values for dispatching strategies. All dispatching strategies aim to quantify the impact of dispatching an ambulance on the future performance of the system. The ambulance is not available for the duration of servicing the call at dispatch. Predicting parameter values such as call rates per demand node for this short-term horizon could therefore improve the effectiveness of the dispatching strategies. A recent review of approaches for call predictions can be found in Mukhopadhyay et al. (2022).

The compared dispatching strategies consider only those ambulances in their calculations that are available at the time of dispatch and omit ambulances that are currently servicing a call. Ambulances that become available shortly after the dispatching decision is taken are therefore not considered in the dispatching strategies. Similarly, the expected service time of the dispatched ambulance is not considered, which implies the assumption that the available ambulances do not differ in their service times. This assumption is not valid due to differences in driving times from ambulance locations to call locations and from hospitals to ambulance stations. Incorporating this temporal dimension could further enhance the dispatching strategies.

We could not find any consideration of how well calls can be classified in the dispatching decision in existing literature. It is reasonable to assume that misclassifications will occur in practice due to, for instance, language barriers, the high-stress situation of the caller, or changes in the patient's condition between the call and the arrival of EMS resources at the scene. If information regarding the classification quality in an EMS system is available, it could be investigated how this information can be considered in the dispatching decision.

The two strategies *Bandara* and *Jagtenberg* proved to be the most effective in our experiments. Neither strategy considers the driving times to non-prioritized calls beyond respecting response time targets. We can therefore not control the balance of the marginal improvement for prioritized calls and the marginal detriment for non-prioritized calls. Investigating whether the driving times to non-prioritized calls can be reduced by considering them in the dispatching decision, without significantly deteriorating the improvements for prioritized calls, could be a valuable extension of the strategies.

We see an area for future research in better understanding the bounds of improvement of the dispatching decision. One approach would be to calculate retrospective optimal solutions. A similar approach has been proposed by Jagtenberg, van den Berg, and van der Mei (2017), and could serve as a starting point for further investigations.

Lastly, we found that the impact of deviating from the *Closest Available* strategy depends on the instance. It is, however, unclear which characteristics of an EMS system increase or decrease the impact of the dispatching decision. Studying the dependency of the impact of the dispatching decision on EMS system characteristics — either based on results of simulating online approaches or on performance bounds calculated through retrospective optimal solutions — could inform decision makers whether the dispatching decision holds potential to improve the quality of their EMS system.

8.5 Comparing Dispatching Strategies: Conclusion

In this chapter, we investigated the potential to impact patient utilities by deviating from the standard *Closest Available* strategy for non-prioritized calls. We addressed the question of which dispatching strategies are suitable for this purpose in the context of the EMS in Baden-Württemberg, and how these strategies influence patient utilities and driving times. Based on a scoping review, we selected and partly adapted eight dispatching strategies. Our simulation experiments demonstrated that deviating from the *Closest Available* strategy for non-prioritized calls can significantly reduce driving times to prioritized calls. Specifically, the strategies proposed by Bandara et al. (2014) and Jagtenberg, Bhulai, and van der Mei (2017) proved to be the most effective in our study regions. These strategies successfully reserve capacity for future prioritized calls by not necessarily dispatching the closest available ambulance to non-prioritized calls, thereby increasing the likelihood that a closer ambulance is available for prioritized calls.

However, our results also highlight the inherent trade-off of this approach. While driving times to prioritized calls could be reduced significantly, driving times to non-prioritized calls increased significantly and in a larger magnitude. While we constrained this increase through response time targets derived from utility functions, the sensitivity of the results to classification errors remains a critical factor. We showed that incorrectly classifying prioritized calls as non-prioritized and consequently assigning a non-closest ambulance can increase driving times to prioritized calls compared to the *Closest Available* strategy.

In conclusion, the dispatching decision showed the potential to significantly improve utilities for critical patients in the EMS in Baden-Württemberg. Implementing dispatching strategies other than the *Closest Available* strategy, however, requires information about how well prioritized calls can be identified and should balance the trade-off between reducing driving times to prioritized calls and increasing driving times to non-prioritized calls. The potential impact of the dispatching strategy additionally appears to depend on the characteristics of the study region.

Part V

Locating and Sizing Ambulance Stations for Heterogeneous Patients

Summary of Part V

Research Question and Objectives

Research question: How can the ambulance station location and sizing decision consider heterogeneous patients?
We derive the following *research objectives* after an initial review of existing modeling approaches:

- Develop a base model that relaxes the full and uniform demand assignment assumptions present in existing models.
- Compare the base model to existing modeling approaches to validate it.
- Extend the base model to consider location-dependent service times.
- Extend the base model to consider multiple patient types with differentiated coverage thresholds and service levels.

Scope and Methodology

We focus on ambulances and their stations as the most used resource in EMS. The decisions of interest are the locations and the sizes of ambulance stations as depicted in Fig. 8.11.

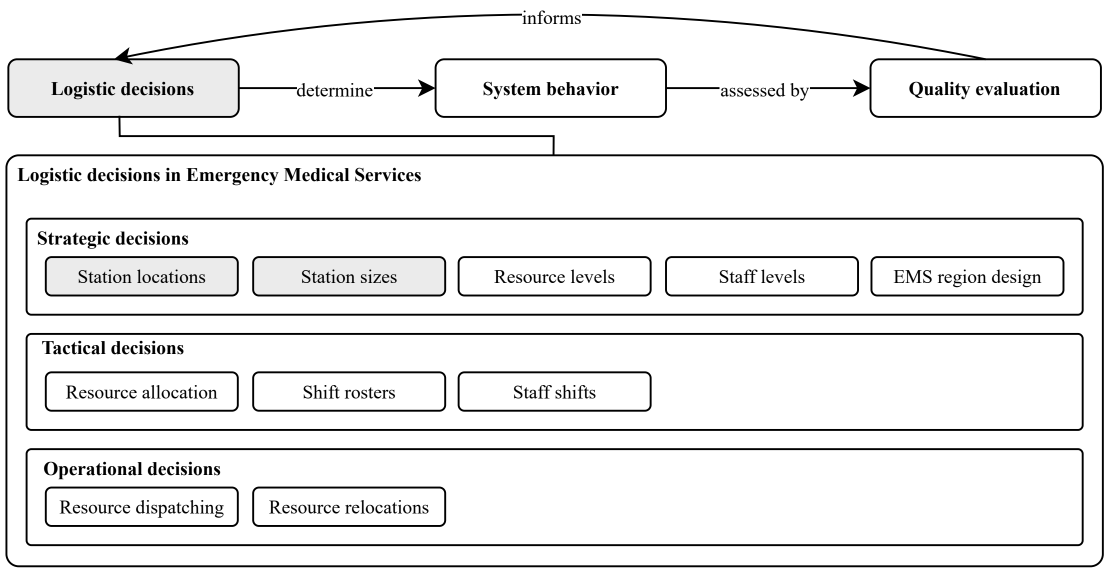


Figure 8.11: Part V investigates the ambulance station location and sizing decision for heterogeneous patients.

Mixed-integer linear programming is employed to formulate a base model and two extensions for the integrated ambulance station location and sizing decision. Fourteen instances are defined based on real-world data from EMS in Baden-Württemberg. Three existing models and the models developed in this part solve the instances in computational experiments. The *simulation* model developed in Part III is used to evaluate the model solutions.

Main Contributions

- Formulation of a capacitated set covering model with system-optimal demand assignment for the ambulance station location and sizing decision
- Comparison of three existing models and the newly developed model using simulation with real-world data
- Extension of the base model to consider location-dependent service times
- Extension of the base model to consider heterogeneous patients with differentiated coverage thresholds and service levels

Key Findings

- The system-optimal demand assignment assumption reduces the number of prescribed resources compared to the full demand assignment assumption.
- The independence assumption in calculating ambulance availabilities leads to fewer prescribed ambulances in our instances compared to the dependence assumption.
- Modeling availabilities based on a station-specific perspective can avoid overcapacities induced by modeling availabilities based on an ambulance-specific perspective.
- Considering differentiated response time thresholds of heterogeneous patients can reduce the number of prescribed resources compared to a single coverage target.
- The effects of considering prehospital time based coverage targets depend on the instance characteristics.

Chapter 9

Developing a Base Model: The QRUBUL

Strategic decisions in EMS determine the fundamental structure and capabilities of the system long-term. They usually have significant financial implications and influence the degrees of freedom of tactical and operational decisions. Strategic decisions, therefore, have a special importance as they determine the bounds of the system's performance. This chapter and the following chapter Chapter 10 focus on the strategic decisions of locating and sizing ambulance stations.

The station location and sizing decisions have three possibly conflicting objectives and one central constraint, depicted in Fig. 9.1. Minimizing the amount of resources necessary to comply with legal regulations is the primary objective of decision makers in practice. The objective of maximizing system performance with a given number of resources while still fulfilling legal regulations is the secondary objective. The vast majority of existing modeling approaches focus on either minimizing resources or maximizing system performance. This chapter focuses on considering heterogeneous patients, and we build upon models that minimize resources to align with the primary objective of decision makers.

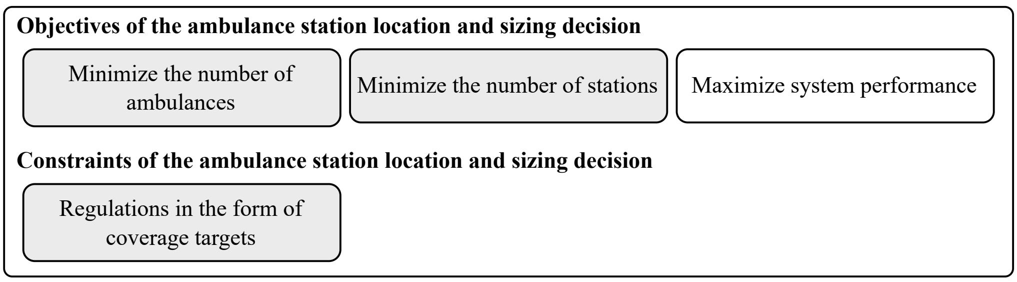


Figure 9.1: Central objectives and constraint of the ambulance station location and sizing decision.

Regulations in EMS are typically based on response time targets requiring a certain percentage of calls to be reached within a response time threshold. We came to the conclusion in Part II that heterogeneous patients exist in EMS, which should be reflected by differentiated planning criteria. The current regulations in Baden-Württemberg contain differentiated targets. At the time of writing, one patient category with a response time target of 12 minutes in 95% and a second patient category with a prehospital time target of 60 minutes in 80% of cases are defined. The regulations additionally include a provision for further patient categories with separate targets that can be defined in the future.

The question, therefore, is: How can the ambulance station location and sizing decision consider heterogeneous patients?

The station location and sizing decisions are usually modeled as one integrated decision, which is referred to as the ambulance location problem. A graph representation of the planning region is the basis for modeling the ambulance location problem. Supply nodes in the graph represent possible station locations, and demand nodes represent call locations. The weights of the edges between supply and demand nodes contain information about driving times, and demand node weights represent call frequencies. A demand node is considered *covered* by a supply node if the edge weight between the nodes is less than or equal to a threshold. This threshold can, for instance, be derived from the response time target in a regulation. Similarly, the *neighborhood* of a node n in the graph refers to the set of nodes that can be reached from n within a given threshold.

The problem of finding the locations of the minimum number of stations necessary to cover all demand nodes is called the Location Set Covering Problem and was introduced by Toregas et al. (1971). The uncertainty in the availability of ambulances needs to be considered to integrate the decision on the size of the stations. An assumption is made on how the demand of demand nodes is assigned to covering stations. This allows calculating the necessary number of ambulances stationed at each station to ensure a certain availability of having an ambulance available when a call arrives.

Our fundamental approach in the following is guided by examining which assumptions of existing models become less valid by considering heterogeneous patients. One assumption in existing models is that demand is fully assigned to each covering station. If two stations cover the same demand node, its demand is considered twice. An alternative assumption is that demand is uniformly assigned to each covering station in relation to the number of ambulances at each station. While this reduces the risk of overcapacities, it can still lead to sizing some stations larger than necessary. Fig. 9.2 visualizes how both assumptions can lead to overcapacities compared to a system-optimal demand assignment if multiple stations cover the same demand node. The simplified example assumes three demand nodes with a demand of 25, 55, and 40, respectively. Two stations, A and B, are considered. Station A covers demand nodes 1 and 2, and station B covers demand nodes 2 and 3. We assume that each ambulance can serve a demand of up to 40. If the demand of node 2 is fully assigned to both stations A and B, 5 ambulances are required. If the demand of node 2 is uniformly assigned, 4 ambulances are required, and if the demand is assigned system-optimally, only 3 ambulances are sufficient.

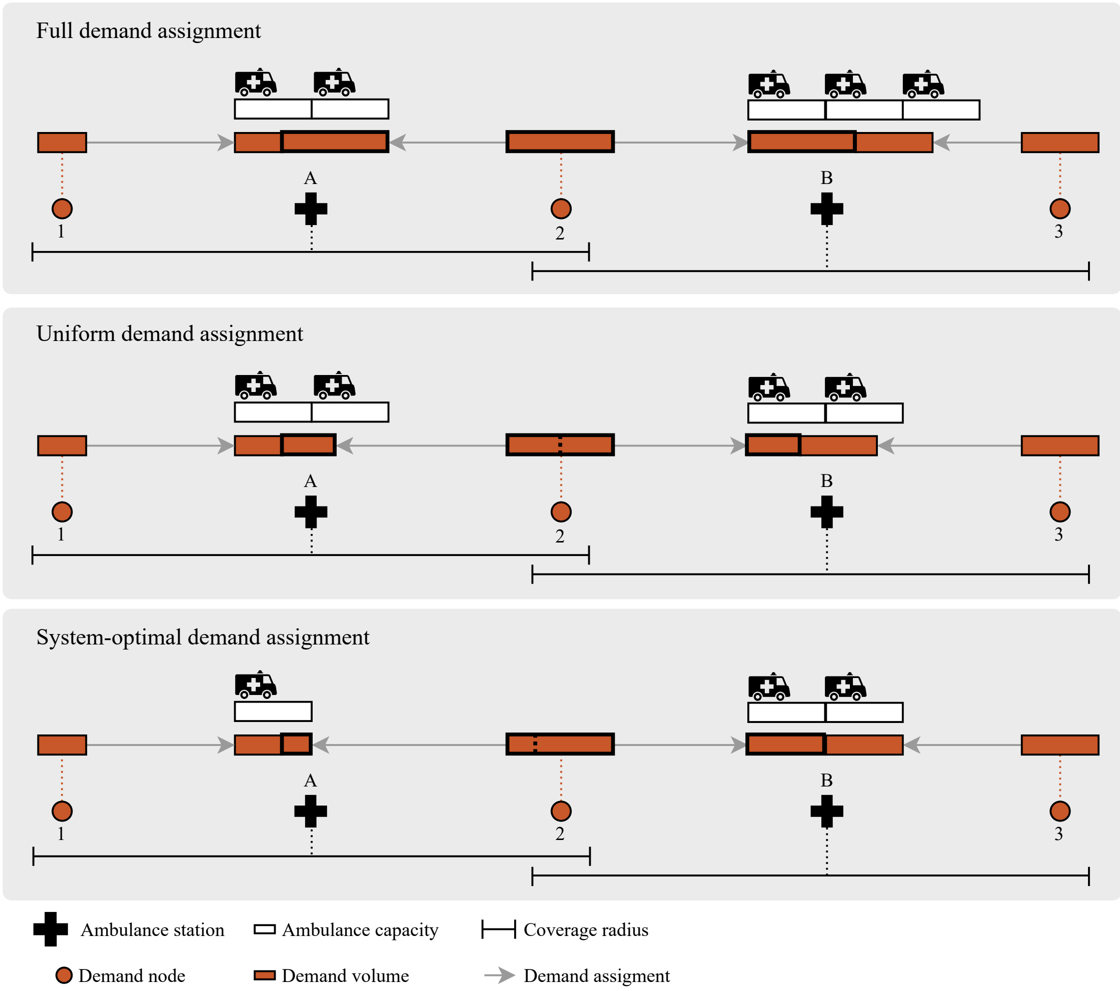


Figure 9.2: The assumptions of full or uniform demand assignment can lead to overcapacities. The effect is expected to increase with larger and differentiated coverage thresholds.

Multiple coverage will increase with differentiated coverage thresholds, increasing the potential overcapacities induced by the full or uniform assignment assumption. We propose the QRUBUL based on a system-optimal demand assignment to address this issue. We adopt the system-optimal wording from the Capacitated Location Set Covering Problem - System Optimal (CLSCP-SO) introduced by Current and Storbeck (1988). The CLSCP-SO is formulated as a general model outside the EMS context, and the QRUBUL extends it by making the capacity of a supply node dependent on the decision on the size of the supply node. We formulate the QRUBUL for a single coverage threshold to compare it to existing modeling approaches and to understand the effects of relaxing demand assignment assumptions. Chapter 10 extends the QRUBUL to integrate differentiated planning targets. The remainder of this chapter is structured as follows: Section 9.1 reviews related literature on the ambulance location problem. Section 9.1.2 presents existing models designed to minimize resources. The QRUBUL model is formulated in Section 9.2. Computational experiments based on real-world data from Baden-Württemberg are conducted in Section 9.3 to compare the QRUBUL to previous models. The results are discussed in Section 9.4 and Section 9.5 concludes the chapter.

9.1 Related Literature on the Ambulance Location Problem

Several literature reviews survey the extensively studied ambulance location problem (Aringhieri et al., 2017; Bélanger et al., 2019; Brotcorne et al., 2003; X. Li et al., 2011; Neira-Rodado et al., 2022; W. Wang et al., 2021). The ambulance location problem determines station and ambulance locations on a strategic level. Dynamic ambulance location or ambulance relocation addresses ambulance movement between stations on an operational level. We focus on models for the strategic variant of the ambulance location problem in the following. Section 9.1.1 presents an overview of model classes in the ambulance location problem. The class of set covering models considering uncertainty in ambulance availability is the most relevant for our work, and Section 9.1.2 presents models of this class to motivate our modeling choices.

9.1.1 An Overview of Modeling Approaches

Ambulance location models can be divided into different classes. Fundamentally, two objectives are at the heart of the ambulance location problem: Minimizing resources and maximizing performance. The earliest model minimizing resources is the Location Set Covering Problem introduced by Toregas et al. (1971). It minimizes the number of required stations to cover all demand nodes. Most models consider stations and one type of ambulance as resources, but there are approaches that model different ambulance types, such as McLay (2009). Church and ReVelle (1974) introduced the earliest model maximizing performance with the Maximal Covering Location Problem, which maximizes the number of covered demand nodes with a given number of stations. Next to coverage as a performance criterion, the average response time has been minimized by Dzator and Dzator (2013), which equates to the classic p-median problem. Erkut et al. (2008) were the first to explicitly consider the patient perspective by maximizing survival based on survival functions. Knight et al. (2012) introduced a model that considers heterogeneous patients with differentiated performance criteria combining survival functions and coverage. While maximizing coverage, minimizing average response times, and maximizing average survival represent an efficiency perspective on EMS, equity-based performance criteria have gained increasing attention in recent years (Chanta et al., 2011; Jagtenberg & Mason, 2020a; Jagtenberg & Mason, 2020b; Toro-Díaz et al., 2015). Ambulance location models almost exclusively focus on response times or functions of response times, but H. Andersson et al. (2020) also explicitly consider prehospital times in their model.

The initial deterministic models have evolved to incorporate uncertainties present in real-world systems. The uncertainty in the availability of ambulances has been integrated by either enforcing an arbitrary multiple coverage, such as the Double Standard Model by Gendreau et al. (1997), or by explicitly calculating the required number of ambulances for each demand node. These availability calculations have been based either on busy fractions, for instance, by Daskin (1983) in the Maximum Expected Covering Location Problem, or on queuing analysis, for instance, by Marianov and ReVelle (1994). Further modeled uncertainties concern driving times (Ingolfsson et al., 2008) and the demand (Beraldi & Bruni, 2009).

More recently, Bélanger et al. (2020) and Toro-Díaz et al. (2013) modeled the location decision integrated with the operational dispatching decision.

Location-independent service times are a common assumption in ambulance location models. Leknes et al. (2017) relax this assumption and consider location-dependent service times, which take hospital locations into account for the calculation of ambulance availabilities. Heterogeneity of sub-regions within a planning region can also take the form of varying call rates, which van Buuren et al. (2018) consider in a post-processing approach. A further modeling assumption present in ambulance location models is the independence of ambulance stations. Grot et al. (2022) relax this assumption in an extension of the Maximum Expected Covering Location Problem.

9.1.2 Set Covering Models Considering Uncertainty in Ambulance Availability

The focus of this chapter is on identifying the required number of stations and ambulances to fulfill a given performance target. Set covering models considering the uncertainty of ambulance availability are a suitable model class for this purpose. This section presents the evolution of existing models to motivate our modeling choices and concludes with our contribution to existing models. We present six models in the following. The last three models will be used in computational experiments in Section 9.3. The first three models introduce modeling concepts on which the latter models build. We follow Borrás and Pastor (2002) in naming the models, which partly deviates from how the models were named in the original publications. Section 9.1.2.1 presents common assumptions of the models and introduces the notation that we use to describe the models subsequently.

9.1.2.1 Assumptions and Notation

The following assumptions underlie all the following models:

Location-independent service times: The time for serving a call is independent of the locations of the demand and supply nodes.

Dispatch from the station: Ambulances are assumed to be dispatched from their station and to return to their station after serving a call.

Strict spatial coverage: EMS regulations prescribe a percentage of calls that need to be reached within a coverage threshold. This percentage is calculated by aggregating the spatial and the temporal dimensions. The following models assume that every demand node needs to be covered by a station, enforcing a strict spatial availability, and use the target percentage to calculate the temporal availability of ambulances.

Deterministic parameters: The travel times, service times, and call rates are assumed to be known and deterministic.

The notations of the original publications have partly been adapted to facilitate comparing the model formulations. We use the following notation for the model formulations:

Sets

- J: Set of supply nodes representing potential station locations
- I: Set of demand demand nodes representing call locations
- A: Set of potential ambulances
- A_j: Set of ambulances stationed at supply node $j \in J$
- N_i: Set of supply nodes covering demand node $i \in I$
- M_n: Set of demand nodes in the neighborhood of $n \in I \cup J$
- O_i: Set of ambulances in the neighborhood of $i \in I$

Variables

- y_j: Binary variable indicating if a station is located at potential station location j
- x_j: Number of ambulances allocated to supply node $j \in J$
- x_{jk}: Binary variable indicating if k ambulances are allocated to supply node $j \in J$
- $\varkappa_a$: Binary variable indicating if ambulance a is deployed
- c_{ik}: Binary variable indicating if demand node $i \in I$ is covered k times
- p_{ij}: Proportion of demand at demand node $i \in I$ assigned to supply node $j \in J$

Parameters

- α: Service level
- f_i: Daily call rate at demand node $i \in I$
- $\bar{t}$: Mean service time per call response
- $\hat{t}$: Upper bound for the service time per call response
- $\dot{t}$: Daily service time in hours
- L: Maximum number of ambulances that can be allocated to a station
- L^{Neigh}: Maximum number of ambulances in the neighborhood of any demand point
- β: Weight of the objective of minimizing the number of ambulances
- γ: Weight of the objective of minimizing the number of stations
- T: Threshold for coverage
- M: A sufficiently large constant

9.1.2.2 The Binomial Probabilistic Location Set Covering Problem

Revelle and Hogan (1989) formulated the Binomial Probabilistic Location Set Covering Problem (BPLSCP). The BPLSCP decides on the number x_j of ambulances stationed at each supply node $j \in J$ and minimizes the number of required ambulances (9.1.1a). Constraint 9.1.1b ensures that a sufficient number b_i of ambulances is stationed within the neighborhood of each demand node $i \in I$ to guarantee a service level α of having an ambulance available when a call arrives.

The required number of ambulances is calculated based on the idea of a busy fraction. A demand zone is defined around each demand node $i \in I$ and contains all demand nodes in the neighborhood M_i of i. For each zone, the necessary hours per day to serve all occurring calls is calculated with $\bar{t}\sum_{k\in M_i} f_k$ and divided by the 24 hours within one day (Eq. (9.1.3)). This represents the busy fraction ρ_i of one ambulance serving the demand in the zone surrounding i. The busy fraction per ambulance if b_i ambulances would serve the demand is therefore $\frac{\rho_i}{b_i}$ under the independence assumption. This enables calculating the necessary number of ambulances with which the probability of having at least one ambulance available is greater than the desired service level α by increasing b_i until Eq. (9.1.2) is satisfied.

The BPLSCP is formulated as follows:

$$\min_{x} \quad \sum_{j\in J} x_j \tag{9.1.1a}$$

$$\text{s.t.} \quad \sum_{j\in N_i} x_j \geq b_i, \quad \forall i \in I, \tag{9.1.1b}$$

$$x_j \geq 0, \quad \forall j \in J, \tag{9.1.1c}$$

$$x_j \in \mathbb{Z}, \quad \forall j \in J \tag{9.1.1d}$$

where b_i is the smallest integer satisfying:

$$1 - \left(\frac{\rho_i}{b_i}\right)^{b_i} \geq \alpha \tag{9.1.2}$$

with:

$$\rho_i = \frac{\bar{t}\sum_{k\in M_i} f_k}{24} \tag{9.1.3}$$

In addition to the common assumptions listed in Section 9.1.2.1, the BPLSCP assumes the independence of ambulances, leading to an availability calculation based on a busy fraction. It additionally views the system from a demand zone perspective and assumes that arriving calls are uniformly assigned to covering ambulances.

9.1.2.3 The Poisson Reliability Location Set Covering Problem

Ball and Lin (1993) introduced the Poisson Reliability Location Set Covering Problem (PRLSCP) which was originally called Reliability-P (Rel-P). It minimizes the number of ambulances (9.1.4a) and models the allocation decision with a binary variable x_{jk} which indicates whether k ambulances are stationed at supply node $j \in J$.

The PRLSCP builds on the assumption that call arrivals follow a Poisson process. The arrival rate is expressed as calls per upper bound service time $\hat{t}$ (s. Eq. (9.1.6)). We can then calculate the probability $P\left(D(j) \geq k\right)$ of k or more calls arriving at one station with Eq. (9.1.5). This is based on the assumption that a station $j \in J$ serves all calls in its neighborhood M_j. This probability $P\left(D(j) \geq k\right)$ is equal to the unavailability probability of a station if k ambulances are allocated to that station.

Constraint 9.1.4b enforces that the joint probability of having at least one ambulance available at one of the stations in the neighborhood N_i of demand node $i \in I$ is greater than the required service level α. The multiplicative calculation of the joint availability probability is linearized using the log function.

Constraint 9.1.4c ensures that exactly one number k or no ambulances are allocated to a supply node $j \in J$.

The PRLSCP model is formulated as follows:

$$\min_{x} \quad \sum_{j \in J} \sum_{k=1}^{L} k x_{jk} \tag{9.1.4a}$$

$$\text{s.t.} \quad \sum_{j \in N_i} \sum_{k=1}^{L} - \log\left(P\left(D(j) \geq k\right)\right) x_{jk} \geq -\log(1-\alpha), \quad \forall i \in I, \tag{9.1.4b}$$

$$\sum_{k=1}^{L} x_{jk} \leq 1, \qquad \forall j \in I, \tag{9.1.4c}$$

$$x_{jk} \in \{0,1\}, \qquad \forall j \in J,\ k = 1, 2, \ldots, L \tag{9.1.4d}$$

where

$$x_{jk} = \begin{cases} 1, & \text{if } k \text{ vehicles are stationed at supply node } j, \\ 0, & \text{otherwise.} \end{cases}$$

with

$$P\left(D(j) \geq k\right) = 1 - \sum_{c=0}^{k-1} e^{-\lambda_j} \frac{1}{c!} \lambda_j^c \tag{9.1.5}$$

and

$$\lambda_j = \frac{\hat{t} \sum_{i \in M_j} f_i}{24} \tag{9.1.6}$$

The PRLSCP builds on the assumption that calls follow a Poisson process. It views the system from a station perspective and assumes that a station will serve the full demand of the demand nodes in its neighborhood.

9.1.2.4 The Queuing Probabilistic Location Set Covering Problem

Marianov and ReVelle (1994) developed the Queuing Probabilistic Location Set Covering Problem (QPLSCP). The model has the same basic structure as the BPLSCP and differs only in how the number of required ambulances b_i covering a demand node $i \in I$ is calculated. The assumption of independence between ambulances underlies the BPLSCP and is relaxed in the QPLSCP. The QPLSCP views each demand zone as an $M/M/k/loss$ queueing system with k servers. This allows calculating the blocking probability using the Erlang-B formula 9.1.8. The required number b_i in the neighborhood of each demand node $i \in I$ is then calculated numerically by increasing b_i until the probability of having at least one ambulance available is at least at the required service level α. The QPLSCP is formulated as follows:

$$\min_{x} \quad \sum_{j \in J} x_j \tag{9.1.7a}$$

$$\text{s.t.} \quad \sum_{j \in N_i} x_j \geq b_i, \quad \forall i \in I, \tag{9.1.7b}$$

$$x_j \geq 0, \quad \forall j \in J, \tag{9.1.7c}$$

$$x_j \in \mathbb{Z}, \quad \forall j \in J \tag{9.1.7d}$$

where b_i is the smallest integer satisfying:

$$\frac{\frac{1}{b_i!}\rho_i^{b_i}}{1 + \rho_i + \frac{1}{2!}\rho_i^2 + \ldots + \frac{1}{b_i!}\rho_i^{b_i}} = \frac{\frac{1}{b_i!}\rho_i^{b_i}}{\sum_{c=0}^{b_i} \frac{1}{c!}\rho_i^c} \leq 1 - \alpha \tag{9.1.8}$$

and

$$\rho_i = \frac{\bar{t} \sum_{k \in M_i} f_k}{24} \tag{9.1.9}$$

The QPLSCP views each demand zone as a queuing system with Poisson distributed call arrivals and does not assume the independence of ambulances. The availability calculation is based on a uniform assignment of calls to covering ambulances.

9.1.2.5 The Queuing Reliability Location Set Covering Problem

The Queuing Reliability Location Set Covering Problem (QRLSCP) is the first of two models introduced by Borrás and Pastor (2002). The model formulation itself is equivalent to the PRLSCP, but the calculation of the availabilities of the supply nodes differs. The calculation of the probabilities $P\left(D(j) \geq k\right)$ of k or more calls arriving at station $j \in J$ adopts the assumption of a lossy system from the QPLSCP. This leads to truncating the Poisson distribution to the cumulative probability of k arriving calls as shown in Eq. (9.1.11). Calculating the unavailability probability in such a way is numerically equivalent to calculating the blocking probability using the Erlang-B formula for k servers. The reasoning for assuming a lossy system is that calls that arrive when no ambulance is available at a station will be served by ambulances from another station and will not wait until an ambulance becomes available again at the station. The QRLSCP is formulated as follows:

$$\min_{x} \quad \sum_{j \in J} \sum_{k=1}^{L} k x_{jk} \tag{9.1.10a}$$

$$\text{s.t.} \quad \sum_{j \in N_i} \sum_{k=1}^{L} -\log\left(P\left(D(j) \geq k\right)\right) x_{jk} \geq -\log(1-\alpha), \quad \forall i \in I, \tag{9.1.10b}$$

$$\sum_{k=1}^{L} x_{jk} \leq 1, \qquad \forall j \in I, \tag{9.1.10c}$$

$$x_{jk} \in \{0,1\}, \qquad \forall j \in J,\ k = 1,2,\ldots,L \tag{9.1.10d}$$

$$P\left(D(j) \geq k\right) = P\left(D(j) = k\right) = \frac{e^{-\lambda_j}\left(\frac{1}{k!}\lambda_j^k\right)}{\sum_{c=0}^{k} e^{-\lambda_j}\left(\frac{1}{c!}\lambda_j^c\right)} \tag{9.1.11}$$

where

$$\lambda_j = \frac{\hat{t}\sum_{i \in M_j} f_i}{24} \tag{9.1.12}$$

The QRLSCP differs from the QPLSCP in that it views each supply node and its neighboring demand nodes as a queuing system, while the QPLSCP views each demand node and its neighboring demand and supply nodes as a queuing system. The assumptions of Poisson-distributed call arrivals and the dependence of ambulance availabilities at one station underlie the queuing system. Each station is assumed to serve the full demand occurring at neighboring demand nodes.

9.1.2.6 The Binomial Reliability Location Set Covering Problem

The Binomial Reliability Location Set Covering Problem (BRLSCP) is the second model introduced by Borrás and Pastor (2002) and again follows the basic structure of the PRLSCP by enforcing a sufficient joint availability of stations covering each demand node $i \in I$. The conceptual difference to the QRLSCP is that the BRLSCP adopts the independence assumption of ambulance availabilities from the BPLSCP. The difference to the BPLSCP is in switching from a demand zone perspective to a station perspective when calculating the unavailability probabilities ρ_{jk} for every supply node $j \in J$ and number of ambulances at the supply node $k \in \{1,\ldots,L\}$.

Eq. (9.1.14) calculates the busy fraction ρ_{jk} per ambulance at a station given k allocated ambulances. A minimum number of k_{0j} allocated ambulances at an opened station needs to be enforced to ensure that $\rho_{jk} \leq 1$ holds. This condition is necessary for the calculation of the joint unavailability probability in constraint 9.1.13b.

The BRLSCP is formulated as follows:

$$\min_{x} \quad \sum_{j \in J} \sum_{k=1}^{L} k x_{jk} \tag{9.1.13a}$$

$$\text{s.t.} \quad \sum_{j \in N_i} \sum_{k=k_{0j}}^{L} -\log\left((\rho_{jk})^k\right) x_{jk} \geq -\log(1-\alpha), \quad \forall i \in I, \tag{9.1.13b}$$

$$\sum_{k=k_{0j}}^{L} x_{jk} \leq 1, \qquad \forall j \in I, \tag{9.1.13c}$$

$$x_{jk} \in \{0,1\}, \qquad \forall j \in J,\ k = 1,2,\ldots,L \tag{9.1.13d}$$

where

$$\rho_{jk} = \frac{\hat{t} \sum_{i \in M_j} f_i}{24k} \tag{9.1.14}$$

and k_{0j} being the smallest k for which $\rho_{jk} \leq 1$, which can be calculated as $k_{0j} = \min\{k \in \{1,2,\ldots,L\} | \rho_{jk} \leq 1\}$.

The BRLSCP bases its availability calculation on the assumption of ambulance independence. It calculates availabilities from the perspective of each station, assuming that each station serves the full demand of neighboring demand nodes.

9.1.2.7 The Upper-Bound Unavailability Location Problem

The reliability models PRLSCP, QRLSCP, and BRLSCP assume that each station serves the full demand of all demand nodes it covers. This can lead to overcapacities if multiple stations cover the same demand node. The Upper-Bound Unavailability Location Model (UBUL) model formulated by Shariat-Mohaymany et al. (2012) relaxes this assumption of full demand assignment and assumes that demand is assigned uniformly to each covering ambulance independently of which station the ambulance is allocated to. The UBUL additionally changes the modeling perspective from a station-specific one to an ambulance-specific one, where the availabilities are calculated per ambulance and not per station.

The ambulance availabilities are based on an exogenously set busy fraction ρ. Constraint 9.1.15b ensures that ambulance utilizations remain below ρ by bounding the demand measured in hours of service per day, which can be assigned to an ambulance.

The busy fraction ρ dictates how many ambulances need to cover each demand node to reach a service level of α for each demand node $i \in I$, formulated as constraint 9.1.15e. Table 9.1 lists the maximum busy fractions which can be set depending on a service level α and a corresponding number of ambulances which need to cover each demand node.

The model needs to be solved repeatedly with different values of b and ρ for a desired service level α to find the values that lead to the lowest objective function value. The objective function is formulated as a weighted sum of the number of ambulances and the number of stations. This extends the previous models, which exclusively focus on minimizing the number of ambulances.

Service level, α **(%)**	**# of required ambulances,** b				
	1	2	3	4	5
75	0.24	0.49	0.62	0.70	0.75
80	0.19	0.44	0.57	0.66	0.71
85	0.14	0.38	0.52	0.61	0.67
90	0.09	0.31	0.45	0.55	0.62
95	0.04	0.21	0.36	0.46	0.54

Table 9.1: Maximum busy fraction values for meeting different reliability levels with specified number of required ambulances (Shariat-Mohaymany et al., 2012)

The UBUL is formulated as follows:

$$\min_{\varkappa, y} \quad \beta \sum_{a \in A} \varkappa_a + \gamma \sum_{j \in J} y_j \tag{9.1.15a}$$

$$\text{s.t.} \qquad 24\rho \geq \sum_{i \in M_a} \left(\dot{t}_i \sum_{k=1}^{L^{Neigh}} c_{ik} \frac{1}{k} \right) + (\varkappa_a - 1)M, \quad \forall a \in A, \tag{9.1.15b}$$

$$\sum_{a \in O_i} \varkappa_a = \sum_{k=1}^{L^{Neigh}} c_{ik} \times k, \quad \forall i \in I, \tag{9.1.15c}$$

$$\sum_{k=1}^{L^{Neigh}} c_{ik} = 1, \qquad \forall i \in I, \tag{9.1.15d}$$

$$\sum_{a \in O_i} \varkappa_a \geq b, \qquad \forall i \in I, \tag{9.1.15e}$$

$$M y_j \geq \sum_{a \in A_j} \varkappa_a, \qquad \forall j \in I, \tag{9.1.15f}$$

$$\varkappa_a, y_j, c_{ik} \in \{0, 1\}, \qquad \forall a \in A,\ \forall j \in I,\ \forall i \in I,\ k = 1, 2, \ldots, L^{Neigh} \tag{9.1.15g}$$

with:

$$L^{Neigh} = \max_{i \in I} \{||N_i||\} \tag{9.1.16}$$

The left-hand side of constraint 9.1.15b gives the maximum daily capacity in service hours, which ensures a busy fraction of at most ρ for each ambulance $a \in A$. The right-hand side counts the service hours of the calls assigned to ambulance a. The daily hours of required service for calls at demand nodes $i \in M_a$ in the set of demand nodes covered by ambulance a are divided by the number of ambulances covering demand node i. Constraints 9.1.15c and 9.1.15d ensure that the number of ambulances covering demand node $i \in I$ is counted correctly. Allocating at least b ambulances in the neighborhood of each demand node $i \in I$ is enforced by constraint 9.1.15e and constraint 9.1.15f counts the number of opened stations. The domains of the binary decision variables are specified in constraint 9.1.15g. The UBUL model follows the assumption of ambulance independence. An upper bound for the busy fraction of each ambulance is the basis for availability calculations, and demand is assumed to be assigned uniformly to covering ambulances.

The UBUL reduces the potential overcapacities induced by the full demand assignment assumption of previous models. Its ambulance-specific modeling perspective might, however, still lead to overcapacities. Let us assume that we solve the model for an instance with a service level of 95% and that the value for b, which leads to the best objective function value, is 2. Each demand node, therefore, needs to be covered by at least 2 ambulances. The instance might contain a remote demand node, for instance, in a rural region, with a sufficiently low call rate such that one ambulance serving this demand node would be busy less than 5% of its service time, assuming that this ambulance covers only this demand node. The service level of 95% would then be satisfied with a single ambulance, but the UBUL would still require the demand node to be covered by two ambulances.

9.1.3 Contribution of the QRUBUL

Table 9.2 provides an overview of the central assumptions and characteristics of existing models in the class of set covering models, considering the uncertainty of ambulance availability, and highlights the contribution of the QRUBUL model developed in this chapter.

Table 9.2: Comparison of set covering models considering uncertainty in ambulance availability for the ambulance location problem

Model	Availability based on	Perspective	Demand assignment	Reference
BPLSCP	busy fraction	zone-specific	uniform	Revelle & Hogan (1989)
PRLSCP	Poisson distribution	station-specific	full	Ball & Lin (1993)
QPLSCP	Erlang-B	zone-specific	uniform	Marianov & ReVelle (1994)
QRLSCP	Erlang-B	station-specific	full	Borrás & Pastor (2002)
BRLSCP	busy fraction	station-specific	full	Borrás & Pastor (2002)
UBUL	busy fraction	ambulance-specific	uniform	Shariat-Mohaymany et al. (2012)
QRUBUL	Erlang-B	station-specific	system-optimal	–

The QRUBUL extends existing set covering approaches by combining queuing-based capacity bounds with system-optimal demand assignment. The QRUBUL addresses limitations in both the UBUL's uniform demand assumption and QRLSCP's full neighborhood coverage requirement. It builds on the idea underlying UBUL of using upper bounds for the demand that can be served by an ambulance or station. It takes a station-specific perspective for modeling availabilities to avoid potential overcapacities caused by an ambulance-specific perspective. At the same time, it relaxes the assumption of uniform demand assignment to ambulances covering a demand node and is not based on the independence assumption of ambulances at the same station. We base the availability calculations on the Erlang-B formula since Shariat-Mohaymany et al. (2012) list using busy fractions instead of a queuing approach as a possible cause for unfulfilled coverage targets in their experiments. Using busy fractions for calculating station availabilities, as in the BRLSCP, assumes independence between ambulances at one station, which we deem to be a less reasonable assumption compared to modeling each station as an $M/M/k/loss$ queuing system. The QRUBUL simultaneously extends the QRLSCP by relaxing the assumption that a station has to serve the full demand in its neighborhood. This is achieved by calculating upper bounds for the demand rate a station can serve with a given number of ambulances, and letting the model assign demand to covering stations.

9.2 Formulating the QRUBUL

This section presents the QRUBUL model. Section 9.2.1 lists the assumptions underlying the QRUBUL and the notation used in the formal model definition. We formulate the QRUBUL in Section 9.2.2.

9.2.1 Assumptions and Notation

Poisson call arrivals: The arrival of calls at demand nodes is assumed to follow a Poisson distribution. This is a common assumption in the ambulance location literature (Ball & Lin, 1993; Borrás & Pastor, 2002; Marianov & ReVelle, 1994).

Loss system: The calculation of station availability is based on the assumption that calls assigned to a station that arrive while all ambulances at the station are busy do not wait for an ambulance of that station to become free, but are assumed lost from a modeling point of view. These calls would be served by the closest available ambulance from another station in practice.

Station independence: Stations and their assigned demand are modeled as independent queuing systems.

Dispatch from the station: Ambulances are assumed to be dispatched from their station and to return to their station after serving a call. This assumption is the basis for the availability calculation of each station. The assumption does not hold in practice, as ambulances can also be dispatched outside their station. The simulation used to evaluate model solutions relaxes this assumption.

System-optimal demand assignment The QRUBUL model decides on the system-optimal assignment of calls to stations. This implies the assumption that the operational dispatching decision is able to reflect this assignment. As with the assumption of uniform demand assignment, this might not be the case in practice. The system-optimal demand assignment ensures that a sufficient overall capacity of ambulances is provided, and simulation can be used to investigate how the dispatching decision needs to be designed to ensure sufficient operational performance.

Deterministic parameters: The travel times, service times, and call rates are assumed to be known and deterministic.

Location-independent service times: The time for serving a call is assumed independent of the locations of the demand and supply nodes. This assumption will be relaxed in Chapter 10 but is initially kept to enable a comparison with existing models.

Strict spatial coverage: The regulations governing EMS prescribe a percentage of calls that need to be reached within a coverage threshold. This percentage is calculated by aggregating both the spatial and the temporal dimensions. The QRUBUL model assumes that every demand node needs to be covered by a station, enforcing a strict spatial availability, and uses the target percentage to calculate the temporal availability of ambulances.

We use the following notation for formulating the QRUBUL:

Sets

- J: Set of supply nodes representing potential station locations
- I: Set of demand demand nodes representing call locations
- N_i: Set of supply nodes covering demand node $i \in I$
- M_n: Set of demand nodes in the neighborhood of $n \in I \cup J$

Variables

- x_{jk}: Binary variable indicating if k ambulances are allocated to supply node $j \in J$
- p_{ij}: Proportion of demand at demand node $i \in I$ assigned to supply node $j \in J$

Parameters

- α: Service level
- f_i: Daily call rate at demand node $i \in I$
- $\hat{t}$: Upper bound for the service time per call response
- L: Maximum number of ambulances that can be allocated to a station
- β: Weight of the objective of minimizing the number of ambulances
- γ: Weight of the objective of minimizing the number of stations

9.2.2 Model Formulation

The model decides how many ambulances to allocate to each station. The binary variable x_{jk} models this decision and takes the value 1, if k ambulances are stationed at supply node j and 0 otherwise. Binary variables require a capacity limit L for the number of

ambulances per station to bound the solution space. The variable p_{ij} defines the proportion of the demand f_i at demand node i assigned to supply node j in its neighborhood N_i.
An upper bound U_k^α constrains the demand that can be assigned to a station. U_k^α depends on the number k of ambulances assigned to a station and the required service level α of having at least one ambulance available at the station. We assume that call arrivals at a demand node follow a Poisson distribution with a rate of λ calls per time unit and that calls which arrive at a station when no ambulance is available are lost from the perspective of that station. This allows calculating the probability of k calls arriving per time unit with a truncated Poisson distribution as

$$P(X=k) = \frac{e^{-\lambda}\frac{\lambda^k}{k!}}{\sum_{c=0}^{k} e^{-\lambda}\frac{\lambda^c}{c!}}$$

, which is numerically equivalent to the Erlang-B formula.
The probability of a station with k ambulances having at least one ambulance available is equal to the probability that fewer than k calls arrive during an upper bound estimation $\hat{t}$ of the time it takes to serve one call. We therefore calculate the rate λ to express the number of calls per upper bound service time $\hat{t}$. Given a daily call rate f, we set $\lambda = \frac{\hat{t}}{24} f$. The upper bound of the daily call rate f which can be served by a station with k ambulances while maintaining an availability of α can now be calculated as

$$U_k^\alpha = \max_{f \in \mathbb{R}^+} \left\{ \lambda = \frac{\hat{t}}{24} f \,\middle|\, \frac{e^{-\lambda}\frac{\lambda^k}{k!}}{\sum_{c=0}^{k} e^{-\lambda}\frac{\lambda^c}{c!}} \leq 1 - \alpha \right\} \tag{9.2.1}$$

The QRUBUL model is formulated as follows:

$$\min_{x,p} \quad \beta \sum_{j \in J} \sum_{k=1}^{L} k x_{jk} + \gamma \sum_{j \in J} \sum_{k=1}^{L} x_{jk} \tag{9.2.2a}$$

$$\text{s.t.} \quad \sum_{i \in M_j} f_i p_{ij} \leq \sum_{k=1}^{L} U_k^\alpha x_{jk}, \quad \forall j \in J, \tag{9.2.2b}$$

$$\sum_{j \in N_i} p_{ij} = 1, \qquad \forall i \in I, \tag{9.2.2c}$$

$$\sum_{k=1}^{L} x_{jk} \leq 1, \qquad \forall j \in J, \tag{9.2.2d}$$

$$x_{jk} \in \{0,1\}, \qquad \forall j \in J,\, k \in \{1,\ldots,L\}, \tag{9.2.2e}$$

$$1 \geq p_{ij} \geq 0, \qquad \forall i \in I,\, j \in N_i \tag{9.2.2f}$$

The objective function 9.2.2a minimizes the weighted sum of the number of ambulances, weighted by β, and the number of stations, weighted by γ. The left-hand side of constraint 9.2.2b calculates the daily calls assigned to station $j \in J$. The right-hand side ensures that this daily call rate does not exceed the upper bound U_k^α if k ambulances are allocated to station $j \in J$. Constraint 9.2.2c ensures that the demand of each demand node $i \in I$ is fully assigned to stations in its neighborhood N_i. Exactly one number k of ambulances or no ambulances can be allocated to a supply node j, which is enforced by constraint 9.2.2d. Constraints 9.2.2e and 9.2.2f define the decision variable domains.

9.3 Computational Experiments for the EMS in Baden-Württemberg

We compare the QRUBUL to previous models for several instances in Baden-Württemberg in the following. The three most recent models, namely the QRLSCP, the BRLSCP, and the UBUL, serve as a comparison. The objective functions of the QRLSCP and the BRLSCP are extended to include the number of opened stations, as in Eq. (9.2.2a), to facilitate the comparison. The models are solved, and the solutions are simulated to compare the prescribed amount of resources as well as the resulting simulated coverage levels between the models.

The computational experiments aim to investigate whether modeling the ambulance location problem based on the assumption of system-optimal demand assignment can reduce the overcapacities that might be introduced by the full or uniform demand assignment assumption without decreasing the coverage in the resulting system. Shariat-Mohaymany et al. (2012) demonstrated with the UBUL that the uniform demand assignment assumption can reduce the prescribed number of ambulances while maintaining coverage compared to the full demand assignment assumption underlying the BRLSCP and QRLSCP models. The QRUBUL differs from the UBUL not only in the demand assignment assumption but also in how the model calculates the availability of ambulances, as shown in Table 9.3. Considering the BRLSCP and QRLSCP allows analyzing the effects of the assumptions regarding demand assignment and ambulance availability in isolation, which provides the necessary context for the comparison of the UBUL and QRUBUL.

	Ambulance availability	
Demand assignment	Independent	Dependent
Full	BRLSCP	QRLSCP
Uniform	*UBUL*	—
System-Optimal	—	QRUBUL

Table 9.3: Structure of model comparison based on assumptions regarding ambulance availability and demand assignment strategies. The UBUL follows an *ambulance-specific* perspective, while the other models follow a station-specific perspective.

9.3.1 Setting up the Experiments

The models are compared using 14 instances in Baden-Württemberg based on data from 2023. The 11 EMS regions depicted in Fig. 9.3 each define one instance. Three larger additional instances consist of neighboring EMS regions (s. Table 9.4).

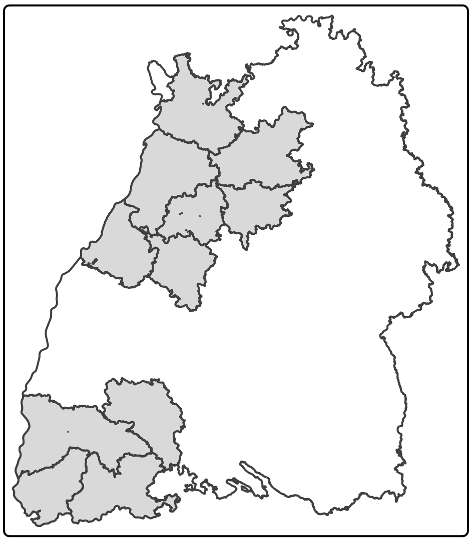


Figure 9.3: EMS regions in Baden-Württemberg that define our instances.

Model Parameters

The models are parametrized as follows for each instance. The *objective function weights* prioritize the minimization of ambulances over the minimization of the number of stations and are set to $\beta = 1$ and $\gamma = \frac{1}{|J|+1}$. The value of γ ensures that the contribution of the number of stations to the objective function value is less than 1. Finding a feasible solution with fewer ambulances will therefore always reduce the objective function value.

K-means clustering of historical call locations is used to calculate the *locations of demand nodes* as described in Section 8.2.1.2. Historical calls are assigned to demand nodes based on the Euclidean distance to calculate the *demand node weights*.

The *supply node locations* are determined using a hexagonal grid with an area of 4 km^2. A hexagonal grid distributes supply nodes uniformly over the area of the instance. We choose this approach since no reason to choose a specific distribution of supply node locations is apparent. The centroids of the hexagons are then snapped to the closest street, resulting in the locations of the supply nodes. The data source for the street network in Baden-Württemberg is OpenStreetMap (OpenStreetMap contributors, 2025). Fig. 9.4 shows the instance of KA with its demand node and supply node locations as an example. Openrouteservice[1], which is based on OpenStreetMap data, is used to estimate the driving times between supply and demand nodes.

[1] https://openrouteservice.org/

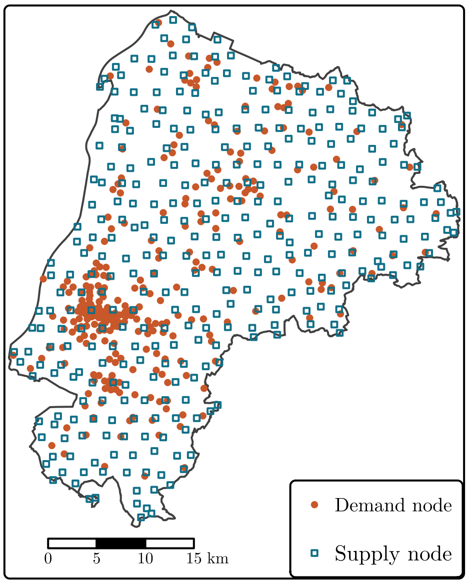


Figure 9.4: Locations of demand and supply nodes in exemplary instance Karlsruhe.

The *maximum station capacity* L is set to 7. This corresponds to the size of the largest current ambulance station in Baden-Württemberg. The *service times* t are based on simulated historical calls with the historical location and allocation of ambulances. The 95th percentile is utilized as the upper bound $\hat{t}$ to mitigate the influence of outliers.

Neighborhood relations between supply and demand nodes are based on the current regulation in Baden-Württemberg. The regulation prescribes a response time target of 12 minutes for the most critical patients. The median time interval in the historical data ahead of the driving time is around 2 minutes. Supply and demand nodes, therefore, are considered neighboring if the driving time between them is 10 minutes or less.

The UBUL model uses a *Big-M* formulation, and the parameter of M needs to be determined. The formulation by Shariat-Mohaymany et al. (2012) uses the same M in constraints 9.1.15b and 9.1.15f. To reduce the magnitude of M, we differentiate between M_1 for constraint 9.1.15b and M_2 for constraint 9.1.15f. M_1 is calculated as $M_1 = \max_{a \in A} \left\{ \sum_{i \in M_a} \dot{t}_i \right\}$. This ensures, that the right-hand side of constraint 9.1.15b is ≤ 0, if $\varkappa_a = 0$. Since $\rho \geq 0$, the constraint will be satisfied. For M_2 we can use the maximum station capacity L. As $|A_j| \leq L \, \forall j \in J$ and $\varkappa_a \in \{0, 1\} \, \forall a \in A$, the constraint 9.1.15f will be satisfied if $y_j = 1$.

The UBUL model requires exogenously setting the minimum number of required ambulances b in the neighborhood of each demand node. The authors of the model suggest varying the parameter and solving the model repeatedly to find the value of b which leads to the lowest objective function value. Preliminary experiments showed that the best bound for $b = 5$ was higher than the best solution for $b < 5$ for all instances. b is therefore varied between 1 and 4 in our experiments. The presented solver times are the sum of solution procedures for the different values of b. All other metrics are based on the solution with the best objective function value.

Instance Characteristics

Table 9.4 lists the characteristics of the instances. The determination of demand and supply node locations, as well as the calculation of the mean and the upper bound of service times, is described in the following. The instances cover rural as well as urban areas, indicated by the call density, which varies from 14.47 calls per km^2 per year to 63.7 calls per km^2 per year. Similarly, the longest mean and upper bound of service times are each around 20% longer than the shortest one. The comparison of the models is therefore based on instances that not only differ in their size but also in their demand structure.

Table 9.4: Description of EMS Location Problem Instances

ID	Calls per km^2	Demand nodes	Supply nodes	$\bar{t}_{\text{service}}$ (h)	$\hat{t}_{\text{service}}$ (h)
1	50.97	138	203	1.18	2.01
2	63.7	146	212	1.28	2.01
3	23.91	141	226	1.31	2.26
4	30.33	180	222	1.22	2.23
5	34.35	197	211	1.27	2.16
6	23.31	220	274	1.23	2.17
7	14.47	237	320	1.32	2.45
8	57.56	287	324	1.23	2.08
9	37.2	257	348	1.2	2.06
10	56.88	285	344	1.24	2.06
11	31.31	341	375	1.1	2.02
12	54.82	445	532	1.21	2.04
13	25.59	822	869	1.17	2.16
14	50.59	866	978	1.22	2.05

Computational and Simulation Setup

The optimization models are solved using Gurobi Optimizer version 12.0.3 on a Windows machine with 16GB RAM and an AMD Ryzen 5 5600X processor with 3700 MHz. A time limit of 10,800 seconds is set for calculating solutions.

The solution of each instance for each model is subsequently simulated with the simulation model described in Chapter 6. For each solution, one year of historical EMS calls is simulated. We employ the variance reduction technique of common random numbers by setting the process times apart from driving times equivalent to the historic process times (Law, 2015, p.588 ff.). In case of missing data, the process times are imputed by drawing from the empirical distributions of existing process times. Therefore, only the travel times differ between the simulation runs of one instance, based on the location and allocation decisions prescribed by the four models. The dispatching decision follows the closest available dispatching strategy in the simulation.

9.3.2 Analyzing the Results

The computational performances of solving the models are analyzed initially to facilitate the interpretation of the following results. We then present the aggregated optimization and simulation results to highlight the key insights.

The subsequent results follow pairwise comparisons of models to analyze the effects of the different assumptions for the demand assignment and ambulance availability calculations. We analyze the differences in the results based on the assumptions of dependence and independence regarding the ambulance availability calculation by comparing the results of the BRLSCP and QRLSCP. We compare the BRLSCP and UBUL to analyze the effects of moving from a full demand assignment assumption to a uniform demand assignment assumption in isolation, and we compare the QRLSCP and QRUBUL to analyze the effect of moving from a full to a system-optimal demand assignment assumption. These comparisons provide context to compare the UBUL and QRUBUL to investigate the difference between the uniform and system-optimal demand assignment assumption.

The models determine coverage based solely on driving times. The simulated coverage, therefore, refers to simulated driving times of 10 minutes or less in the following. Supply nodes to which at least one ambulance has been allocated are referred to as stations when interpreting the results of the optimization models and the simulation.

9.3.2.1 Computational Performance

Table 9.5 presents the aggregated results of computational performances, and Table 9.6 lists the detailed results of solver times and MIPGaps per instance for each model. The BRLSCP and QRLSCP find an optimal solution for all instances, while the UBUL finds one for 11, and the QRUBUL for 8 out of the 14 instances. The BRLSCP and QRLSCP additionally require less computational effort to find optimal solutions compared to the UBUL and QRUBUL. The highest solver times of the UBUL are caused by the need to repeatedly solve the model for different values of b.

Table 9.5: Comparison of aggregated computational performance.

Model	Instances	Feasible	Optimal	Mean MIP gap of non-optimal feasible solutions (%)	Max MIP gap of non-optimal feasible solutions (%)	Mean solver time (s)	Max solver time (s)
BRLSCP	14	14	14	0	0	2.57	15
QRLSCP	14	14	14	0	0	8	61
UBUL	14	14	11	3.76	4.23	6856.79	12704
QRUBUL	14	14	8	2.12	3.01	5023.07	10800

Table 9.5 allows an analysis of the solution qualities and the computational performance across the instances. There is no indication that certain characteristics make an instance harder or easier to solve across all models apart from the instance size. All instances are solved very quickly by the BRLSCP and QRLSCP. The solver times differ significantly between the UBUL and QRUBUL apart from the largest three instances.

Table 9.6: Comparison of computational performance for each instance.

	Solver Time (s)				MIP Gap (%)			
ID	BRLSCP	QRLSCP	UBUL	QRUBUL	BRLSCP	QRLSCP	UBUL	QRUBUL
1	3	**0**	155	70	0	0	0	0
2	**0**	2	578	27	0	0	0	0
3	**0**	5	10833	10	0	0	0	0
4	**0**	**0**	10828	52	0	0	0	0
5	**0**	9	10890	1032	0	0	0	0
6	**0**	1	11477	579	0	0	0	0
7	**0**	**0**	10867	2812	0	0	0	0
8	5	**2**	1574	10800	0	0	0	1.37
9	15	**4**	1350	941	0	0	0	0
10	**0**	12	228	10800	0	0	0	1.11
11	**2**	12	449	10800	0	0	0	0.48
12	9	**1**	12704	10800	0	0	3.29	**2.17**
13	**1**	3	11572	10800	0	0	**0.01**	3.01
14	**1**	61	12490	10800	0	0	4.23	**2.97**

9.3.2.2 Aggregated Results

Fig. 9.5 depicts the ranks of the four models based on the prescribed resource level, and Fig. 9.6 depicts the ranks based on the simulated coverage. The detailed numeric results for all instances and all models can be found in Table C.2.

Figure 9.5: Ranks of models per instance by prescribed resources, ranked from fewest (rank = 1) to most (rank = 4).

The dependence assumption leads to more prescribed resources compared to the independence assumption, as shown by comparing the QRLSCP and BRLSCP. The uniform demand assignment assumption reduces the prescribed resources in all but two instances compared to the full demand assignment assumption, as shown by the comparison of the UBUL to the BRLSCP and QRLSCP. The system-optimal demand assignment assumption of the QRUBUL leads to fewer prescribed resources in all instances compared to the BRLSCP and QRLSCP. The QRUBUL leads to equal or fewer prescribed ambulances in 10 out of the 14 instances compared to the UBUL. The reduction caused by the system-optimal demand assignment assumption and the station-specific modeling perspective might be offset by basing the ambulance availability calculation on the dependence assumption in the QRUBUL compared to the independence assumption of the UBUL.

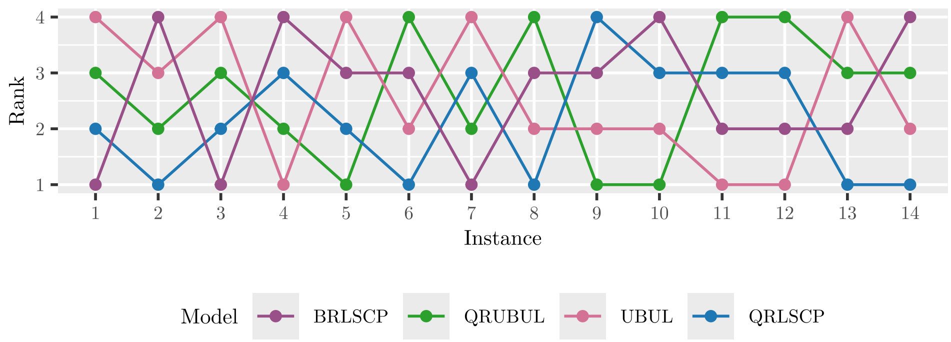


Figure 9.6: Ranks of models per instance by simulated driving time coverage ranked from highest (rank = 1) to lowest (rank = 4).

The number of prescribed resources clearly does not solely determine simulated coverages, and no conclusion can be drawn as to which modeling assumption leads to a higher or lower coverage. The BRLSCP, UBUL, and QRUBUL each achieve the highest coverage in three instances, and the QRLSCP in five instances. Each model achieves the lowest coverage in at least one instance. This indicates that the reduction in prescribed resources of the QRUBUL does not systematically result in a reduced simulated coverage.
To conclude, the QRUBUL prescribes fewer resources in all instances and leads to comparable simulated coverages compared to the BRLSCP and QRLSCP. The QRUBUL leads to fewer prescribed resources in the majority of instances and a similar simulated coverage compared to the UBUL, and requires, on average, less computational effort due to a lower number of binary variables and avoiding the need to repeatedly solve the model.

9.3.2.3 The Full Versus the System-Optimal Demand Assignment Assumption: QRLSCP and QRUBUL

The comparison of the QRLSCP and QRUBUL allows investigating the effects of switching from the full demand assignment assumption to the system-optimal demand assignment assumption, as shown in Table 9.7.

Table 9.7: Comparison of the QRLSCP and QRUBUL. The UBUL follows an *ambulance-specific* perspective, while the other models follow a station-specific perspective.

	Ambulance availability	
Demand assignment	Independent	Dependent
Full	BRLSCP	**QRLSCP**
Uniform	*UBUL*	—
System-Optimal	—	**QRUBUL**

Fig. 9.7 shows the relative differences in objective function values and simulated coverage of the QRUBUL compared to the QRLSCP. We refer to relative differences as follows. The relative difference of a value V_A of model A compared to a value V_B of model B is calculated as $\frac{V_A - V_B}{V_B}$. Values of interest are the simulated driving time coverages and objective function values.

The QRUBUL prescribes fewer ambulances in all instances. This confirms that modeling based on the system-optimal demand assignment assumption reduces the prescribed number of ambulances. The simulation results allow for analyzing whether the reduction in ambulances decreases the system performance measured in the proportion of covered calls. The QRUBUL decreases coverage in 9 out of the 14 instances. The difference in simulated coverage exceeds 1 percentage point in 2 out of these 9 instances, indicating that the difference in simulated coverage is relatively small compared to the differences in objective function value.

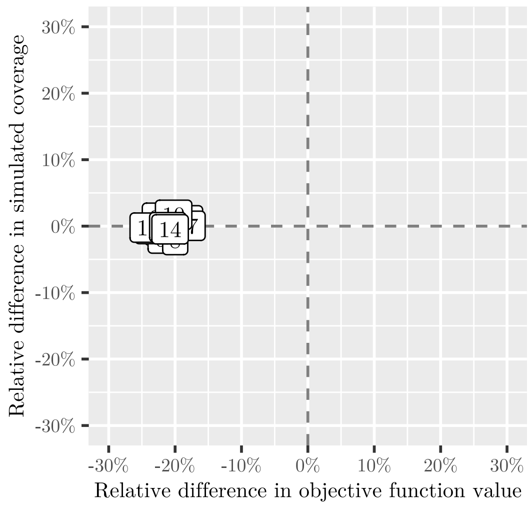


Figure 9.7: Relative differences in simulated coverages and objective function values of the QRUBUL compared to the QRLSCP. Numbers of labels indicate the instance ID (s. Table 9.4).

9.3.2.4 The Independence Versus the Dependence Assumption: BRLSCP and QRLSCP

Comparing the BRLSCP and QRLSCP serves to analyze the effects of the assumption underlying the ambulance availability calculation as shown in Table 9.8. This analysis serves as a context for the following comparison of the UBUL and QRUBUL.

Table 9.8: Comparison of the BRLSCP and QRLSCP. The UBUL follows an *ambulance-specific* perspective, while the other models follow a station-specific perspective.

	Ambulance availability	
Demand assignment	Independent	Dependent
Full	**BRLSCP**	**QRLSCP**
Uniform	*UBUL*	—
System-Optimal	—	QRUBUL

Fig. 9.8 shows the relative differences in objective function value and simulated coverage of the QRLSCP compared to the BRLSCP. Basing the ambulance availability calculation on the dependence assumption leads to more prescribed resources in all instances. The

simulated coverages do not show a systematic difference between the two models. An explanation for the similarity in coverage is that both models lead to an overcapacity of ambulances, which results in similar coverage levels.

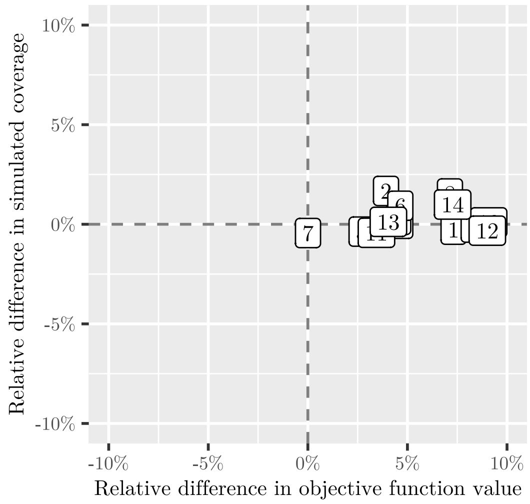


Figure 9.8: Relative differences in simulated coverages and objective function values of the QRLSCP compared to the BRLSCP. Numbers of labels indicate the instance ID (s. Table 9.4).

9.3.2.5 The Uniform Versus the System-Optimal Demand Assignment Assumption: UBUL and QRUBUL

We cannot compare the uniform and system-optimal demand assignment assumptions in isolation since the UBUL and QRUBUL also differ in their assumptions for the ambulance availability calculations and their modeling perspective (s. Table 9.9).

Table 9.9: Comparison of the UBUL and QRUBUL. The UBUL follows an *ambulance-specific* perspective, while the other models follow a station-specific perspective.

	Ambulance availability	
Demand assignment	Independent	Dependent
Full	BRLSCP	QRLSCP
Uniform	***UBUL***	—
System-Optimal	—	**QRUBUL**

The dependence assumption, on which the QRUBUL is based, leads to more prescribed resources compared to the independence assumption, on which the UBUL is based. An indication that the system-optimal demand assignment assumption and station-specific perspective reduce the number of prescribed resources would therefore be if the QRUBUL prescribes a similar or lower number of resources compared to the UBUL. Fig. 9.9 shows the relative differences in objective function values and simulated coverages of the QRUBUL compared to the UBUL, and Table 9.10 details the numeric results.

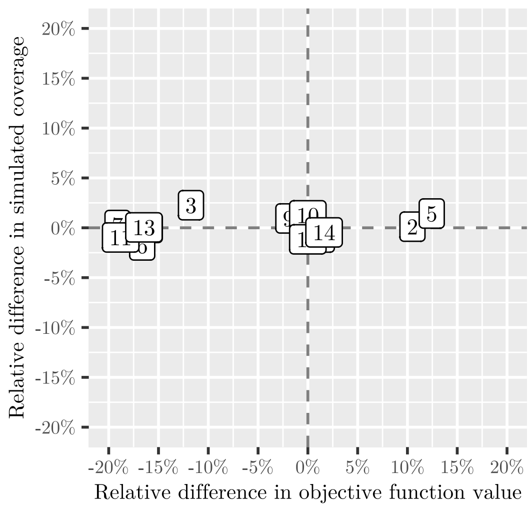


Figure 9.9: Relative differences in simulated coverages and objective function values of the QRUBUL compared to the UBUL. Numbers of labels indicate the instance ID (s. Table 9.4).

The QRUBUL prescribes equal or fewer resources in ten out of the fourteen instances and more in the remaining four instances. This supports our hypothesis that the system-optimal demand assignment assumption and the station-specific modeling perspective can reduce the number of prescribed resources compared to the uniform demand assignment assumption with an ambulance-specific modeling perspective.

Table 9.10: Prescribed resources and simulated coverage for the UBUL and QRUBUL

	Stations		Ambulances		Driving time coverage		$\Delta^{max}_{driving\_coverage}$	
ID	UBUL	QRUBUL	UBUL	QRUBUL	UBUL	QRUBUL	UBUL	QRUBUL
1	**16**	**16**	**34**	**34**	**92.07%**	92.92%	**-0.85%pt.**	0%pt.
2	**18**	**18**	**38**	42	**93.47%**	93.57%	**-0.1%pt.**	0%pt.
3	**17**	**17**	34	**30**	**88.33%**	90.33%	**-2%pt.**	0%pt.
4	**22**	**22**	44	**37**	93.63%	**93.61%**	0%pt.	**-0.02%pt.**
5	18	**16**	**32**	36	**90.54%**	91.83%	**-1.29%pt.**	0%pt.
6	**20**	**20**	42	**35**	90.5%	**88.9%**	0%pt.	**-1.6%pt.**
7	**25**	**25**	47	**38**	**95.19%**	95.46%	**-0.27%pt.**	0%pt.
8	33	**32**	**71**	72	88.26%	**87.39%**	0%pt.	**-0.87%pt.**
9	**24**	**24**	51	**50**	**90.74%**	91.59%	**-0.85%pt.**	0%pt.
10	29	**28**	**67**	**67**	**89.97%**	91.09%	**-1.13%pt.**	0%pt.
11	**41**	**41**	85	**69**	96.66%	**95.72%**	0%pt.	**-0.94%pt.**
12	46	**44**	**101**	**101**	92.38%	**91.31%**	0%pt.	**-1.07%pt.**
13	86	**85**	170	**142**	**96.24%**	96.28%	**-0.04%pt.**	0%pt.
14	89	**84**	**184**	187	90.74%	**90.32%**	0%pt.	**-0.42%pt.**

The QRUBUL differs from the UBUL in both the demand assignment assumption and the modeling perspective, and we can try to discern which difference has a stronger influence on the differences in the solutions.

We first analyze the demand assignment assumptions. Differences in demand assignment assumptions affect the solution if demand nodes are covered by multiple stations. We would therefore expect a correlation between the proportions of demand nodes with multiple neighboring stations and the differences in objective function value if the demand assignment assumption has a strong influence on the solutions. Fig. 9.10 depicts the proportion of demand nodes with multiple neighboring stations in the solutions of the QRUBUL and the difference in objective function value of the QRUBUL compared to the UBUL.

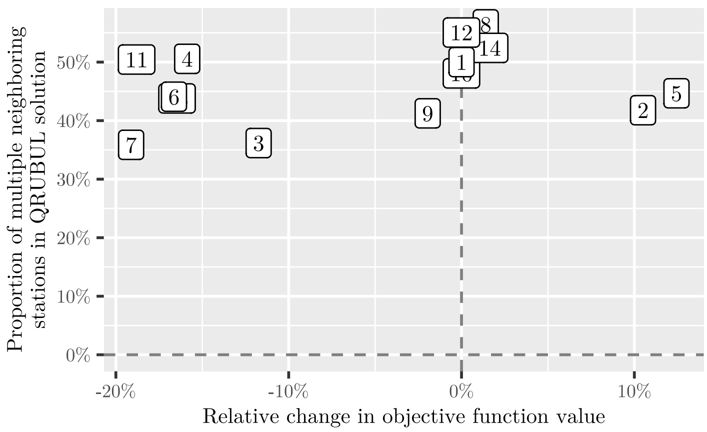


Figure 9.10: Relative differences in objective function values of the QRUBUL compared to the UBUL and the proportions of demand nodes with more than one neighboring station in the QRUBUL solution.

No correlation between the proportion of demand nodes with multiple neighboring stations and the difference in objective function value is apparent. We therefore have to conclude that, at least for the instances in our experiments, switching from the uniform to the system-optimal demand assignment assumption does not appear to have a significant effect on the prescribed resources compared to other factors.

Secondly, we can analyze the effects of switching from an ambulance-specific to a station-specific perspective for the availability calculations. As reasoned in Section 9.1.2.7, the ambulance-specific perspective can lead to overcapacities due to enforcing a minimum number of covering ambulances for each demand node. Setting the lower bound of covering ambulances to 2 in the UBUL led to the best objective function value in all instances. This might induce overcapacities if the demand of some nodes is sufficiently low to be covered by a single ambulance. We would therefore expect to see a correlation between the proportion of demand nodes that are covered by a single ambulance and the difference in objective function value. Fig. 9.11 depicts the proportions of demand nodes whose demand is assigned to a station with a single ambulance in the QRUBUL solution and the relative differences in objective function values of the QRUBUL compared to the UBUL.

We can see a clear correlation between the difference in objective function value and the proportion of demand nodes whose demand is assigned to a station with a single ambulance in the QRUBUL solution. This supports the assumption that the ambulance-specific perspective can lead to overcapacities due to the necessary lower bound of covering ambulances.

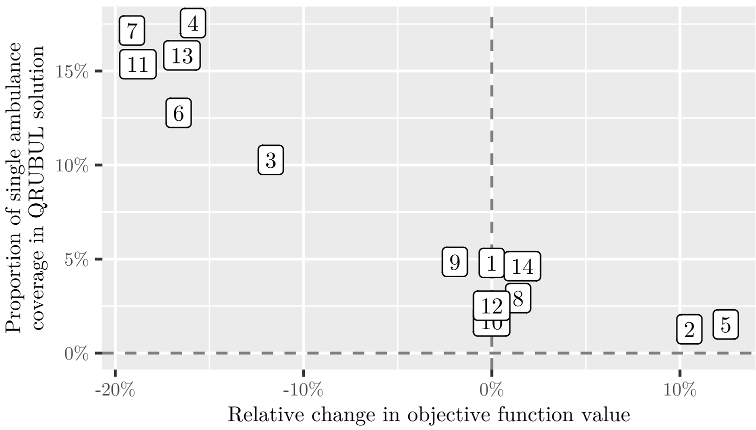


Figure 9.11: Relative differences in objective function values of the QRUBUL compared to the UBUL and the proportions of demand nodes covered by a single ambulance in the QRUBUL solution.

The following conclusions can be drawn based on the computational experiments:

- The system-optimal demand assignment effectively reduces overcapacities compared to the full demand assignment assumption without systematically reducing simulated coverage levels.
- The dependence assumption in the calculation of ambulance availabilities leads to more prescribed ambulances compared to the independence assumption in our instances.
- Modeling availabilities based on a station-specific perspective can reduce overcapacities compared to an ambulance-specific perspective due to avoiding a lower bound on the number of covering ambulances per demand node.

9.4 The QRUBUL: Discussion

This chapter aimed to develop a model for the ambulance location problem that relaxes the full and uniform demand assignment assumptions. The full and uniform demand assignment assumptions can lead to overcapacities. The overcapacities are assumed to increase when considering heterogeneous patients with differing coverage thresholds, since this will increase the multiple coverage of demand nodes by different stations. The relaxation of these assumptions is therefore seen as a first step towards a planning approach that considers heterogeneous patients.

Contributions

We have proposed the QRUBUL based on a system-optimal demand assignment. The QRUBUL extends the QRLSCP. The calculation of the availability of a station is based on the Erlang-B formula as in the QRLSCP, but the full demand assignment assumption is relaxed to the system-optimal demand assignment assumption. The QRLSCP enforces a

sufficient joint availability of stations covering a demand node while the QRUBUL enforces a minimum availability for each station. If every station is the sole covering station for at least one demand node, there is no difference between the two approaches since each station would need to fulfill the availability requirement in both models.
Lastly, the QRUBUL is closely related to the UBUL. Both models build on upper bounds for the availability of ambulances. The QRUBUL differs from the UBUL by basing the availability calculations on the Erlang-B formula instead of on busy fractions and extends it by relaxing the uniform demand assignment assumption to the system-optimal demand assignment assumption. Additionally, the QRUBUL takes a station-specific instead of an ambulance-specific perspective for calculating ambulance availabilities.
The computational experiments confirmed that the QRUBUL reduces the prescribed resources compared to the QRLSCP without systematically reducing coverage levels in simulation experiments. This supports the hypothesis that a system-optimal demand assignment can reduce overcapacities introduced by the full demand assignment assumption. The QRUBUL prescribed equal or fewer resources compared to the UBUL in ten out of fourteen instances. The results indicate that the reduction in resources stems more from switching from the ambulance-specific perspective of the UBUL to a station-specific perspective than from the difference between the uniform and the system-optimal demand assignment assumption. At the same time, the results do not suggest that the system-optimal demand assignment assumption leads to a reduced simulated coverage compared to the uniform demand assignment assumption.

Limitations

The QRUBUL model is still based on several limiting assumptions. Service times are assumed to be location-independent. This assumption will be relaxed in Chapter 10.
The system-optimal demand assignment assumption implies that the operational dispatching decision is able to dispatch ambulances in such a way that the necessary availability can be achieved. It is unclear to what degree this can be achieved in practice.
The availabilities of the ambulances at one station are not assumed to be independent, but the availability of a station is still assumed to be independent of other stations. At the same time, the ambulance availability calculation is based on a loss-system assumption. If no ambulance is available when a call arrives at a station, the call is not assumed to wait for an ambulance to become available but to be served by another station. The model does not consider this possible cascading effect of forwarded calls.
Ambulances are assumed to be dispatched only from their station. Ambulances can also be dispatched, for instance, on the way from a hospital back to their station in practice.
The QRUBUL enforces strict spatial coverage, requiring each demand node to be covered by at least one station. The regulation would allow relaxing strict spatial coverage if a sufficient proportion of calls is reached within the coverage threshold.

The QRUBUL contains several deterministic parameters that are not deterministic in practice. Specifically, the driving times and call rates are assumed to be deterministic in the QRUBUL, but exhibit variance in the historical data.
We used car driving times from Openrouteservice to estimate the ambulance driving times between supply and demand nodes. These driving times might not necessarily be equal to ambulance driving times, especially when considering the use of sirens.
We assume a 24/7 allocation of ambulances to stations in our simulation experiments. This might mask differences between models, especially during periods of low demand.

Methodological Reflections

We made multiple methodological choices that influence the interpretability of the results. The QRUBUL differs from the UBUL not only in the demand assignment assumption but also in how the availability of ambulances and stations is calculated. The QRUBUL availability calculation is based on queueing theory. The advantage of this approach is being able to calculate all necessary parameters based on instance data. The UBUL model requires setting the exogenous parameter of the busy fraction, which necessitates repeatedly solving the model to find the best value. Developing a model based on the system-optimal demand assignment and on busy fractions would, however, have eliminated the influence of the different availability calculations when comparing the models and might have allowed a clearer understanding of how the modeling perspective (station-specific or ambulance-specific) and the demand assignment (system-optimal or uniform) affect the model solutions.
The two objectives of minimizing the number of ambulances and minimizing the number of stations were handled by the weighted sum approach. The chosen weights prioritize the objective of minimizing the number of ambulances. This could also have been achieved by employing a lexicographic approach, where the model is initially solved only for the objective of minimizing ambulances. The identified minimum number of ambulances would then be set as an upper bound constraint, and the model would be solved again with the objective of minimizing the number of stations. It would be interesting to compare the computational performance of both approaches.
We have focused on extensions of the Location Set Covering Problem since these models directly fit our goal of minimizing resources while maintaining coverage targets. It could, however, also have been possible to reformulate models that were initially designed to maximize a performance criterion with a given number of ambulances. This could have broadened the comparison of our model to existing approaches.
The instances in our computational experiments are based on real-world data from the EMS in Baden-Württemberg, Germany. While this enables comparing the models with realistic data, it limits the transferability of conclusions to other EMS systems.
We set a time limit for solving the optimization models. Not all instances of all models could be solved to optimality. Solving all instances of all models to optimality would

strengthen the conclusions drawn from the results. The solution procedure of the UBUL could also be improved when the model is solved sequentially for different values of b. Solving the model for the initial value of b gives us an initial objective function value if a feasible solution can be found. This objective function value can be used as an upper bound for the objective function value in the second iteration, where the model is solved with the next value of b. The best objective function of the solved iterations can be used as an upper bound from the third iteration onwards. This would tighten the model formulation and could reduce the solution times.

The simulation of the solutions is based on the same data used to parametrize the instances. This facilitates attributing differences in the simulation results to differences in the models by eliminating the influence of changes in the data between the optimization and the simulation. Simulating unseen call data would, however, enable analyzing how well the model solutions perform for unseen data. This would be interesting, especially from a practical planning point of view, where the models would be used to plan for future unknown call demand.

Capturing the ambulance utilizations in the simulation would enable analyzing ambulance and station availabilities and comparing them to the availabilities assumed by the optimization models. This could enhance especially the comparison of the dependence and independence assumptions in calculating the ambulance availabilities in the models.

The results of the models are influenced by the parameter choices in our computational experiments, and the results need to be interpreted in light of these choices. We used a clustering approach to determine the demand node locations to increase the number of demand nodes in regions with a high call density and avoid unnecessary computational effort due to locating demand nodes in regions with no or very few calls. The demand nodes could also have been determined using, for instance, a square or hexagonal grid structure or a mixture of a grid and clustering approach. It could also be interesting to compare how a lower or higher number of demand nodes affects the results, especially the simulation results.

The upper bound service times used in the experiments are based on the 95th percentile of historic service times. This conservative choice of the upper bound service time might lead to a conservatively large number of ambulances. The relatively small differences in simulated coverage levels compared to the differences in objective function values suggest that all models might still prescribe more ambulances than necessary, and a sensitivity analysis for this parameter might have generated further insights.

The demand at each demand node equals the average daily historic call rate. Emergency call rates fluctuate significantly during the day and between weekdays and weekends. The ambulance location problem determines the size of each station, and the next step would be to determine shift schedules and the number of ambulances per shift. Given a known shift schedule, for instance, three 8-hour shifts per day, the demand could also be taken from the shift with the highest demand to ensure sufficient capacity of the stations.

Outlook

Our work presents several avenues for further research. The QRLSCP and QRUBUL are based on the assumption that call arrivals follow a Poisson distribution. This is a common assumption in the ambulance location problem, but assessing the validity of this assumption based on real-world data would either validate existing modeling approaches or open up further avenues of research. Krueger and Schimmelpfeng (2013) concluded that the arrival of calls at dispatch centers more closely resembles a Pólya distribution than a Poisson distribution. The explanation for the deviation in call arrivals at the dispatch center from the Poisson distribution is that one emergency can lead to multiple emergency calls to the dispatch center. This reasoning does not hold for the ambulance location problem, where we are only interested in the emergencies themselves. Testing whether emergency occurrences follow a Poisson distribution based on real-world data would, however, still be a valuable contribution. While we use the assumption of Poisson call arrivals in the QRUBUL, the model is valid for any discrete call arrival distribution with an additive property and a scaling property.

The QRUBUL is based on deterministic driving times and call rates. As these parameters exhibit variance in the historic data, modeling them as stochastic parameters might improve the solutions. There are approaches to integrate stochastic demand into the ambulance location problem, such as Beraldi and Bruni (2009), and also for considering stochastic travel times, as in van den Berg et al. (2016). These approaches could provide a starting point to integrating stochastic demand and travel times in the QRUBUL.

Another direction of further research concerns the solution procedures. We have used solution procedures implemented in off-the-shelf solvers, and not all instances could be solved to optimality within a time limit of 10,800 seconds on the available machine. The ambulance location problem is a strategic problem where it is not necessary to solve instances very quickly. It could, however, be beneficial to solve larger instances, for instance, when planning for the EMS in a federal state, to avoid over- or undersupply induced by disaggregating the EMS system based on political borders within the federal state. More efficient solution procedures could enable solving larger instances in a reasonable time, which would be valuable, especially from a practical planning point of view.

The optimization and simulation results showed differences in comparative model behavior across the instances. It is yet unclear how instance characteristics influence the comparative model behavior and whether specific characteristics would make one or another modeling approach more suitable. Investigating how the model results depend on instance characteristics could enable a better understanding of which modeling approaches should be used for which instance.

The QRUBUL aims to model the decision maker's primary objective of minimizing resources while fulfilling legal regulations by following a set covering approach. Set covering approaches build on coverage targets, and we developed utility functions in Part II to address limitations of coverage targets. It could therefore be interesting to follow a performance maximization approach, which could allow to explicitly integrate utility functions into a model for the ambulance station location and sizing problem. We perceive a set covering approach for heterogeneous patients as a necessary first component of providing decision support for decision makers. Set covering models can have more than one optimal solution, and integrating utility functions into a performance maximization approach could therefore improve patient utilities without increasing resources.

9.5 The QRUBUL: Conclusion

Existing set covering models for the ambulance location problem are based on either the full or uniform demand assignment assumption. These assumptions can lead to overcapacities, which are likely to increase when considering heterogeneous patients with differentiated coverage thresholds. We propose the QRUBUL based on the system-optimal demand assignment assumption to overcome this problem and build the foundation for a model that considers heterogeneous patients. Real-world data from the EMS in Baden-Württemberg is used to design and conduct computational experiments. The experiments compare the QRUBUL to the existing BRLSCP, QRLSCP, and UBUL by solving several instances and simulating the resulting solutions. The results confirm that moving from a full demand assignment assumption to the system-optimal demand assignment assumption reduces the number of prescribed ambulances while resulting in comparable coverages in simulation experiments. The results did not indicate a significant difference between the uniform and system-optimal demand assignment assumption, but the QRUBUL still led to fewer prescribed resources and comparable simulated coverages in the majority of instances compared to the UBUL due to taking a station-specific modeling perspective for the ambulance availability calculations.

The QRUBUL constitutes a first step towards a model considering heterogeneous patients as present, for instance, in the current EMS regulations in Baden-Württemberg. The QRUBUL is still based on several limiting assumptions, such as the strict spatial coverage assumption and location-independent service times. The location-independent service time assumption will be relaxed in the following chapter Chapter 10 before the model is extended to consider heterogeneous patients.

Chapter 10

Extending the QRUBUL to Heterogeneous Regions and Patients

Developing the QRUBUL in Chapter 9 was the first step towards an ambulance location model considering heterogeneous patients. The QRUBUL relaxes the full and uniform demand assignment assumptions present in previous models. The QRUBUL is still based on several limiting assumptions listed in Section 9.2.1, and it is yet unclear which assumptions become less justifiable when considering heterogeneous patients. We therefore examine which assumptions are impacted when considering heterogeneous patients.

Considering heterogeneous patients does not impact the assumption of Poisson-distributed call arrivals. The scaling property of the Poisson distribution allows calculating the call arrival rates of different patient types at a demand node by multiplying by the expected proportion of calls of a patient type.

Considering heterogeneous patients does not change the justification for viewing every station and its assigned demand as an independent queuing system with losses. It becomes generally less justifiable with lower service levels, as a larger amount of calls is assumed lost, and forwarding calls between stations is not considered in the model. Since heterogeneous patients do not necessarily lead to lower service levels, this issue is not connected to considering heterogeneous patients compared to considering a singular patient type.

The assumption of dispatching ambulances only from their station does not become more or less reasonable compared to considering one patient type. It is clear that this assumption does not align with the actual EMS practice, where ambulances can be dispatched, for example, when returning to their station from a hospital, but considering heterogeneous patients does not necessarily increase the likelihood of dispatching ambulances outside their station.

Using deterministic parameters for travel times, service times, and call rates is a common assumption in ambulance location models. We see no indication that the validity of using deterministic parameters is affected by considering heterogeneous patients.

Assuming location-independent service times is already an unrealistic assumption when considering only one patient type. Fig. 10.1 depicts two exemplary hospital—station pairs, one in the more rural northeastern part of the EMS region, and one pair in the city center. The station in the northeastern part will have significantly longer transport times and return times in the case of transporting a patient. Location-independent service times additionally ignore that the driving times reduce if a station is located closer to its assigned demand. This leads to different service times for stations, assuming that the remaining components of the service time are comparable.

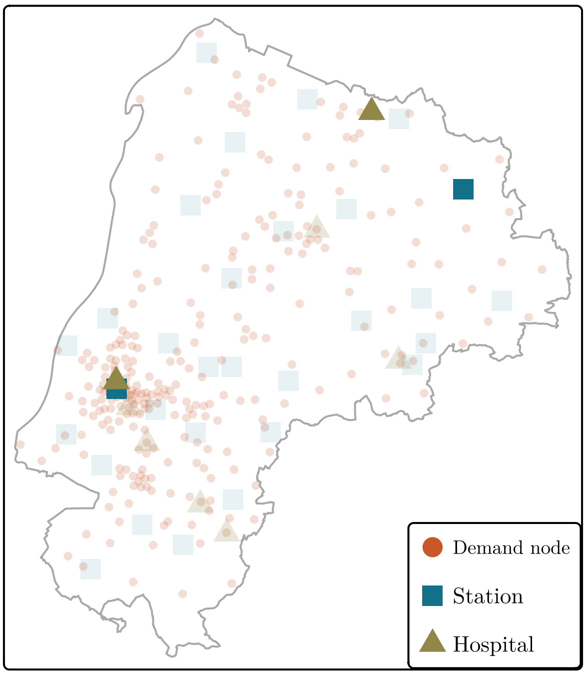


Figure 10.1: Motivating example for relaxing the location-independent service time assumption

Patient types can be differentiated by their coverage threshold, leading to higher differences in driving times compared to considering a single patient type. This makes the assumption of location-independent service times even less justifiable. We extend the QRUBUL to the Queuing Reliability Upper-Bound Unavailability Location Problem for Heterogeneous Regions (QRUBUL-HR) to relax the assumption of location-independent service times in Section 10.2.2.

The last assumption of the QRUBUL is the assumption of strict spatial coverage. We assume that every demand node must be reachable from a station within its coverage threshold and that the service level for covered calls is influenced only by the allocated ambulance capacities. This assumption can become more restrictive when considering heterogeneous patients who differ in their coverage thresholds and service levels. We extend the QRUBUL-HR to the Queuing Reliability Upper-Bound Unavailability Location Problem for Heterogeneous Regions and Heterogeneous Patients (QRUBUL-HR-HP) to consider heterogeneous patients in Section 10.2.3. We will discuss different alternatives for integrating heterogeneous patients, which partly relax the assumption of strict spatial coverage.

The remainder of this chapter is structured as follows. Section 10.1 reviews related approaches to consider heterogeneous regions and heterogeneous patients in the ambulance location problem. We present the formulations of the QRUBUL-HR and QRUBUL-HR-HP in Section 10.2. The computational experiments conducted in Section 10.3 firstly compare the QRUBUL-HR to the QRUBUL to investigate the impact of relaxing the location-independent service time assumption. We secondly compare the QRUBUL-HR-HP in two variants with different sets of patient types to the QRUBUL-HR to analyze how considering heterogeneous patients impacts the required ambulances and stations. Our approach and findings are discussed in Section 10.4, and Section 10.5 concludes the chapter.

10.1 Related Models Considering Heterogeneous Regions and Patients

We presented a general overview of the literature on the ambulance location problem in Section 9.1, and this section focuses on approaches that consider heterogeneous patients or heterogeneous regions.

Knight et al. (2012) presented, to the best of our knowledge, the first approach to consider heterogeneous patients in an ambulance location model. They formulate the Maximal Expected Survival Location Model for Heterogeneous Patients (MESLMHP) which maximizes a weighted sum of expected performance measured by different criteria depending on the patient type. They compare two variants of the MESLMHP. The first variant is based on three patient types with response time coverage thresholds of 8, 14, and 21 minutes, respectively. The second variant uses the survival calculated as a function of response time with the survival function depicted in Fig. 5.1 as the performance criterion for one patient type. Three additional patient types with the coverage thresholds of 8, 14, and 21 minutes are considered as well. Their experiments demonstrate that considering heterogeneous patients in the ambulance location problem can increase performance for the most critical patient types. The MESLMHP differs from our approach firstly in that it maximizes performance with a given number of ambulances and secondly by basing ambulance availabilities on busy fractions instead of on queuing theory.

McCormack and Coates (2015) follow a similar approach to Knight et al. (2012) by maximizing the weighted sum of performance criteria for two patient types. The performance criterion for the first patient type is calculated with the same survival function employed by Knight et al. (2012), and the performance criterion for the second patient type is based on a coverage threshold of 8 minutes.

Leknes et al. (2017) introduced the first model considering heterogeneous patients and heterogeneous regions by integrating location-dependent service times. Their Maximum Expected Performance Location Problem for Heterogeneous Regions (MEPLP-HR) maximizes the weighted sum of expected performance calculated by differentiated performance criteria per patient type. They conduct computational experiments based on one patient type for whom the performance is calculated using a survival function, and two additional patient types with coverage thresholds of 12 and 25 minutes, respectively. They calculate location-dependent service times T_{ij} for the combinations of demand nodes i and supply nodes j based on travel times T^T_{ji} between supply and demand nodes, travel times T^T_{hj} between nodes and hospitals and an on-scene time T^S. The probability p^H for transporting a patient is considered, since not every call results in transporting a patient to a hospital. The service times are then calculated as:

$$T_{ij} = T^T_{ji} + T^S + p^H(T^T_{ih} + T^T_{hj}) + (1 - p^H)T^T_{ij}$$

Ambulance availabilities are calculated using a Markov queuing model based on a Poisson process. The non-linear relations between the ambulance location and allocation decisions and the ambulance availabilities are linearized using special ordered sets of type 2.

H. Andersson et al. (2020) employ the MEPLP-HR in a case study and extend the considered patient types by a fourth type whose performance criterion is whether patients of that type can reach a hospital within 90 minutes. To the best of our knowledge, this constitutes the first consideration of a prehospital time-based coverage threshold.

Both heterogeneous patients and heterogeneous regions have been considered in existing ambulance location models. These models have all been designed to maximize performance with a given number of ambulances and stations. To the best of our knowledge, the QRUBUL-HR and QRUBUL-HR-HP are the first models to integrate heterogeneous patients and heterogeneous regions into a set covering approach for the ambulance location problem.

10.2 Formulating Models for Heterogeneous Regions and Patients

10.2.1 Assumptions and Notation

The following assumptions underlie the QRUBUL-HR and QRUBUL-HR-HP formulations. All but the last two assumptions are adopted from the QRUBUL and are described in Section 9.2.1, and we have discussed the impact of considering heterogeneous patients on the validity of the assumptions in the introduction to this chapter. We list the assumptions in the following for a compact and comprehensive presentation of our model, including the assumptions, the notation, and the formulation.

Poisson call arrivals: The arrival of calls at demand nodes is assumed to follow a Poisson distribution.

Loss system: The calculation of station availability is based on the assumption that calls assigned to a station that arrive while all ambulances at the station are busy do not wait for an ambulance of that station to become free, but are assumed lost from a modeling point of view.

Dispatch from the station: Ambulances are assumed to be dispatched from their station and to return to their station after serving a call. This assumption is the basis for the availability calculation of each station.

System-optimal demand assignment: The QRUBUL decides on the system-optimal assignment of calls to stations.

Deterministic parameters The travel times, service times, and call rates are assumed to be known and deterministic.

Partial strict spatial coverage: EMS regulations prescribe a percentage of calls that need to be reached within a coverage threshold. This percentage is calculated by aggregating both the spatial and the temporal dimensions. The QRUBUL-HR assumes that every demand node needs to be covered by a station, enforcing a strict spatial availability, and uses the target percentage to calculate the temporal availability of ambulances. The QRUBUL-HR-HP partially relaxes this assumption. The assumption still holds for patient categories with the highest prescribed percentage. The strict spatial coverage is relaxed for demand nodes with no supply nodes in their neighborhood for patient categories with a lower percentage.

Location-dependent service times: We consider individual service times for each pair of demand and supply nodes. The QRUBUL-HR-HP additionally considers the patient type for the service time parameter.

We use the following notation for formulating the QRUBUL-HR and QRUBUL-HR-HP:

Sets

- J: Set of supply nodes representing potential station locations
- I: Set of demand demand nodes representing call locations
- Θ: Set of patient types
- N_i: Set of supply nodes covering demand node $i \in I$
- N_i^θ: Set of supply nodes in the neighborhood of i for patient type θ.
- M_n: Set of demand nodes in the neighborhood of $n \in I \cup J$
- M_j^θ: Set of demand nodes in the neighborhood of j for patient type θ

Variables

- x_{jk}: Binary variable indicating if k ambulances are allocated to supply node $j \in J$
- p_{ij}: Proportion of demand at demand node $i \in I$ assigned to supply node $j \in J$
- p_{ij}^{θ}: Proportion of demand of patient type $\theta \in \Theta$ at demand node $i \in I$ assigned to supply node $j \in J$

Parameters

- α: Service level
- f_i: Daily call rate at demand node $i \in I$
- $\hat{t}_{ij}$: Upper bound for the service time if a call at demand node $i \in I$ is served by an ambulance stationed at supply node $j \in J$
- $\hat{t}_{ij}^{\theta}$: Upper bound for the service time if a call of patient type $\theta \in \Theta$ at demand node $i \in I$ is served by an ambulance stationed at supply node $j \in J$
- L_j: Maximum number of ambulances that can be allocated to station $j \in J$
- β: Weight of the objective of minimizing the number of ambulances
- γ: Weight of the objective of minimizing the number of stations

10.2.2 The Queuing Reliability Upper Bound Location Model for Heterogeneous Regions

The QRUBUL-HR extends the QRUBUL to relax the assumption of location-independent service times and has a similar model structure. The main decision is the allocation of ambulances to supply nodes modeled by the decision variable x_{jk}. The value equals 1, if exactly $k \in \{1, \ldots, L_j\}$ ambulances are stationed at supply node $j \in J$, and 0 otherwise. The QRUBUL-HR additionally determines the proportion p_{ij} of demand f_i from demand node $i \in I$ assigned to supply node j. Demand can only be assigned to supply nodes in the neighborhood N_i of demand node i. The objective of the decision remains minimizing the required resources.

The first and main difference between the QRUBUL and QRUBUL-HR lies in the relaxation of location-independent service times for the calculation of ambulance capacities. This is achieved by calculating the number of calls per upper-bound service time assigned to a supply node within the model. The calls per day f_i of demand node i are scaled to calls per hour through multiplying with $\frac{1}{24}$. The calls per hour are scaled to calls per service time by multiplying by the location-dependent upper bound service time in hours $\hat{t}_{ij}$. This allows calculating the calls per service time assigned to a supply node. The assigned calls per service time need to remain below a threshold Λ_k^{α} given that k ambulances are allocated to supply node j. This ensures that at least one ambulance will be available when a call arrives at the station with a probability of α. The threshold Λ_k^{α} is calculated using the truncated Poisson distribution:

$$\Lambda_k^\alpha = \max\left\{\lambda \in \mathbb{R} \middle| \frac{e^{-\lambda}\frac{\lambda^k}{k!}}{\sum_{c=0}^{k} e^{-\lambda}\frac{\lambda^c}{c!}} \leq 1-\alpha\right\}$$

The second difference between the QRUBUL and QRUBUL-HR lies in using station-specific capacity limits L_j for each supply node j instead of one global capacity limit L. The capacity limit L_j of a supply node j can be set to a value smaller than L, if less than L ambulances are sufficient to serve the demand of all demand nodes $i \in M_j$ in the neighborhood of j. L_j is calculated as follows:

$$L_j = \min\left\{\min\left\{k \in \mathbb{N} \middle| \lambda_j^{max} \leq \Lambda_k^\alpha\right\}, L\right\}$$

with

$$\lambda_j^{max} = \sum_{i \in M_j} \frac{1}{24} f_i \hat{t}_{ij}$$

The number of decision variables x_{jk} depends on the values of L_j. The solution space of the QRUBUL-HR will therefore be reduced if the value of L_j is smaller than L for at least one supply node j.

The QRUBUL-HR is formulated as follows:

$$\min_{x,p} \quad \beta \sum_{j \in J} \sum_{k=1}^{L_j} k x_{jk} + \gamma \sum_{j \in J} \sum_{k=1}^{L_j} x_{jk} \tag{10.2.1a}$$

$$\text{s.t.} \quad \sum_{i \in M_j} \frac{1}{24} f_i \hat{t}_{ij} p_{ij} \leq \sum_{k=1}^{L_j} \Lambda_k^\alpha x_{jk}, \quad \forall j \in J, \tag{10.2.1b}$$

$$\sum_{j \in N_i} p_{ij} = 1, \quad \forall i \in I, \tag{10.2.1c}$$

$$\sum_{k=1}^{L_j} x_{jk} \leq 1, \quad \forall j \in J, \tag{10.2.1d}$$

$$0 \leq p_{ij} \leq 1, \quad \forall i \in I,\ j \in N_i, \tag{10.2.1e}$$

$$x_{jk} \in \{0,1\}, \quad \forall j \in J,\ k \in \{1,\ldots,L_j\} \tag{10.2.1f}$$

The objective function 10.2.1a minimizes the weighted sum of the number of allocated ambulances, weighted by β, and the number of opened stations, weighted by γ. Constraint 10.2.1b ensures that at least one ambulance is available with a service level of at least α when a call arrives at a station. The full demand of each demand node needs to be assigned to a station covering the demand node, which is modeled by constraint 10.2.1c. Constraint 10.2.1d ensures that each supply node can be allocated exactly one number of ambulances or no ambulances. The domains of the decision variables are defined in constraints 10.2.1e and 10.2.1f.

10.2.3 The Queuing Reliability Upper Bound Location Model for Heterogeneous Regions and Heterogeneous Patients

The QRUBUL-HR-HP extends the QRUBUL-HR to consider multiple patient types. A patient type $\theta \in \Theta$ is characterized by how the neighborhoods M_i^θ and N_i^θ of supply nodes and demand nodes $i \in I \cup J$ are calculated. Additionally, each patient type can exhibit a separate service level α^θ of calls that are required to be covered.
We assume that the dispatching decision during EMS operations can distinguish between patient types. This is a strong assumption, but it allows investigating how differentiating between patient types can affect the required resources if the dispatching decision can distinguish between them. The assumption allows assigning calls from one demand node $i \in I$ to different supply nodes depending on the patient type θ. The decision variable modeling the call assignment p_{ij}^θ is therefore defined for each combination of demand node i, supply node $j \in I$, and patient type θ. The daily call rates f_i^θ as well as the service times $\hat{t}_{ij}^\theta$ are also defined for each patient type θ in the QRUBUL-HR-HP. The differentiation of service times can, for instance, reflect that the average treatment time on scene or the probabilities of requiring transportation to a hospital differ between patient types.
Considering differentiated service levels α^θ leads to the question of how they can be reflected in the model. We formulate three variants, two of which are excluded from further analysis due to either model complexity or strong necessary assumptions. We will present the core model formulation, which is used for the subsequent computational experiments, as well as the two alternative variants. We assume that $\alpha^1 \geq \alpha^2 \geq \ldots \geq \alpha^{|\Theta|}$.

10.2.3.1 Formulating the Core Model

The core QRUBUL-HR-HP formulation considers differentiated service levels α^θ by setting $\alpha = \alpha^1$. This ensures a sufficient capacity at each station to reach the required service level α^θ for every patient type $\theta \in \Theta$. The patient categories θ with $\alpha^\theta < \alpha^1$ will be covered at a service level higher than required, since we still assume strict spatial coverage. In addition, a demand node i can have an empty neighborhood $N_i^\theta = \emptyset$ for patient types $\theta \in \Theta : \alpha^\theta < \alpha^1$, especially when the neighborhoods are based on prehospital times. An instance would therefore be infeasible even though it could fulfill the coverage targets if the proportion of demand with empty neighborhoods is sufficiently small. We can calculate the proportion u^θ of demand with empty neighborhoods for each patient type as

$$u^\theta = \frac{\sum_{i \in I : N_i^\theta = \emptyset} f_i^\theta}{\sum_{i \in I} f_i^\theta}.$$

The instance can meet the coverage targets at the required service level α if $\alpha(1 - u^\theta) \geq \alpha^\theta$ is fulfilled. The strict spatial coverage assumption can therefore be partially relaxed as long as the above condition holds and the demand of patient type θ at demand node i can be assigned to supply nodes $j \in J$ outside the neighborhood N_i^θ of i, if $N_i^\theta = \emptyset$.

Relaxing the strict spatial coverage assumption can lead to assigning demand to a supply node from outside the supply nodes' neighborhoods. The possible assignment of demand to supply nodes outside the supply nodes' neighborhoods M_j^θ requires adapting the calculation of station-specific capacity limits L_j by altering the calculation of λ_j^{max} to:

$$\lambda_j^{max} = \sum_{\theta \in \Theta} \sum_{i \in M_j^\theta} \frac{1}{24} f_i^\theta \hat{t}_{ij}^\theta + \sum_{\theta \in \Theta} \sum_{i \in I : N_i^\theta = \emptyset} \frac{1}{24} f_i^\theta \hat{t}_{ij}^\theta$$

This adds the total demand with empty neighborhoods to the calculation of the maximum call rate assignable to supply node j. We can then calculate the station-specific capacity limits L_j as in the QRUBUL-HR with:

$$L_j = \min \left\{ \min \left\{ k \in \mathbb{N} \middle| \lambda_j^{max} \leq \Lambda_k^\alpha \right\}, L \right\}$$

The QRUBUL-HR-HP, which is used in the subsequent computational experiments, is formulated as:

$$\begin{aligned}
\min_{x,p} \quad & \beta \sum_{j \in J} \sum_{k=1}^{L_j} k x_{jk} + \gamma \sum_{j \in J} \sum_{k=1}^{L_j} x_{jk} && && (10.2.2a) \\
\text{s.t.} \quad & \sum_{\theta \in \Theta} \sum_{i \in I} \frac{1}{24} f_i^\theta \hat{t}_{ij}^\theta p_{ij}^\theta \leq \sum_{k=1}^{L_j} \Lambda_k^\alpha x_{jk}, && \forall j \in J, && (10.2.2b) \\
& \sum_{j \in N_i^\theta} p_{ij}^\theta = 1, && \forall i \in I,\, \theta \in \Theta : N_i^\theta \neq \emptyset, && (10.2.2c) \\
& \sum_{j \in J} p_{ij}^\theta = 1, && \forall i \in I,\, \theta \in \Theta : N_i^\theta = \emptyset, && (10.2.2d) \\
& \sum_{k=1}^{L_j} x_{jk} \leq 1, && \forall j \in J, && (10.2.2e) \\
& 0 \leq p_{ij}^\theta \leq 1, && \forall i \in I,\, j \in J,\, \theta \in \Theta, && (10.2.2f) \\
& x_{jk} \in \{0,1\}, && \forall j \in J,\, k \in \{1, \dots, L_j\} && (10.2.2g)
\end{aligned}$$

where Λ_k^α is calculated as

$$\Lambda_k^\alpha = \max \left\{ \lambda \in \mathbb{R} \middle| \frac{e^{-\lambda} \frac{\lambda^k}{k!}}{\sum_{c=0}^k e^{-\lambda} \frac{\lambda^c}{c!}} \leq 1 - \alpha \right\} \qquad (10.2.3)$$

The objective function 10.2.2a minimizes the weighted sum of the number of allocated ambulances, weighted by β, and the number of opened stations, weighted by γ. Constraint 10.2.2b ensures that at least one ambulance is available with a service level of at least α when a call arrives at a station. The full demand of patient type θ at each demand node i with a non-empty neighborhood N_i^θ needs to be assigned to a neighboring station j which is modeled by constraint 10.2.2c. If the neighborhood N_i^θ of demand node i for patient type θ is empty, the demand can be assigned to any supply node j as modeled by constraint 10.2.2d. Constraint 10.2.2e ensures that each supply node j can be allocated exactly one number k of ambulances or no ambulances. The domains of the decision variables are defined in constraints 10.2.2f and 10.2.2g.

10.2.3.2 Formulating an Alternative Model to Exploit the Spatial Coverage Dimension

The core QRUBUL-HR-HP formulation can lead to calls of patient types θ being covered with a service level α higher than their required service level α^θ. This is due to requiring strict spatial coverage, if the neighborhood N_i^θ is non-empty. We can relax the strict spatial coverage requirement for patient types $\theta \in \Theta : \alpha^\theta < \alpha^1$. This can enable finding solutions with fewer resources, as it increases the possibilities for assigning demand.
The first step in relaxing the strict spatial coverage requirement is calculating the minimum proportion l^θ of demand of patient type θ which needs to be assigned to supply nodes j inside the neighborhood N_i of demand nodes i as $l^\theta = \min\{l \in \mathbb{R} | l\alpha \geq \alpha^\theta\}$. Constraints 10.2.2c and 10.2.2d in the core QRUBUL-HR-HP can then be replaced by the following two constraints:

$$\sum_{j \in J} p_{ij}^\theta = 1 \qquad \forall i \in I, \theta \in \Theta \tag{10.2.4}$$

$$\sum_{j \in N_i} p_{ij}^\theta \geq l^\theta \qquad \forall i \in I, \theta \in \Theta \tag{10.2.5}$$

Changing the assignment logic additionally requires changing how the maximum assignable call rate λ_j^{max} is calculated to:

$$\lambda_j^{max} = \sum_{\theta \in \Theta} \sum_{i \in M_j^\theta} \frac{1}{24} f_i^\theta \hat{t}_{ij}^\theta + \sum_{\theta \in \Theta} \sum_{i \in I \setminus M_j^\theta} (1 - l^\theta) \frac{1}{24} f_i^\theta \hat{t}_{ij}^\theta$$

The QRUBUL-HR-HP formulation is based on the already strong assumption that the dispatching decision in EMS operations can distinguish between patient types and dispatch ambulances from stations in the respective neighborhoods, and not necessarily always the closest available ambulance. The alternative formulation, as shown above, additionally assumes that ambulances will be actively dispatched from outside the neighborhood in order to fulfill the demand assignments. This assumption was deemed too unrealistic to merit the potential gain in allocated resources to further pursue this model variant.

10.2.3.3 Formulating an Alternative Model to Exploit the Temporal Coverage Dimension

The temporal dimension of coverage is determined by the availability of ambulances at stations. The model considers this by calculating the upper bounds for assignable call rates Λ_k^α. The upper bound Λ_k^α for the call rate assignable to supply node j needs to be calculated on the highest service level α of the patient types θ of calls assigned to j. If only calls of patient types θ with $\alpha^\theta < \alpha^1$ are assigned to a supply node j, more calls could be assigned to j than in the core formulation of QRUBUL-HR-HP.

We will go through the necessary adaptations of the core model formulation step by step in the following, and then present the complete alternative formulation.

The following formulation is based on a strict spatial coverage for all patient types $\theta \in \Theta$, which means that constraint 10.2.2d needs to be removed from the core model formulation, and the following condition has to hold for feasibility: $N_i^\theta \neq \emptyset \, \forall i \in I, \theta \in \Theta$.

We introduce the binary variable z_{jk}^θ to capture, if calls of type θ are assigned to supply node j and if k ambulances are allocated to j at the same time. z_{jk}^θ equals 1 if that is the case and 0 otherwise. This allows adapting constraint 10.2.2b to:

$$\sum_{\theta \in \Theta} \sum_{i \in I} \frac{1}{24} f_i^\theta \hat{t}_{ij}^\theta p_{ij}^\theta \leq \sum_{k=1}^{L_j} \Lambda_k^{\alpha^{\theta^\star}} z_{jk}^{\theta^\star} + (1 - z_{jk}^{\theta^\star}) \sum_{\theta \in \Theta} \sum_{i \in I} \frac{1}{24} f_i^\theta \hat{t}_{ij}^\theta \qquad \forall j \in J, \theta^\star \in \Theta \qquad (10.2.6)$$

The second summand on the right-hand side of the constraint ensures that the constraint is satisfied for $z_{jk}^\theta = 0$, and ensures setting the call rate bound based only on patient types of assigned calls.

We introduce the binary variable π_j^θ to capture whether calls of patient type θ are assigned to supply node j. The following constraint ensures that π_j^θ is set to 1, if calls of patient type θ are assigned to supply node j and that it can be set to 0 otherwise:

$$|M_j^\theta| \pi_j^\theta \geq \sum_{i \in M_j^\theta} p_{ij}^\theta \qquad \forall j \in J, \theta \in \Theta \qquad (10.2.7)$$

The following linear constraints connect z_{jk}^θ to the decision variables x_{jk} and π_j^θ:

$$z_{jk}^\theta \geq \pi_j^\theta + x_{jk} - 1 \qquad \forall j \in J, k \in \{1, \ldots, L_j\}, \theta \in \Theta \qquad (10.2.8)$$

$$z_{jk}^\theta \leq \pi_j^\theta \qquad \forall j \in J, k \in \{1, \ldots, L_j\}, \theta \in \Theta \qquad (10.2.9)$$

$$z_{jk}^\theta \leq x_{jk} \qquad \forall j \in J, k \in \{1, \ldots, L_j\}, \theta \in \Theta \qquad (10.2.10)$$

The first constraint 10.2.8 ensures that z_{jk}^θ is set to 1, if calls of patient type θ are assigned to supply node j and k ambulances are allocated to j. The second constraint 10.2.9 ensures that z_{jk}^θ is set to 0, if no calls of patient type θ are assigned to supply node j. The third constraint 10.2.10 ensures that z_{jk}^θ is set to 0, if not k ambulances are allocated to supply node j.

Enforcing a strict spatial coverage enables calculating the maximum assignable call rate λ_j^{max} for each supply node j for the calculation of the capacity limits L_j as follows:

$$\lambda_j^{max} = \sum_{\theta \in \Theta} \sum_{i \in M_j^\theta} \frac{1}{24} f_i^\theta \hat{t}_{ij}^\theta$$

The station-specific capacities L_j can then be calculated as:

$$L_j = \min \left\{ \min \left\{ k \in \mathbb{N} \middle| \lambda_j^{max} \leq \Lambda_k^{\alpha^1} \right\}, L \right\}, \quad \forall j \in J$$

The complete variant of the QRUBUL-HR-HP which exploits differentiated service levels along the temporal coverage dimension is formulated as:

$$\min_{x,p,z,\pi} \quad \beta \sum_{j \in J} \sum_{k=1}^{L_j} k x_{jk} + \gamma \sum_{j \in J} \sum_{k=1}^{L_j} x_{jk} \tag{10.2.11a}$$

$$\text{s.t.} \quad \sum_{\theta \in \Theta} \sum_{i \in M_j^\theta} \frac{1}{24} f_i^\theta \hat{t}_{ij}^\theta p_{ij}^\theta \leq \sum_{k=1}^{L_j} \Lambda_k^{\alpha^{\theta^\star}} z_{jk}^{\theta^\star} + (1 - z_{jk}^{\theta^\star}) D_j, \quad \forall j \in J, \theta^\star \in \Theta, \tag{10.2.11b}$$

$$\sum_{j \in N_i^\theta} p_{ij}^\theta = 1, \quad \forall i \in I,\ \theta \in \Theta, \tag{10.2.11c}$$

$$|M_j^\theta| \pi_j^\theta \geq \sum_{i \in M_j^\theta} p_{ij}^\theta, \quad \forall j \in J, \theta \in \Theta, \tag{10.2.11d}$$

$$z_{jk}^\theta \geq \pi_j^\theta + x_{jk} - 1, \quad \forall j \in J, k \in \{1, \dots, L_j\}, \theta \in \Theta, \tag{10.2.11e}$$

$$z_{jk}^\theta \leq \pi_j^\theta, \quad \forall j \in J, k \in \{1, \dots, L_j\}, \theta \in \Theta, \tag{10.2.11f}$$

$$z_{jk}^\theta \leq x_{jk}, \quad \forall j \in J, k \in \{1, \dots, L_j\}, \theta \in \Theta, \tag{10.2.11g}$$

$$\sum_{k=1}^{L_j} x_{jk} \leq 1, \quad \forall j \in J, \tag{10.2.11h}$$

$$0 \leq p_{ij}^\theta \leq 1, \quad \forall i \in I,\ j \in J,\ \theta \in \Theta, \tag{10.2.11i}$$

$$x_{jk} \in \{0, 1\}, \quad \forall j \in J,\ k \in \{1, \dots, L_j\}, \tag{10.2.11j}$$

$$z_{jk}^\theta \in \{0, 1\}, \quad \forall j \in J, k \in \{1, \dots, L_j\}, \theta \in \Theta, \tag{10.2.11k}$$

$$\pi_j^\theta \in \{0, 1\}, \quad \forall j \in J, \theta \in \theta \tag{10.2.11l}$$

with : $D_j = \sum_{\theta \in \Theta} \sum_{i \in M_j^\theta} \frac{1}{24} f_i^\theta \hat{t}_{ij}^\theta \quad \forall j \in J$

There are two reasons why this variant of the QRUBUL-HR-HP is not explored further in the computational experiments. The first reason is the increased model complexity by adding $\sum_{j \in J} L_j |\Theta| + |J||\Theta|$ binary variables to the model. Preliminary experiments showed that the core formulation could not be solved to optimality for the instances used in the computational experiments with the available resources in terms of machine and solution time. The variant will additionally only affect the results if there are any stations, for which $\pi_j^1 = 0$, which is unlikely given that the patient category with the highest service level also exhibits a low response time threshold in our real-world instances.

10.3 Computational Experiments for Baden-Württemberg

We conduct computational experiments in the following to investigate how the extensions of the QRUBUL perform in real-world instances. We firstly compare the QRUBUL to the QRUBUL-HR to investigate the effect of relaxing the assumption of location-independent service times. Secondly, we compare two variants of the QRUBUL-HR-HP with different patient types to the QRUBUL-HR to investigate the effects of considering heterogeneous patients. The setup of the computational experiments remains in large parts as described in Section 9.3.1. Section 10.3.1 describes deviations and necessary extensions of the experimental setup before we present the results in Section 10.3.2.

10.3.1 Setting up the Experiments

Relaxing the assumption of location-dependent service times necessitates determining the *location-dependent service times* $\hat{t}_{ij}^{\theta}$ for every combination of demand nodes $i \in I$, supply nodes $j \in J$, and patient types $\theta \in \Theta$. We omit the differentiation between patient types for our computational experiments to focus on the impact of the difference in driving times, and set $\hat{t}_{ij}^{\theta} = \hat{t}_{ij} \, \forall \theta \in \Theta$. We follow the approach of Leknes et al. (2017) in calculating the location-dependent service times. Fig. 10.2 depicts the components of the service time from dispatching an ambulance until it arrives back at its station.

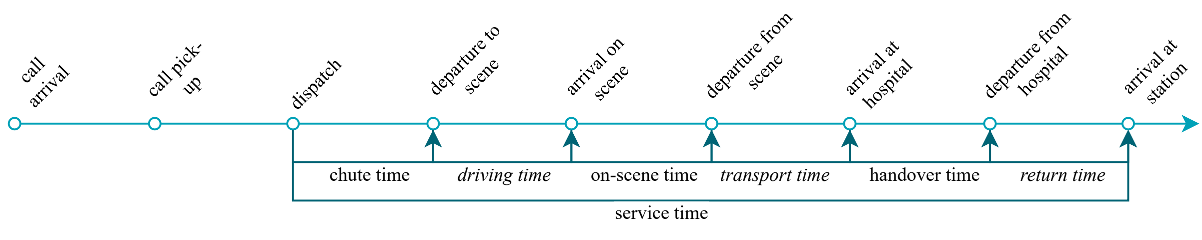


Figure 10.2: Components of the service time.

The *chute time*, *on-scene time*, and *handover time* are assumed to be location-independent and are calculated as the 95th percentile of historical process times to stay consistent with using upper-bound service times. We denote the 95th percentile of the sum of these times as T^S, and the 95th percentile of the sum of *chute time* and *on-scene-time* as $\tilde{T^S}$. The *driving time* T_{ji} from supply node $j \in J$ to demand node $i \in I$, the *transport time* T_{iH_i} from demand node i to its closest hospital H_i, and the *return time* T_{H_ij} from that hospital H_i back to the supply node j are calculated using Openrouteservice[1]. The sum of these components results in the location-dependent service time with transport. Not every call response results in transporting the patient to a hospital. If no transport occurs, the ambulance returns to its station after the *on-scene time*. We assume that the *driving time* between a supply and a demand node is symmetrical, which allows us to calculate the location-dependent service without transport as the sum of *chute time*, *driving time*, *on-scene-time*, and again the *driving time*. The proportion p^H of call responses with transport is based on historical data for each instance. The location-dependent service time is then calculated as the weighted sum of the location-dependent service time with transport and the location-dependent service time without transport:

$$\hat{t}_{ij} = p^H \left(2T_{ji} + T^S + T_{iH_i} + T_{H_ij}\right) + (1 - p^H)\left(2T_{ji} + \tilde{T^S}\right)$$

The QRUBUL is based on a location-independent service time $\hat{t}$. This service time is calculated as the average of the location-dependent service times between all neighboring pairs of demand nodes and supply nodes as $\hat{t} = \frac{1}{|I|}\sum_{i\in I}\frac{1}{|N_i|}\sum_{j\in N_i}\hat{t}_{ij}$. This facilitates attributing differences in the solutions and simulation results of the QRUBUL and QRUBUL-HR to relaxing the location-independent service time assumption.
The QRUBUL-HR-HP considers heterogeneous *patient types* $\theta \in \Theta$ which need to be defined for the computational experiments. We consider the following three sets:

- *undiff*$= \{A\_R\}$
- *rt-2*$= \{A\_R, B\_R\}$
- *rt-1-pt-1*$= \{A\_R, A\_P\}$

The first set *undiff* serves as a benchmark and contains a single patient type. The second set *rt-2* aims to analyze the effect of differentiating between response time targets, and the third set *rt-1-pt-1* aims to analyze the effects of considering prehospital time coverage targets. The set *rt-2* therefore contains two patient types based on response time coverage targets, and the set *rt-1-pt-1* contains one patient type based on a response time coverage and one patient type based on a prehospital time coverage. Table 10.1 lists the three patient types with their coverage threshold and service levels.

Table 10.1: Coverage requirements by patient type

Patient type	Coverage threshold	Coverage time	Service level
A_R	12	response	95%
A_P	60	prehospital	80%
B_R	23	response	80%

[1]https://openrouteservice.org/

The current regulation for EMS in Baden-Württemberg is the basis for the patient types *A_R* and *A_P* as they define two patient categories with the listed coverage thresholds and service levels. The coverage threshold for patient type *B_R* is derived from the utility function of patient category 3 (s. Chapter 5). The threshold is the first integer response time with a utility value above 0.5. The service level is set to the same service level as for the patient type *A_P*.

The *neighborhoods* $N_i^{A\_P}$ for the patient type with a prehospital time-based coverage threshold are determined as follows: For each demand node i, the transport time to its closest hospital is calculated using Openrouteservice. This transport time, the median historic on-scene time, and the median chute time are subtracted from the coverage threshold to result in an individual driving time threshold for each demand node i. Supply nodes j with a driving time to i below this threshold are considered as covering i.

The diagnosis categories developed in Chapter 5 are assigned to the patient types to calculate the patient type-dependent call rates f_i^θ based on historical data. Table 10.2 lists which diagnosis category is assigned to which patient type. The idea for *rt-2* is that the two diagnosis categories 1 and 2, with patients in life-threatening situations, are prioritized in the planning. For *rt-1-pt-1*, the diagnosis categories 2 and 4, whose utility depends on the prehospital time, are assigned to patient type *A_P* while the remaining categories are assigned to patient type *A_R*. Diagnosis category 7 is omitted as it is not possible to distinguish between categories 7 and 8 based on the diagnoses and as a result, no historical calls are classified as category 7 (s. Chapter 5).

Table 10.2: Patient type definitions based on diagnosis categories

Diagnosis category	*rt-2*	*rt-1-pt-1*
1	A_R	A_R
2	A_R	A_P
3	B_R	A_R
4	B_R	A_P
5	B_R	A_R
6	B_R	A_R
8	B_R	A_R

The service times $\hat{t}_{ij}^\theta$ are assumed to be the same for each patient type $\theta \in$ *rt-2* $\cup$ *rt-1-pt-1* to reduce the differences between the QRUBUL-HR and QRUBUL-HR-HP and to facilitate attributing different results to considering differentiated coverage targets alone.

The settings for solving the optimization models and running the simulations are kept as described in Section 9.3.1.

10.3.2 Analyzing the Results

The following sections analyze the results of the computational experiments. We firstly analyze the effects of considering location-dependent service times by comparing the QRUBUL and QRUBUL-HR. Secondly, we analyze the effects of considering heterogeneous patients by comparing two variants of the QRUBUL-HR-HP with different patient types to the variant without patient type differentiation.

10.3.2.1 Effects of Considering Heterogeneous Regions

Effects on Computational Performance

We introduced station-specific capacity bounds L_j for the supply nodes $j \in J$ in the QRUBUL-HR to reduce the number of integer variables in the model. This proved effective in our instances, where the average number of integer variables in all instances decreased by about 70% from the QRUBUL to the QRUBUL-HR. Table 10.3 presents the aggregated computational performances. Surprisingly, the reduction in the number of integer variables did not appear to reduce solution times. The reduction stems from avoiding defining variables that will never be set to 1 in a solution. A possible explanation for not reducing the solution times is that Gurobi Optimizer might be capable of detecting and removing these variables during preprocessing. Overall, the QRUBUL-HR requires an increased computational effort as indicated by fewer instances solved to proven optimality, an increased mean and maximum MIP-Gap, and a higher mean solver time.

Table 10.3: Comparison of aggregated computational performance over all instances between the QRUBUL and QRUBUL-HR.

Model	Instances	Feasible	Optimal	Mean MIP gap of non-optimal feasible solutions (%)	Max MIP gap of non-optimal feasible solutions (%)	Mean solver time (s)	Max solver time (s)
QRUBUL	14	14	10	2.69	3.72	4020.93	10800
QRUBUL-HR	14	14	8	2.74	4.87	5746.93	10801

Effects on Aggregated Optimization and Simulation Results

Considering location-dependent service times can have three effects, which might all be present in the same instance. If location-dependent service times of a station are shorter than the average service times, overcapacities might be avoided, and the number of resources might be reduced. If, on the other hand, location-dependent service times of a station are higher than the average service times, undercapacities might be avoided, and the number of resources might be increased to avoid uncovered calls.

Considering location-dependent service times can also reduce the number of prescribed ambulances through the following effect. The number of prescribed ambulances depends on the average call rates per service time. Reducing the average service times reduces the number of calls per service time and could therefore reduce the number of prescribed ambulances. The average service time of a station depends on both the locations of assigned demand nodes and the locations of their closest hospitals. The station location and demand assignment decisions will therefore impact the average service times. The number of prescribed ambulances could, as a consequence, be reduced if the decisions are taken such that the average service time is reduced.

Table 10.4 lists the mean simulated coverages, mean number of prescribed stations, and mean number of prescribed ambulances in the fourteen instances. Considering location-dependent service times slightly increases the simulated coverage and slightly reduces the prescribed resources. These results indicate that considering heterogeneous service times can increase coverage and reduce prescribed resources.

Table 10.4: Average results of experiments with QRUBUL and QRUBUL-HR.

Model	Mean coverage (%)	Mean stations	Mean ambulances
QRUBUL	91.81	34.07	62.5
QRUBUL-HR	92.03	34	62.14

Effects on Service Times

We further investigate whether the average service times of stations are reduced by considering location-dependent service times. To this end, we calculate the total call rates per service time in the solution of each instance. For the QRUBUL-HR this is equivalent to $\sum_{j\in J}\sum_{k=1}^{L_j}\sum_{i\in M_j}\frac{1}{24}f_i\hat{t}_{ij}p_{ij}x_{jk}$ with the values of the decision variables in the solution. For the QRUBUL we calculate $\sum_{j\in J}\sum_{k=1}^{L_j}\sum_{i\in M_j}\frac{1}{24}f_i\hat{t}p_{ij}x_{jk}$. Table 10.5 lists the call rates per service time as well as the relative reductions.

Table 10.5: Comparison of call rates per service time in QRUBUL-HR and QRUBUL solutions.

ID	QRUBUL-HR	QRUBUL	Difference (%)
1	11.54	12.25	5.75
2	16.77	17.2	2.48
3	6.57	6.74	2.4
4	8.93	9.15	2.41
5	10.72	11.03	2.75
6	8.71	8.86	1.72
7	6.57	6.85	4.02
8	24.34	25.4	4.16
9	16.46	17.49	5.93
10	25.95	26.92	3.61
11	16.79	17.72	5.24
12	38.9	40.06	2.89
13	33.61	36.09	6.86
14	68.29	72.29	5.53

The call rates per service time in the QRUBUL-HR solution are lower than in the QRUBUL solution in every instance. This observation supports the supposition that station locations are chosen such that the average service times are reduced, which can reduce the number of prescribed ambulances.

Effects per Instance

Next, we will analyze the changes in prescribed resources and simulated coverages for each instance to investigate how consistent the effects are for our instances. Fig. 10.3 presents an aggregated comparison of the relative change in objective function values and simulated coverages of the QRUBUL-HR compared to the QRUBUL, and Table 10.6 lists the numeric results.

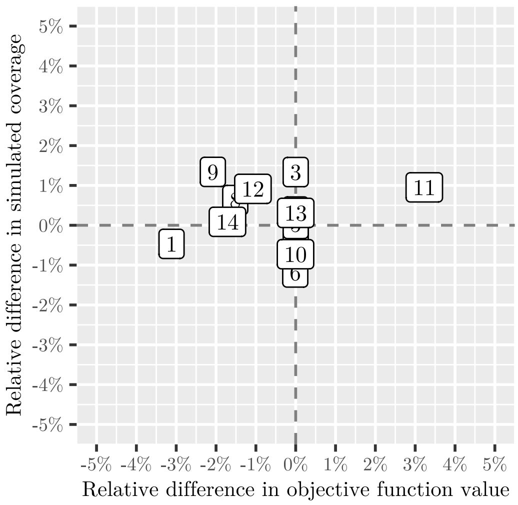


Figure 10.3: Comparison of simulated coverages and objective function values of the QRUBUL-HR compared to the QRUBUL. Numbers of labels indicate the instance ID (s. Table 9.4)

Table 10.6: Comparison of prescribed resources and simulated coverages between the QRUBUL and QRUBUL-HR

	Opened Stations		Ambulances		Driving time coverage		$\Delta^{max}_{driving\_coverage}$	
ID	QRUBUL	QRUBUL-HR	QRUBUL	QRUBUL-HR	QRUBUL	QRUBUL-HR	QRUBUL	QRUBUL-HR
1	**16**	**16**	32	**31**	93.42%	**92.98%**	0%pt.	**-0.44%pt.**
2	**18**	**18**	**40**	**40**	93.96%	**93.7%**	0%pt.	**-0.26%pt.**
3	**17**	**17**	**28**	**28**	**89.21%**	90.41%	**-1.19%pt.**	0%pt.
4	**22**	**22**	**35**	**35**	**92.84%**	92.91%	**-0.07%pt.**	0%pt.
5	**16**	**16**	**32**	**32**	**90.65%**	90.68%	**-0.03%pt.**	0%pt.
6	21	**20**	**33**	**33**	89.49%	**88.43%**	0%pt.	**-1.06%pt.**
7	**25**	**25**	**36**	**36**	**95.17%**	95.5%	**-0.33%pt.**	0%pt.
8	34	**33**	66	**65**	**88.38%**	88.94%	**-0.56%pt.**	0%pt.
9	**24**	**24**	48	**47**	**90.24%**	91.45%	**-1.21%pt.**	0%pt.
10	29	**28**	**61**	**61**	90.33%	**89.68%**	0%pt.	**-0.66%pt.**
11	**42**	44	**62**	64	**95.33%**	96.24%	**-0.91%pt.**	0%pt.
12	**44**	**44**	93	**92**	**90.14%**	90.97%	**-0.82%pt.**	0%pt.
13	**86**	**86**	**134**	**134**	**96.24%**	96.54%	**-0.3%pt.**	0%pt.
14	**83**	**83**	175	**172**	**89.93%**	90%	**-0.08%pt.**	0%pt.

The differences in objective function values and coverages are within the low single-digit percentages. The QRUBUL-HR does not lead to a lower coverage with more prescribed resources for any instance. The simulated coverage increases in ten out of the fourteen instances, and decreases in the remaining four. The effect on the number of resources depends on the instance. Fewer resources are prescribed in seven instances, equal resources are prescribed in six instances, and more resources are prescribed in one instance.

Instances 1, 2, 6, and 10 exhibit a reduced simulated coverage for the QRUBUL-HR solution. One fewer ambulance is prescribed in instance 1, the same number of ambulances and stations are prescribed in instance 2, and one fewer station is prescribed in both instances 6 and 10. The reduction in coverage can therefore not be explained by a systematic reduction in resources, especially since there are instances with reduced resources and an increased simulated coverage. The decrease in simulated coverage could be due to existing model assumptions, such as the binary coverage, or due to the differences between demand node locations and actual call locations.

Instance 11 is the only instance for which the QRUBUL-HR prescribes more resources. The instance was solved to proven optimality by the QRUBUL, but not by the QRUBUL-HR, which might exaggerate the difference in prescribed resources. The lower bound of the QRUBUL-HR (62.41) is, however, larger than the optimal objective function value of the QRUBUL (62.11). For the other instances that were not solved to optimality by both models, the best bound of the QRUBUL-HR solution is lower than the objective function value of the QRUBUL solution. Instance 11 does not exhibit any noticeable characteristics that serve to explain the increase in prescribed resources, and hints at potential for further research into the connection between instance characteristics and model behavior.

Conclusion of Considering Heterogeneous Regions

Overall, considering location-dependent service times slightly improves the simulated coverage with fewer resources. The magnitude of the improvements is small, and the effects vary between instances. We expect the benefits of considering location-dependent service times to increase when considering heterogeneous patients due to the increase in the variance of driving times, and therefore deem it reasonable to consider location-dependent service times in the QRUBUL-HR-HP.

10.3.2.2 Effects of Considering Heterogeneous Patients

The QRUBUL and QRUBUL-HR are steps in developing the QRUBUL-HR-HP. Both models aimed at lifting limitations of previous models that we deem problematic when considering heterogeneous patients. The analyses, therefore, focused on whether lifting the limitations improved the solutions of previous models.

The QRUBUL-HR-HP can model arbitrary sets of patient types defined by coverage definitions and service levels. The main focus of the following analysis lies on how considering more than one patient type impacts computational performance and prescribed resource levels. Considering differentiated coverage thresholds in the ambulance location problem implies that the operational dispatching decision can distinguish between those patient types and can assign incoming calls in similar proportions to the demand assignment proportions prescribed by the QRUBUL-HR-HP. We simulate the model solutions with the current closest available dispatching strategy to provide an indication of how crucial differentiated dispatching strategies are when basing the station location and sizing decisions on differentiated patient types.

We conduct experiments with three model variants of the QRUBUL-HR-HP. The variants are based on the three sets *undiff*, *rt-2*, and *rt-1-pt-1*, defined in Section 10.3.1, and will be referred to by those set names in the following. The *undiff* variant considers a single patient type, and is equivalent to the QRUBUL-HR. The *rt-2* variant considers two patient types, *A_R* and *B_R*, based on response time targets. The *rt-1-pt-1* variant considers two patient types, *A_R* and *A_P*, one based on a response time target, and the other on a prehospital time target.

Effects on Computational Performance

Table 10.7 lists the aggregated results of the computational performances of the three model variants, and Table 10.8 contains the detailed results for each instance.

Table 10.7: Comparison of aggregated computational performances over all instances of the QRUBUL-HR-HP variants.

Variant	Instances	Feasible	Optimal	Mean MIP gap of non-optimal feasible solutions (%)	Max MIP gap of non-optimal feasible solutions (%)	Mean solver time (s)	Max solver time (s)
undiff	14	14	8	2.74	4.87	5746.93	10801
rt-2	14	14	3	3.57	6.91	9913.21	10814
rt-1-pt-1	14	14	2	2.27	5.26	9356.71	10808

Table 10.8: Summary of optimization results per instance for QRUBUL-HR-HP

	Opened Stations			Ambulances			Solver Time (s)			MIP Gap (%)		
ID	undiff	rt-2	rt-1-pt-1	undiff	rt-2	rt-1-pt-1	undiff	rt-2	rt-1-pt-1	undiff	rt-2	rt-1-pt-1
1	**16**	**16**	**16**	31	**26**	30	**19**	2297	10800	0	0	2.28
2	**18**	19	19	40	**36**	39	**161**	10800	10800	0	4.39	**1.3**
3	**17**	**17**	20	28	**24**	28	**16**	10144	905	0	0	0
4	**22**	**22**	23	35	**30**	32	8404	10800	**475**	0	1.08	0
5	**16**	**16**	17	32	**27**	30	**324**	10800	10800	0	2.9	**2.5**
6	**20**	**20**	24	33	**28**	34	**1195**	10800	10800	0	2.77	**1.23**
7	25	**24**	36	36	**30**	41	10800	**7530**	10800	1.24	0	**1.07**
8	33	**32**	34	65	**55**	60	**10800**	**10800**	**10800**	**0.45**	3.37	0.51
9	**24**	**24**	28	47	**40**	49	**2792**	10800	10800	0	3.33	**1.72**
10	**28**	**28**	30	61	**53**	59	**2743**	10800	10800	0	3.72	**2.2**
11	44	**40**	44	64	**54**	60	**10800**	**10800**	**10800**	2.65	**1.91**	2.73
12	**44**	45	**44**	92	**80**	88	10801	**10800**	**10800**	**1.59**	5.58	2.5
13	86	**83**	99	134	**112**	133	10801	**10800**	10806	4.87	**3.3**	3.98
14	**83**	85	93	172	**150**	171	**10801**	10814	10808	**3.32**	6.91	5.26

Considering multiple patient types decreases the computational performance, likely due to the increase in the number of decision variables. While the *undiff* variant could solve 8

out of the 14 instances to optimality within the solution time limit, the variants *rt-2* and *rt-1-pt-1* could only solve 3 and 2 instances to optimality, respectively. The maximum MIP-Gap likewise increased, as did the mean solver times.
The number of considered patient types should therefore be chosen with consideration, especially when solving real-world-sized instances, as the number of continuous variables grows proportionally with the number of patient types.
The increased MIP-Gaps and decreased number of optimally solved instances need to be considered when interpreting the following analyses of objective function values.

Effects on Prescribed Resources

Fig. 10.4 depicts the minimum relative change in objective function values of the variants *rt-2* and *rt-1-pt-1* compared to the *undiff* variant. To strengthen the conclusions drawn from this comparison, we consider both the best bound and the best objective function value to calculate the relative differences and calculate the minimum relative difference as follows: If the best objective function value for an instance of the *rt-2* or *rt-1-pt-1* variant is lower than the best bound of the *undiff* variant, we calculate the minium relative difference between the best objective function value of the *rt-2* or *rt-1-pt-1* variant and the best bound of the *undiff* variant. If the best bound of the *rt-2* or *rt-1-pt-1* is higher than the best objective function value of the *undiff* we calculate the minimum relative difference based on the best bound of the *rt-2* or *rt-1-pt-1* and the best objective function value of the *undiff*. If neither is the case, the minimum relative difference is set to 0.

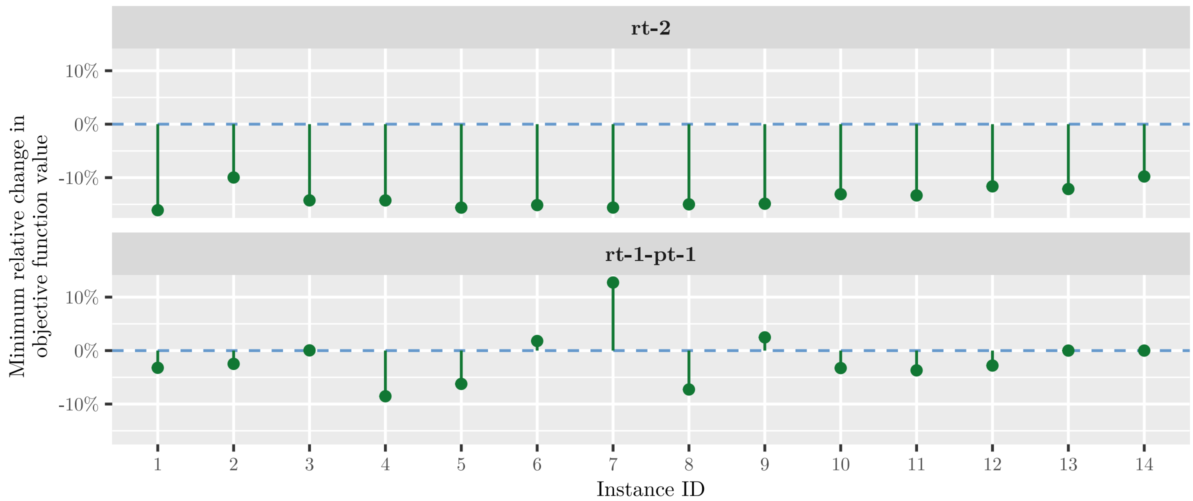


Figure 10.4: Minimum relative change of the objective function value of the QRUBUL-HR-HP variants compared to the QRUBUL-HR solution.

Introducing a second patient type with a larger response time-based threshold in the *rt-2* variant decreases the prescribed resources in all instances. This is at least partly a consequence of calculating station availabilities with the truncated Poisson distribution, where the relationship between the number of ambulances at a station and the maximum call rate assignable to that station for a given service level is non-linear, and doubling the number of ambulances more than doubles the maximum assignable call rate. The increased coverage threshold allows assigning more demand to the same station, which leads to a decrease in the prescribed resources. This effect also influences the distribution of station sizes depicted in Fig. 10.5. The *rt-2* variant prescribes more stations with 5, 6, or 7 ambulances, and the proportion of small stations with only 1 ambulance also increases.

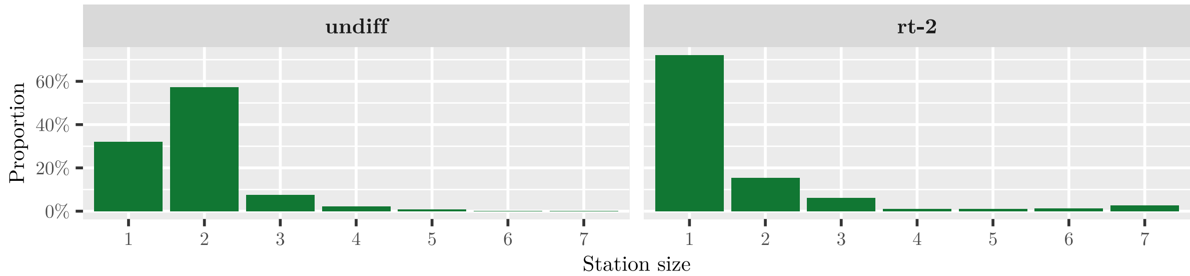


Figure 10.5: Distribution of station sizes for the QRUBUL-HR-HP *rt-2* variant compared to the *undiff* variant.

Introducing a second patient type based on a prehospital time coverage does not decrease the number of prescribed resources in all instances. We see an increase in resources in instances 6, 7, and 9. An increase in the prescribed resources occurs if a sufficiently large proportion of demand nodes $i \in I$ has fewer supply nodes in their prehospital time-based neighborhood $N_i^{A\_P}$ compared to their response time neighborhood $N_i^{A\_R}$. The farther a hospital is located away from a demand node, the less driving time from the supply node to the demand node remains to stay within the prehospital time limit.

Fig. 10.6 depicts the proportion of demand nodes i of each instance, for which $|N_i^{A\_P}| < |N_i^{A\_R}|$ holds. We see that the instances 6, 7, and 9 are among the instances with the highest proportions of demand nodes with $|N_i^{A\_P}| < |N_i^{A\_R}|$. Instance 13 also exhibits a large proportion, but the relatively high MIP Gap does not permit a conclusive statement regarding the change in prescribed resources.

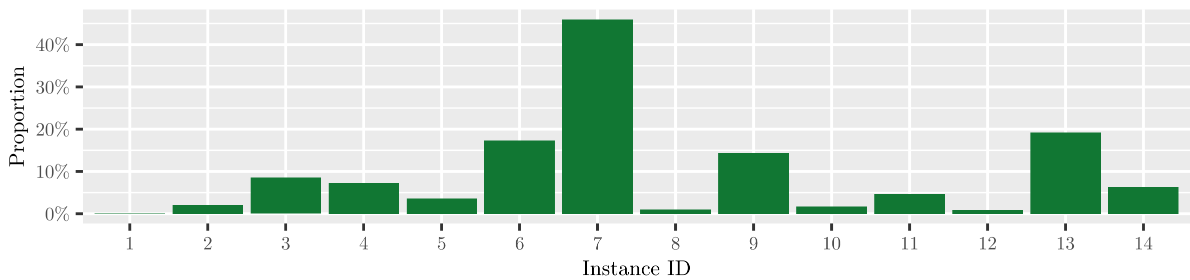


Figure 10.6: Proportion of demand nodes of *rt-1-pt-1* variant of the QRUBUL-HR-HP with a shorter prehospital time than response time based neighborhood.

Effects on Simulated Coverages

Next, we will analyze the results of simulating the solutions of the three QRUBUL-HR-HP variants with the closest available dispatching strategy. The closest available dispatching strategy will always dispatch the closest available ambulance, while the QRUBUL-HR-HP assumes that some demand can be assigned to stations that are not the closest one. This might have the effect of reducing the ambulance availability at some stations and therefore reducing the coverage in the simulation. Fig. 10.7 depicts how the simulated coverage for the two patient types A_R and B_R changes in the *rt-2* variant compared to the *undiff* variant.

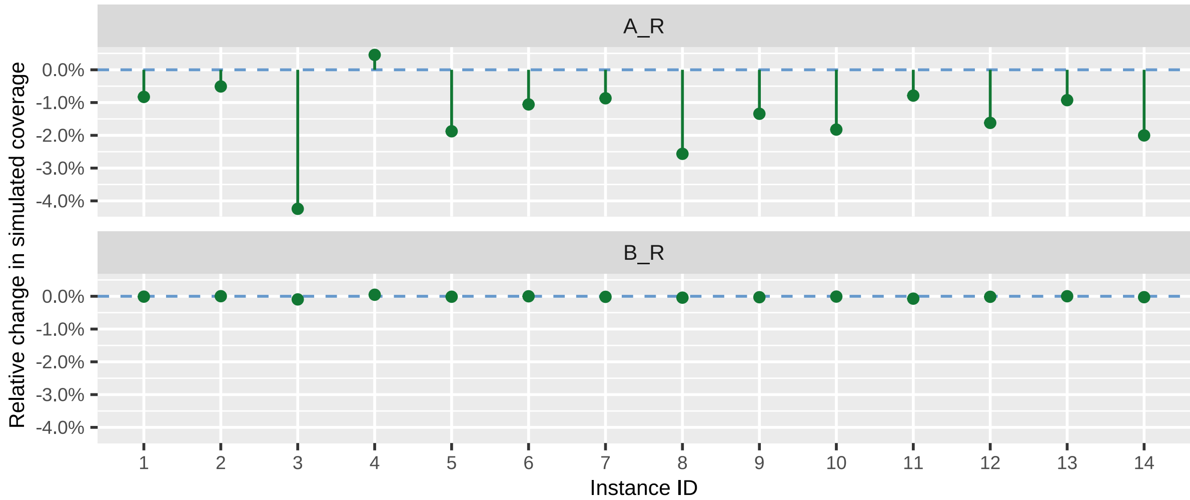


Figure 10.7: Relative change of simulated coverage for the patient categories of the QRUBUL-HR-HP *rt-2* variant compared to the *undiff* variant.

We can observe that the coverage for patients of type B_R changes only very slightly in some instances. This indicates that the QRUBUL-HR-HP prescribes sufficient resources within the larger response time coverage threshold such that the availability of ambulances does not change for B_R patients.

The coverage for A_R patients, on the other hand, decreases in all instances but one. This is in line with the expectation that the closest available dispatching strategy will decrease the availability of ambulances at some stations since it does not differentiate between patient types and can therefore not adhere to the demand assignment assumed in the QRUBUL-HR-HP solution. The closest available dispatching strategy, therefore, does not appear fully compatible with the station locations and sizes, which have been based on differentiated patient types.

An unexpected result is the increase in coverage for both patient types in instance 4, considering that the same number of stations, but five fewer ambulances are prescribed (s. Table 10.8). Two possible causes might lead to this result. The parameter choices, especially basing the service times on 95th percentiles of historic data, might be conservative and lead to prescribing more ambulances than necessary in this instance. A second possible cause is that even if two station location configurations both cover all demand nodes,

they might still lead to different coverages in the simulation due to using binary coverage thresholds and deterministic travel times, or due to the differences between demand node and actual call locations.
Fig. 10.8 depicts the changes in simulated coverage for the two patient types *A_P* and *A_R* when comparing the solution of the *rt-1-pt-1* to the solution of the *undiff* variant.

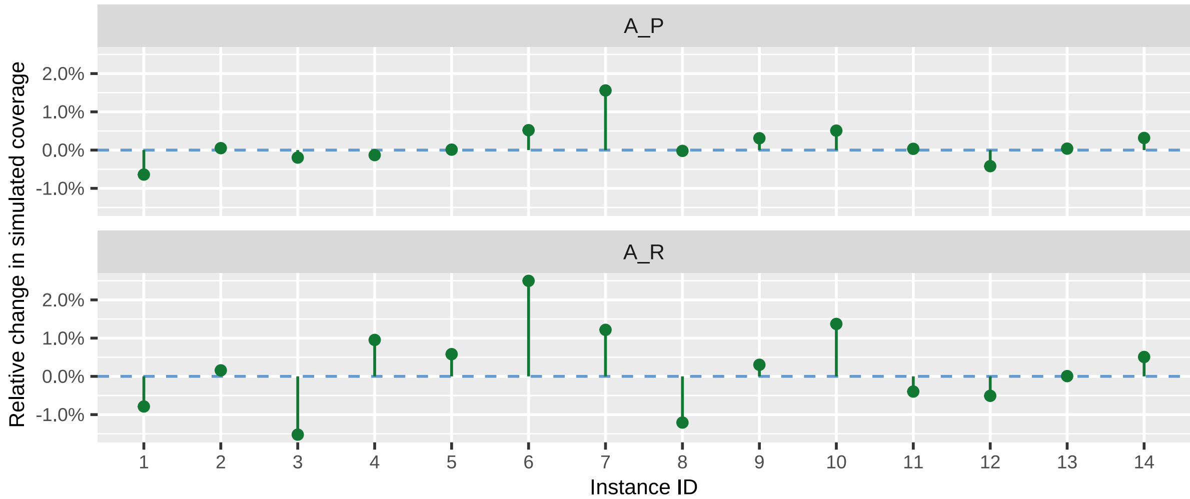


Figure 10.8: Relative change of simulated coverage for the patient categories of the QRUBUL-HR-HP *rt-1-pt-1* variant compared to the *undiff* variant.

The QRUBUL-HR-HP prescribes more resources for instances 6, 7, and 9 in the *rt-1-pt-1* variant compared to the *undiff* variant. In all of these instances, the prehospital time-based coverage of patients of type *A_P* increases in the simulation. The QRUBUL-HR-HP therefore appears able to consider prehospital time-based coverage targets by deriving individual driving time-based neighborhoods for demand nodes.
The coverage for *A_R* patients also increases for the instances 6, 7, and 9, likely as a result of the additional resources. The coverage for *A_R* patients changes in other instances as well, with both increases and decreases. An equal number of ambulances is prescribed in instance 3, and fewer ambulances in all other instances. On the other hand, the number of stations increases in all instances except instances 1, 11, and 12. No clear correlation between the change in ambulances or stations and the simulated coverage is apparent, which makes it likely that the change in coverage is either due to the closest available dispatching strategy or the parameter choices.

Conclusion of Considering Heterogeneous Patients

To conclude, considering multiple patient types increases the computational complexity. Considering a second patient type based on a larger response time threshold reduces the number of prescribed resources in all instances. The simulated coverage, however, decreases for the patients with the smaller response time threshold when dispatching by the closest available dispatching strategy. Considering a second patient type based on a prehospital time threshold has instance-dependent effects on the number of prescribed resources. Depending on the hospital locations, the number of prescribed resources either increases or decreases. The results additionally indicate that the QRUBUL-HR-HP does not yet model all factors that influence the simulated coverage.

10.4 Extension to Heterogeneous Regions and Patients: Discussion

This chapter firstly aimed to relax the assumption of location-independent service times in the QRUBUL. Secondly, it aimed to consider heterogeneous patients in a set-covering approach to the ambulance location problem. In this context, patients are differentiated by their coverage definition and service level.

Contributions

To these ends, two novel models are proposed. The QRUBUL-HR relaxes the assumption of location-independent service times and presents, to the best of our knowledge, the first set covering model for the ambulance location problem which integrates location-dependent service times.

We developed the QRUBUL-HR-HP to consider heterogeneous patients. To the best of our knowledge, the QRUBUL-HR-HP is the first set-covering model for the ambulance location problem that considers heterogeneous patients. The model is able to handle arbitrary patient types defined by how the neighborhood of each demand node is defined and by the service level.

Next to the core QRUBUL-HR-HP formulation, we proposed two alternative formulations. One formulation exploits the heterogeneity of patient types along the spatial coverage dimension, and the second formulation exploits the heterogeneity of patient types along the temporal dimension. The alternative model formulations were not pursued further in computational experiments due to strong assumptions and increased model size.

We validated both the QRUBUL-HR and the QRUBUL-HR-HP through computational experiments based on real-world data. The experiments demonstrated that the QRUBUL-HR can avoid over- and undercapacities created by the location-independent service time assumption in the majority of instances. Differentiating between two response time-based patient types in the QRUBUL-HR-HP led to a decrease in prescribed resources compared to planning only based on the patient type with a stricter coverage threshold and service level. The effects of considering a prehospital time-based coverage definition depend on the instance, specifically on the hospital locations. The farther the hospitals are located away from demand nodes, the shorter the remaining driving time from supply nodes to demand nodes, and the more likely that additional resources are needed. The simulation experiments demonstrated that the closest available dispatching strategy might not be compatible with station locations and sizes that are based on heterogeneous patients.

Practical Implications

The current regulations for EMS in Baden-Württemberg contain planning targets for heterogeneous patient types, and the QRUBUL-HR-HP enables planning station locations and their sizes based on heterogeneous patient types with the goal of identifying the minimum number of resources necessary to fulfill regulatory requirements. By considering arbitrary patient types, the model can also be used to evaluate the effects of different patient categorizations before they are formally defined in a regulation.

Considering location-dependent service times in the QRUBUL-HR and QRUBUL-HR-HP enables evaluating the impact of changes in hospital locations on the resources required for EMS. The German hospital landscape is currently undergoing a reform with a trend to reduce the number of hospitals. Fewer hospitals potentially lead to longer driving times from demand nodes to hospitals, which increases service times. This impacts firstly how well prehospital time-based coverages can be fulfilled and secondly how many calls can be served by an ambulance. The QRUBUL-HR-HP enables evaluating these impacts and can support in adapting the EMS to changes in hospital locations.

Limitations

The QRUBUL-HR-HP relaxes assumptions of the QRUBUL, but is still based on limiting assumptions. Several of those remain unchanged from the QRUBUL and have been discussed in the previous chapter in Section 9.4. We will therefore touch only briefly on these unchanged limitations. The QRUBUL-HR-HP assigns demand from demand nodes to supply nodes and therefore assumes that the dispatching decision can follow a similar demand assignment. The availability of stations is assumed to be independent of other stations, and calls arriving when no ambulance is available at a station are assumed lost. We further assume that ambulances are dispatched only from their station, and the QRUBUL-HR-HP models both driving times and call rates as deterministic parameters.

Distinguishing between patient types in the QRUBUL-HR-HP introduces the additional assumption that it is possible to distinguish between patient types in the dispatching decision. It is unlikely that dispatchers can distinguish between patient types with perfect accuracy based on the limited information available during an emergency call.

The calculation of location-dependent service times in the QRUBUL-HR and QRUBUL-HR-HP assumes that patients are transported to the closest hospital. The historical call data show that this is not always the case. Possible reasons for not choosing the closest hospital are overcrowding situations in emergency departments and hospitals, specific treatment requirements that are not available in every hospital, or patient preferences.

Methodological Reflections

We made multiple methodological choices in developing the QRUBUL-HR-HP and in performing the computational experiments. Similar to the limiting assumptions, several methodological choices remain unchanged from the development of the QRUBUL. These choices are discussed in Section 9.4 and will only be listed briefly in the following. The calculation of ambulance availabilities at a station remains based on the dependence assumption of ambulances. We continue employing the weighted-sum approach for handling the two objectives of minimizing the number of stations and the number of ambulances. The setup of the computational experiments remains largely consistent with the previous chapter. This includes the definition of demand node locations, calculating driving times using Openrouteservice, calculating call rates based on historic daily averages, and using the 95th percentile of historic process times in calculating the service times. The optimization models are again not necessarily solved to optimality due to setting a time limit for the calculations. All computational experiments are based on the EMS system in Baden-Württemberg, which limits the transferability of the findings to other EMS systems. Lastly, the simulation remains based on the same data used to parametrize the instances for the optimization models, and we assume a 24/7 allocation of ambulances.

Considering differentiated patient types requires defining them for the computational experiments. We defined three patient types, partly based on current regulations in Baden-Württemberg and partly based on the patient categories and utility functions developed in Part II. We then combined two patient types each into two sets of patient types for the computational experiments to investigate the impacts of firstly differentiating between response time-based coverages and secondly of considering prehospital time-based coverages. An alternative would have been to base all patient types on the patient categories developed in Part II and derive coverage thresholds from the utility functions. Similarly, sets with more than two patient types might be interesting to investigate, even though this would likely further decrease the computational performance due to the additional variables.

Outlook

The presented work offers several avenues for further research. The call rates per patient types are modeled as deterministic parameters in the QRUBUL-HR-HP. It is unlikely that EMS can distinguish between patient types without misclassifications during operations in the dispatching decision. This opens up the avenue of investigating how differing classification accuracies can be considered by the QRUBUL-HR-HP.

The travel times and call rates are deterministic parameters in the QRUBUL-HR-HP. The historical data show that these parameters vary in practice, which leads to the question of how they could be modeled as stochastic parameters and whether a stochastic modeling approach improves the solutions.

The experiments demonstrated that the closest available dispatching strategy might not work well in an EMS system designed for heterogeneous patients. Investigating how the alternative dispatching strategies investigated in Part IV perform with station locations and sizes prescribed by the QRUBUL-HR-HP would therefore be another interesting avenue for further research.

The model size increases with additional patient types, reducing the computational performance. Investigating custom solution procedures or heuristic approaches could therefore enable efficient solving of larger instances of the QRUBUL-HR-HP.
The QRUBUL-HR-HP models the strategic problem of station locations and sizes. Adapting the core ideas of station capacities based on location-dependent service times and heterogeneous patients to the tactical problem of allocating ambulances to stations for different shifts appears as a promising topic for further research.
Finally, a logical next step is seen in extending the QRUBUL-HR-HP to integrate the objective of maximizing a performance criterion, such as patient utility, with the number of resources prescribed by the current version of the QRUBUL-HR-HP while maintaining coverage targets. Maximizing a performance criterion additionally presents opportunities for investigating trade-offs between efficiency and fairness by using different aggregations of the performances provided for each demand node by a given solution.

10.5 Extension to Heterogeneous Regions and Patients: Conclusion

This chapter aimed to relax the assumption of location-independent service times and consider heterogeneous patients in a set-covering approach to the ambulance location problem. We developed two models to address these objectives and evaluated them in computational experiments based on real-world data from the EMS in Baden-Württemberg.
The QRUBUL-HR extends the QRUBUL to integrate location-dependent service times into a set-covering framework. Computational experiments demonstrate that this extension avoids over- and undercapacities in the majority of instances compared to the base QRUBUL. Considering location-dependent service times additionally opens up the possibility to investigate the impacts of changes in hospital locations on resource requirements in EMS due to increased transport times.
The QRUBUL-HR-HP enables considering arbitrary sets of patient types defined by coverage thresholds and service levels. The model supports both response time and prehospital time-based coverage definitions. Our experiments show that differentiating between patient types can reduce prescribed resources when planning based on multiple coverage thresholds rather than only the strictest requirement.
The QRUBUL-HR-HP presents a first step in a planning approach based on heterogeneous patients, as required by current regulations for EMS in Baden-Württemberg. Integrating the objective of maximizing performance, for instance, based on patient utility functions, and investigating alternative dispatching strategies in combination with solutions of the QRUBUL-HR-HP are seen as promising next steps towards a comprehensive planning approach for heterogeneous patients.

Chapter 11

Conclusion and Outlook

EMS are an integral part of healthcare systems. They respond to medical emergencies outside medical facilities, treat patients on scene, and, if necessary, transport them to facilities for further care. The pressure on EMS in Germany is increasing due to a rising number of emergency calls, an increase in cost, and staff shortages.

This thesis studies EMS from a logistic perspective, and arose from the research project EVRALOG-BW[1] which was conducted with the EMS in Baden-Württemberg, Germany. The starting point for this thesis is the debate about response time targets. Response time targets are currently the most common evaluation metric for EMS logistics, and the majority of modeling approaches base their objective on response time targets. It is, however, unclear whether response time targets are suitable to quantify the impact of logistic decisions in EMS on patient outcomes. This thesis, therefore, set out to address the following fundamental question:

How can the impact on patient outcomes be considered when evaluating EMS systems from a logistic perspective, and what does this imply for making logistic decisions?

The following sections conclude the thesis by summarizing our approach, presenting the key findings related to our research questions, and pointing out avenues for further research. We divide our research into the three topics *quality evaluation*, *system behavior*, and *logistic decisions*, and follow this structure both for the conclusion and the outlook. *Quality evaluation* covers quantitatively evaluating the behavior of EMS systems from a logistic perspective. *System behavior* encompasses modeling how logistic decisions impact the behavior of EMS systems. *Logistic decisions* describes modeling approaches that result in suggestions for how to make logistic decisions.

[1]Funded by the Ministry of the Interior, Digitalisation and Local Government in Baden-Württemberg, Germany. Further information: https://healthcarelab.ksri.kit.edu/english/projects.php

11.1 Conclusion

Part II of this thesis describes our research for the topic *quality evaluation*, and Fig. 11.1 summarizes our research question and key findings.

Quality evaluation

RQ1: How can the quality of EMS logistics be quantified?

Key Findings:

- Existing quality indicators in EMS logistics focus on response times and do not sufficiently capture the effectiveness of logistic decisions.
- The importance of time and the relevant time interval along the rescue chain depend on the patient's needs.
- The current medical evidence and availability of connected outcome and process data is insufficient to quantify the effectiveness of EMS logistics for the majority of patients data-based.
- Utility functions defined by medical experts for patient categories are a step towards quantifying the effectiveness of EMS logistics for heterogeneous patients.

Figure 11.1: Research question and findings for the topic system evaluation.

The quality of healthcare in general, and specifically EMS systems as part of healthcare systems, is difficult to define and even more difficult to quantify. One approach to defining quality in healthcare is the disaggregated approach of defining quality dimensions. Logistic decisions can impact the quality of EMS in several quality dimensions. Quality indicators are quantitative measures that aim to make quality dimensions measurable. Evaluating the quality of EMS logistics should explicitly consider the patient perspective, which is captured in the quality dimension of effectiveness. Effectiveness connects the output of logistic decisions, such as response or prehospital times, to patient outcomes. Existing quality indicators fail to quantify the effectiveness of EMS logistics for the vast majority of patients. One reason for the lack of quality indicators of effectiveness is the insufficient empirical evidence for a data-based quantification of effectiveness for most emergencies. A second reason is that effectiveness bridges the disciplines of logistics and medicine, making it challenging for either discipline in isolation to develop comprehensive quality indicators of effectiveness. Interdisciplinary approaches, such as expert consensus processes, can yield quality indicators that connect the disciplines despite the lack of empirical data. Utility functions map time to a utility value and can be defined for categories of patients by medical experts. Utility functions overcome limitations of existing response time target-based quality indicators and present a step towards understanding how logistic decisions in EMS impact the heterogeneous types of patients in EMS.

System behavior is the topic of Part III of this thesis, and Fig. 11.2 summarizes our research question and key findings

System behavior

RQ2: How can the effects of logistic decisions on the quality of EMS be predicted?

Key Findings:
- Simulation models combining agent-based and discrete-event paradigms are suitable for studying the effects of logistic decisions on EMS system behavior.
- Simulation models tend to be designed for a specific EMS system, hindering the transferability of models between EMS systems.
- Our simulation model closely resembles historical system behavior for emergency rescue operations in Baden-Württemberg, and additional information and data regarding the choice of transport destinations could further enhance the applicability of the model.

Figure 11.2: Research question and key findings for the topic system behavior

Simulation is a suitable and commonly employed method for modeling the logistic behavior of EMS systems. Simulation models tend to be designed for one specific EMS system, and often the specific aspect of the EMS system under study. This makes it challenging to transfer and adopt existing modeling approaches. One challenge in designing truly generic and transferable models lies in combining the ad-hoc emergency rescue and scheduled patient transportation tasks of EMS systems. The second challenge lies in modeling the individual behavior of different resource types and their coordination during joint operations. Abstracting resource behavior and employing coordination concepts of task leaders and task backlogs in simulation models can tackle these challenges. Hybrid modeling approaches combining agent-based and discrete-event modeling paradigms appear suitable the model the logistic behavior of EMS systems.

Parts IV and V investigate *logistic decisions*. Part IV explores the dispatching decision, and Fig. 11.3 lists the guiding research question and key findings.

Logistic decisions

RQ3: How can the ambulance dispatching decision consider heterogeneous patients?

Key Findings:
- Driving times to prioritized calls can be reduced significantly by not always dispatching the closest available ambulance to non-prioritized calls.
- The impact of not always dispatching the closest available ambulance depends on the region under study.
- Not always dispatching the closest-available ambulance significantly increases driving times to non-prioritized patients in a larger magnitude than the decrease in driving times to prioritized patients on our experiments.
- If actually prioritized calls are classified as non-prioritized in the dispatching decision too often, driving times to actually prioritized calls can increase beyond always dispatching the closest available ambulance.

Figure 11.3: Research question and key findings for investigating the dispatching decision.

The dispatching decision can consider heterogeneous patients through differentiated dispatching strategies that do not always dispatch the closest available ambulance. Our scoping review reveals a variety of dispatching strategies with differing attributes in the literature. Examining the attributes of dispatching strategies enables identifying suitable strategies for a specific EMS system. The lack of benchmark instances and a missing classification of EMS systems hinder the comparison of the effects of dispatching strategies without custom simulation experiments. We approach the dispatching decision by differentiating between prioritized calls, to which the closest available ambulance is always dispatched, and non-prioritized calls, to which not necessarily the closest available ambulance is dispatched. Differentiated dispatching strategies can significantly reduce driving times to prioritized calls. They are, at the same time, very sensitive to how well prioritized calls can be identified as such at the time of dispatch. Decreasing driving times to prioritized patients additionally comes at the cost of increasing driving times to non-prioritized calls. Simulation models enable quantifying this trade-off as well as the impacts of the classification quality.

Part V studies the strategic ambulance station location and sizing decision. Fig. 11.4 displays the research question and key findings for this part of the thesis.

Logistic decisions

RQ4: How can the ambulance station location and sizing decision consider heterogeneous patients?

Key Findings:

- The system-optimal demand assignment assumption reduces the number of prescribed resources compared to the full demand assignment assumption.
- The independence assumption in calculating ambulance availabilities leads to fewer prescribed ambulances in our instances compared to the dependence assumption.
- Modeling availabilities based on a station-specific perspective can avoid overcapacities induced by modeling availabilities based on an ambulance-specific perspective.
- Considering differentiated response time thresholds of heterogeneous patients can reduce the number of prescribed resources compared to a single coverage target.
- The effects of considering prehospital time based coverage targets depend on the instance characteristics.

Figure 11.4: Research question and key findings for investigating the station location and sizing decision.

Strategic planning in EMS generally has to consider the conflicting objectives of minimizing cost and maximizing the quality for patients, while fulfilling legal planning requirements. This applies to the decision on locating and sizing ambulance stations as well. The number and sizes of stations impact cost, the locations impact quality through their effect on driving times, and legal requirements exist in the form of coverage targets that prescribe which proportion of calls has to be reached within a certain response or prehospital time. We approach the decision based on the hierarchy of firstly minimizing cost to fulfill legal requirements, and secondly maximizing quality without increasing cost, while continuing

to fulfill legal requirements. The scope of this thesis encompasses minimizing cost to fulfill legal requirements. The model class of set covering models considering ambulance availabilities is suitable for this scope.
Heterogeneous patients can be considered in the ambulance station location and sizing decision through patient types and type-specific coverage targets. Considering heterogeneous patients can affect the validity of assumptions in existing modeling approaches beyond integrating differentiated patient types. We identify that the existing demand assignment and service time assumptions become less valid when considering heterogeneous patients. We propose a model based on the system-optimal demand assignment assumption to reduce overcapacities caused by the full and uniform demand assignment assumptions of existing modeling approaches. Extending the model to consider location-specific service times can reduce both over- and undercapacities caused by the assumption of uniform service times, but the effects depend on the instance characteristics. Solving these challenges in model assumptions enables formulating the QRUBUL-HR-HP as the first set covering model for locating and sizing ambulance stations, considering patient-type specific coverage targets that can be based on response or prehospital times.

Coming back to the fundamental question of this thesis, we found that EMS systems should be evaluated by explicitly considering the patient perspective. Patients in EMS are heterogeneous, and logistic decisions impact them differently. The differing needs of patients should be considered when making logistic decisions. Combining medical and logistic knowledge is necessary to design an EMS system that enables the highest level of quality that the available resources can achieve.

We have discussed the specific limitations of our research in each chapter. A general limitation of research in EMS logistics that applies to this thesis as well is the lack of both a classification of EMS systems and publicly available and widely adopted benchmark instances. It is reasonable to assume that our methodological approaches can apply to EMS systems outside Baden-Württemberg, and the variance in instance characteristics in our computational experiments enhances the generalizability of our findings. The universal validity of findings for EMS systems based on numeric experiments with a set of EMS instances can, however, not be claimed.

11.2 Outlook

Our findings point to several directions for further research in all three of the covered topics, and we provided detailed descriptions at the end of each chapter. In the following, we synthesize the main avenues for future research in each topic.

Quality Evaluation

The understanding of how logistics decisions impact patient outcomes could be improved by *connecting outcome and patient data*. The example of survival functions for sudden cardiac arrest demonstrates how time can be interpreted from a patient's perspective, but no equivalent for survival functions exists for the vast majority of emergencies. Combining process data from EMS with outcome data from on-scene treatment and especially from hospitals would allow an evidence-based approach to mapping time to patient outcome.

We propose to evaluate call responses using utility functions, which result in a utility value for a call. There exist several proposals in the field of Operations Research of how individual evaluation metrics can be aggregated with either an efficiency- or fairness-forward perspective. We believe, however, that an *interdisciplinary approach to aggregating individual evaluations* could deepen the understanding of how individual measures can be aggregated to a system evaluation. Operations Research can provide tools for quantifying abstract concepts and can assess the logistic consequences of aiming to improve a given system evaluation metric. System evaluation metrics for healthcare systems should, however, reflect the moral values of a society, and the toolbox of Operations Research is not suited to identifying what these are.

System Behavior

The main remaining challenge in modeling the behavior of EMS systems remains the reusability and transferability of simulation models. The cause of this challenge is both the broad diversity of worldwide EMS systems and a missing *comprehensive classification of EMS systems from a logistic perspective*. This leads to the development of models tailored only to specific EMS systems and difficulties in assessing which existing approaches might be transferable to a specific system. Such a classification could additionally enable a more modular design of simulation components that could accelerate developing new simulation models or adapting models if an EMS system changes.

Logistic Decisions

An overarching issue in modeling logistic decisions in EMS is the difficulty in comparing modeling approaches due to the variety of instances used in papers. The sensitivity of health-related data hinders publishing detailed instance data. A lack of knowledge about which instance characteristics impact the comparative performance of modeling approaches compounds the issue. We see a first step in addressing this issue in *classifying EMS instances*. In contrast to classifying EMS systems, classifying EMS instances should refer more to the spatial and temporal distribution of calls than to the resources and their behavior. Such a classification can facilitate a *development of benchmark instances* as a second step. Such a database of benchmark instances covering EMS instances with different characteristics would significantly improve the comparability of modeling approaches and would strengthen findings based on computational experiments.

Prescriptive models map input parameters to a decision. We follow existing literature in parameterizing our models using historical data. Decisions in practice are made for the future, and *combining call prediction models with prescriptive models* would likely increase the value of decision support based on prescriptive models. Investigating the benefit of parameterizing prescriptive models using models for call predictions and identifying the required spatial and temporal granularity of predictions could additionally guide the further development of predictive models.

EMS logistics remains a fascinating field with ample room for future research. We hope to have contributed to its evolution with this thesis, and look forward to future advances in research and application.

List of Figures

List of Tables

Bibliography

Aboueljinane, L., & Frichi, Y. (2022). A simulation optimization approach to investigate resource planning and coordination mechanisms in emergency systems. *Simulation Modelling Practice and Theory*, *119*, Article 102586. https://doi.org/10.1016/j.simpat.2022.102586

Aboueljinane, L., Sahin, E., & Jemai, Z. (2013). A review on simulation models applied to emergency medical service operations. *Computers & Industrial Engineering*, *66*(4), 734–750. https://doi.org/10.1016/j.cie.2013.09.017

Aboueljinane, L., Sahin, E., Jemai, Z., & Marty, J. (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. https://doi.org/10.1016/j.simpat.2014.05.007

Ahituv, N., & Berman, O. (1987). Devising a cooperation policy for emergency networks. *Journal of the Operational Research Society*, *38*(11), 1015–1029. https://doi.org/10.1057/jors.1987.172

Albert, L. A. (2023). A mixed-integer programming model for identifying intuitive ambulance dispatching policies. *Journal of the Operational Research Society*, *74*(11), 2300–2311. https://doi.org/10.1080/01605682.2022.2139646

Al-Shaqsi, S. Z. K. (2010). Response time as a sole performance indicator in EMS: Pitfalls and solutions. *OPEN Access Emergency Medicine*, *2*, 1–6. https://doi.org/10.2147/OAEM.S8510

Amorim, M., Ferreira, S., & Couto, A. (2018). Emergency medical service response: Analyzing vehicle dispatching rules. *Transportation Research Record*, *2672*(32), 10–21. https://doi.org/10.1177/0361198118781645

Andersson, H., Granberg, T. A., Christiansen, M., Aartun, E. S., & Leknes, H. (2020). Using optimization to provide decision support for strategic emergency medical service planning – three case studies. *International Journal of Medical Informatics*, *133*, Article 103975. https://doi.org/10.1016/j.ijmedinf.2019.103975

Andersson, T., & Värbrand, P. (2007). Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society*, *58*(2), 195–201. https://doi.org/10.1057/palgrave.jors.2602174

Ansari, S., McLay, L. A., & Mayorga, M. E. (2017). A maximum expected covering problem for district design. *Transportation Science*, *51*(1), 376–390. https://doi.org/10.1287/trsc.2015.0610

Aringhieri, R., Bocca, S., Casciaro, L., & Duma, D. (2018). A simulation and online optimization approach for the real-time management of ambulances. *2018 Winter Simulation Conference (WSC)*, 2554–2565. https://doi.org/10.1109/WSC.2018.8632231

Aringhieri, R., Carello, G., & Morale, D. (2007). *Ambulance location through optimization and simulation: The case of Milano urban area* (tech. rep.). Università degli Studi di Milano, Polo didattico e di ricerca di Crema; Crema. https://air.unimi.it/handle/2434/40782

Aringhieri, R., Carello, G., & Morale, D. (2016). Supporting decision making to improve the performance of an Italian emergency medical service. *Annals of Operations Research*, *236*(1), 131–148. https://doi.org/10.1007/s10479-013-1487-0

Aringhieri, R., Elena, B. M., Khodaparasti, S., & van Essen, J. T. (2017). Emergency medical services and beyond: Addressing new challenges through a wide literature review. *Computers & Operations Research*, *78*, 349–368. https://doi.org/10.1016/j.cor.2016.09.016

Azami-Aghdash, S., Moosavi, A., Gharaee, H., Sadeghi, G., Mousavi Isfahani, H., Ghasemi Dastgerdi, A., & Mohseni, M. (2021). Development of quality indicators to measure pre-hospital emergency medical services for road traffic injury. *BMC Health Services Research*, *21*(1), 235–247. https://doi.org/10.1186/s12913-021-06238-1

Ball, M. O., & Lin, F. L. (1993). Reliability model applied to emergency service vehicle location. *Operations Research*, *41*(1), 18–36. https://doi.org/10.1287/opre.41.1.18

Bandara, D., Mayorga, M. E., & Mclay, L. A. (2012). Optimal dispatching strategies for emergency vehicles to increase patient survivability. *International Journal of Operational Research*, *15*(2), 195–214. https://doi.org/10.1504/IJOR.2012.048867

Bandara, D., Mayorga, M. E., & Mclay, L. A. (2014). Priority dispatching strategies for EMS systems. *Journal of the Operational Research Society*, *65*(4), 572–587. https://doi.org/10.1057/jors.2013.95

Bélanger, V., Lanzarone, E., Nicoletta, V., Ruiz, A., & Soriano, P. (2020). A recursive simulation-optimization framework for the ambulance location and dispatching problem. *European Journal of Operational Research*, *286*(2), 713–725. https://doi.org/10.1016/j.ejor.2020.03.041

Bélanger, V., Ruiz, A., & Soriano, P. (2019). Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. *European Journal of Operational Research*, *272*(1), 1–23. https://doi.org/10.1016/j.ejor.2018.02.055

Beraldi, P., & Bruni, M. E. (2009). A probabilistic model applied to emergency service vehicle location. *European Journal of Operational Research*, *196*(1), 323–331. https://doi.org/10.1016/j.ejor.2008.02.027

Berlin, G. N., & Liebman, J. C. (1974). Mathematical analysis of emergency ambulance location. *Socio-Economic Planning Sciences*, *8*(6), 323–328. https://doi.org/10.1016/0038-0121(74)90036-6

Billhardt, H., Lujak, M., Sánchez-Brunete, V., Fernández, A., & Ossowski, S. (2014). Dynamic coordination of ambulances for emergency medical assistance services. *Knowledge-Based Systems*, *70*, 268–280. https://doi.org/10.1016/j.knosys.2014.07.006

Blozik, E., Reich, O., Rapold, R., Scherer, M., & Swiss Quality Indicator for Primary Care (SQIPRICA) Working Group. (2018). Evidence-based indicators for the measurement of quality of primary care using health insurance claims data in Switzerland: Results of a pragmatic consensus process. *BMC Health Services Research*, *18*(1), 743–754. https://doi.org/10.1186/s12913-018-3477-z

Borrás, F., & Pastor, J. T. (2002). The ex-post evaluation of the minimum local reliability level: An enhanced probabilistic location set covering model. *Annals of Operations Research*, *111*(1/4), 51–74. https://doi.org/10.1023/A:1020941400807

Brotcorne, L., Laporte, G., & Semet, F. (2003). Ambulance location and relocation models. *European Journal of Operational Research*, *147*(3), 451–463. https://doi.org/10.1016/S0377-2217(02)00364-8

Bürger, A., Wnent, J., Bohn, A., Jantzen, T., Brenner, S., Lefering, R., Seewald, S., Gräsner, J.-T., & Fischer, M. (2018). The effect of ambulance response time on survival following out-of-hospital cardiac arrest. *Deutsches Ärzteblatt International*, *115*(33-34), 541–548. https://doi.org/10.3238/arztebl.2018.0541

Campbell, S., Roland, M., & Buetow, S. (2000). Defining quality of care. *Social Science and Medicine*, *51*(11), 1611–1625. https://doi.org/10.1016/S0277-9536(00)00057-5

Carter, G. M., Chaiken, J. M., & Ignall, E. (1972). Response areas for two emergency units. *Operations Research*, *20*(3), 571–594. https://doi.org/10.1287/opre.20.3.571

Carvalho, A. S., Captivo, M. E., & Marques, I. (2020). Integrating the ambulance dispatching and relocation problems to maximize system's preparedness. *European Journal of Operational Research*, *283*(3), 1064–1080. https://doi.org/10.1016/j.ejor.2019.11.056

Chanta, S., Mayorga, M. E., & Albert, L. A. (2012). A hybrid tabu search for locating and dispatching emergency response units using an imbedded queuing model. *62nd IIE Annual Conference and Expo 2012*, 3179–3186. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84900317025&partnerID=40&md5=f40552426db4409642c818388f21be12

Chanta, S., Mayorga, M. E., Kurz, M. E., & McLay, L. A. (2011). The minimum p -envy location problem: A new model for equitable distribution of emergency resources. *IIE Transactions on Healthcare Systems Engineering*, *1*(2), 101–115. https://doi.org/10.1080/19488300.2011.609522

Chen, T.-T., Ma, M. H.-M., Chen, F.-J., Hu, F.-C., Lu, Y.-C., Chiang, W.-C., & Ko, P. C.-I. (2015). The relationship between survival after out-of-hospital cardiac arrest and process measures for emergency medical service ambulance team performance. *Resuscitation*, *97*, 55–60. https://doi.org/10.1016/j.resuscitation.2015.04.035

Church, R., & ReVelle, C. (1974). The maximal covering location problem. *Papers of the Regional Science Association*, *32*(1), 101–118. https://doi.org/10.1007/bf01942293

Coster, J. E., Irving, A. D., Turner, J. K., Phung, V.-H., & Siriwardena, A. N. (2018). Prioritizing novel and existing ambulance performance measures through expert and lay consensus: A three-stage multimethod consensus study. *Health Expectations*, *21*(1), 249–260. https://doi.org/10.1111/hex.12610

Current, J. R., & Storbeck, J. E. (1988). Capacitated covering models. *Environment and Planning B: Planning and Design*, *15*(2), 153–163. https://doi.org/10.1068/b150153

Da Ros, F., Di Gaspero, L., Roitero, K., La Barbera, D., Mizzaro, S., Della Mea, V., Valent, F., & Deroma, L. (2024). Supporting fair and efficient emergency medical services in a large heterogeneous region. *Journal of Healthcare Informatics Research*, *8*(2), 400–437. https://doi.org/10.1007/s41666-023-00154-1

Dantas, R. A. N., Torres, G. D. V., Salvetti, M. D. G., Dantas, D. V., & Mendonça, A. E. O. D. (2015). Instrument for assessing the quality of mobile emergency pre-hospital care: Content validation. *Revista da Escola de Enfermagem da USP*, *49*(3), 380–386. https://doi.org/10.1590/S0080-623420150000300004

Daskin, M. S. (1983). A maximum expected covering location model: Formulation, properties and heuristic solution. *Transportation Science*, *17*(1), 48–70. https://doi.org/10.1287/trsc.17.1.48

Degel, D., Wiesche, L., Rachuba, S., & Werners, B. (2015). Time-dependent ambulance allocation considering data-driven empirically required coverage. *Health Care Management Science*, *18*(4), 444–458. https://doi.org/10.1007/s10729-014-9271-5

Department of Health & Social Care. (2023, October). *Handbook to the NHS constiution for England: Updated 1 october 2023* (tech. rep.). https://www.gov.uk/government/publications/supplements-to-the-nhs-constitution-for-england/the-handbook-to-the-nhs-constitution-for-england

Deutsche Interdisziplinäre Vereinigung für Intensiv- und Notfallmedizin. (n.d.). MIND. Retrieved November 22, 2024, from https://www.divi.de/component/edocman/220308-divi-mind-40-codes-pdf/download?Itemid=0

Donabedian, A. (1980). *The definition of quality and approaches to its assessment*. Health Administration Press.

Dunford, J., Domeier, R. M., Blackwell, T., Mears, G., Overton, J., Rivera-Rivera, E. J., & Swor, R. (2002). Performance measurements in emergency medical services. *Prehospital Emergency Care*, *6*(1), 92–98. https://doi.org/10.1080/10903120290938896

Dzator, M., & Dzator, J. (2013). An effective heuristic for the p-median problem with application to ambulance location. *Opsearch*, *50*(1), 60–74. https://doi.org/10.1007/s12597-012-0098-x

El Sayed, M. J. (2012). Measuring quality in emergency medical services: A review of clinical performance indicators. *Emergency Medicine International*, *2012*, Article 161630. https://doi.org/10.1155/2012/161630

Elzhov, T. V., Mullen, K. M., Spiess, A.-N., & Bolker, B. (2023). *Minpack.lm: R interface to the levenberg-marquardt nonlinear least-squares algorithm found in MINPACK, plus support for bounds*. Manual. https://CRAN.R-project.org/package=minpack.lm

Enayati, S., Mayorga, M. E., Toro-Díaz, H., & Albert, L. A. (2019). Identifying trade-offs in equity and efficiency for simultaneously optimizing location and multipriority dispatch of ambulances. *International Transactions in Operational Research*, *26*(2), 415–438. https://doi.org/10.1111/itor.12590

Erkut, E., Ingolfsson, A., & Erdoğan, G. (2008). Ambulance location for maximum survival. *Naval Research Logistics (NRL)*, *55*(1), 42–58. https://doi.org/10.1002/nav.20267

Federal Ministry of Health. (2025, December). Key figures for statutory health insurance 2013 to 2025 (dt.: Kennzahlen der gesetzlichen Krankenversicherung 2013 bis 2025). Retrieved December 16, 2025, from https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Statistiken/GKV/Kennzahlen_Daten/KF2025Bund_Dezember_2025.pdf

Fischer, M., Kehrberger, E., Marung, H., Moecke, H., Prückner, S., Trentzsch, H., Urban, B., & Fachexperten der Eckpunktepapier-Konsensus-Gruppe. (2016). Eckpunktepapier 2016 zur notfallmedizinischen Versorgung der Bevölkerung in der Prähospitalphase und in der Klinik. *Notfall + Rettungsmedizin*, *19*(5), 387–395. https://doi.org/10.1007/s10049-016-0187-0

Fitzsimmons, J. A. (1971). An emergency medical system simulation model. *Proceedings of the 5th Conference on Winter Simulation*, 18–25. https://doi.org/10.1145/800294.811420

Fitzsimmons, J. A. (1973). A methodology for emergency ambulance deployment. *Management Science*, *19*(6), 627–636. https://doi.org/10.1287/mnsc.19.6.627

Forsythe, G. E., Malcolm, M. A., Moler, C. B., & Forsythe, G. E. (1977). *Computer methods for mathematical computations*. Prentice-Hall.

Frank, E., Hall, M. A., & Witten, I. H. (2016). *The WEKA workbench. Online appendix for "data mining: Practical machine learning tools and techniques"* (4th ed.). Morgan Kaufmann.

Fujiwara, O., Makjamroen, T., & Gupta, K. K. (1987). Ambulance deployment analysis: A case study of Bangkok. *European Journal of Operational Research*, *31*(1), 9–18. https://doi.org/10.1016/0377-2217(87)90130-5

Gendreau, M., Laporte, G., & Semet, F. (1997). Solving an ambulance location model by tabu search. *Location Science*, *5*(2), 75–88. https://doi.org/10.1016/S0966-8349(97)00015-6

Gesetz über den Rettungsdienst. (2024, August). https://www.landesrecht-bw.de/bsbw/document/jlr-RettDGBW2024pP6

Gholami-Zanjani, S. M., Pishvaee, M. S., & Torabi, S. A. (2018). OR models for emergency medical service (EMS) management. *International Series in Operations Research and Management Science*, *262*, 395–421. https://doi.org/10.1007/978-3-319-65455-3_16

Goetz, J. L., Keltner, D., & Simon-Thomas, E. (2010). Compassion: An evolutionary analysis and empirical review. *Psychological Bulletin*, *136*(3), 351–374. https://doi.org/10.1037/a0018807

Goldberg, J., Dietrich, R., Chen, J. M., Mitwasi, M., Valenzuela, T., & Criss, E. (1990). A simulation model for evaluating a set of emergency vehicle base locations: Development, validation, and usage. *Socio-Economic Planning Sciences*, *24*(2), 125–141. https://doi.org/10.1016/0038-0121(90)90017-2

Grot, M., Becker, T., Steenweg, P. M., & Werners, B. (2022). Enhanced coverage by integrating site interdependencies in capacitated EMS location models. *Health Care Management Science*, *25*(1), 42–62. https://doi.org/10.1007/s10729-021-09562-4

Gunes, E., & Szechtman, R. (2005). A simulation model of a helicopter ambulance service. *Proceedings of the Winter Simulation Conference, 2005*, 951–957. https://doi.org/10.1109/WSC.2005.1574344

Habicht, S., & Hahnen, D. (2024). Strategien gegen den Personalmangel im Rettungsdienst. *Notfall + Rettungsmedizin*, *27*(7), 537–541. https://doi.org/10.1007/s10049-023-01250-w

Haghani, A., Tian, Q., & Hu, H. (2004). Simulation model for real-time emergency vehicle dispatching and routing. *Transportation Research Record: Journal of the Transportation Research Board*, *1882*(1), 176–183. https://doi.org/10.3141/1882-21

Harbour, J. L. (1997). *The basics of performance measurement*. Productivity Press. http://www.loc.gov/catdir/enhancements/fy0801/97183217-d.html

Harewood, S. I. (2002). Emergency ambulance deployment in Barbados: A multi-objective approach. *Journal of the Operational Research Society*, *53*(2), 185–192. https://doi.org/10.1057/palgrave.jors.2601250

Haugland, H., Rehn, M., Klepstad, P., Krüger, A., & The EQUIPE-collaboration group. (2017). Developing quality indicators for physician-staffed emergency medical services: A consensus process. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *25*(1), 14–22. https://doi.org/10.1186/s13049-017-0362-4

Haugland, H., Uleberg, O., Klepstad, P., Krüger, A., & Rehn, M. (2019). Quality measurement in physician-staffed emergency medical services: A systematic literature review. *International Journal for Quality in Health Care*, *31*(1), 2–10. https://doi.org/10.1093/intqhc/mzy106

Henderson, S. G., & Mason, A. J. (2004). Ambulance service planning: Simulation and data visualisation. In M. L. Brandeau, F. Sainfort, & W. P. Pierskalla (Eds.), *Operations research and health care: A handbook of methods and applications* (pp. 77–102). Kluwer Academy Publishers. https://doi.org/10.1007/1-4020-8066-2_4

Howard, I., Cameron, P., Wallis, L., Castrén, M., & Lindström, V. (2019). Identifying quality indicators for prehospital emergency care services in the low to middle income setting: The South African perspective. *African Journal of Emergency Medicine: Revue Africaine De La Medecine D'urgence*, *9*(4), 185–192. https://doi.org/10.1016/j.afjem.2019.07.003

Hyman, J. M. (1983). Accurate monotonicity preserving cubic interpolation. *SIAM Journal on Scientific and Statistical Computing*, *4*(4), 645–654. https://doi.org/10.1137/0904045

Iannoni, A. P., Morabito, R., & Saydam, C. (2009). An optimization approach for ambulance location and the districting of the response segments on highways. *European Journal of Operational Research*, *195*(2), 528–542. https://doi.org/10.1016/j.ejor.2008.02.003

Ibri, S., Nourelfath, M., & Drias, H. (2012). A multi-agent approach for integrated emergency vehicle dispatching and covering problem. *Engineering Applications of Artificial Intelligence*, *25*(3), 554–565. https://doi.org/10.1016/j.engappai.2011.10.003

Ibri, S., Nourelfath, M., & Drias, H. (2013). A contribution for integrating dispatching and covering in the emergency vehicles management services. *International Journal of Advanced Operations Management*, *5*(2), 121–137. https://doi.org/10.1504/IJAOM.2013.053533

Inakawa, K., Furuta, T., & Suzuki, A. (2010). Effect of ambulance station locations and number of ambulances to the quality of the emergency service. *The Ninth International Symposium on Operations Research and Its Applications (ISORA'10)*, 340–347. https://api.semanticscholar.org/CorpusID:16597479

Ingolfsson, A., Budge, S., & Erkut, E. (2008). Optimal ambulance location with random delays and travel times. *Health Care Management Science*, *11*(3), 262–274. https://doi.org/10.1007/s10729-007-9048-1

Ingolfsson, A., Erkut, E., & Budge, S. (2003). Simulation of single start station for Edmonton EMS. *Journal of the Operational Research Society*, *54*(7), 736–746. https://doi.org/10.1057/palgrave.jors.2601574

Institute of Medicine. (1990). *Medicare: A strategy for quality assurance: Volume 1*. National Academies Press. https://doi.org/10.17226/1547

Iskander, W. H. (1989). Simulation modeling for emergency medical service systems. *Proceedings of the 21st Conference on Winter Simulation*, 1107–1111. https://doi.org/10.1145/76738.76879

Jagtenberg, C. J., Bhulai, S., & van der Mei, R. D. (2015). An efficient heuristic for real-time ambulance redeployment. *Operations Research for Health Care*, *4*, 27–35. https://doi.org/10.1016/j.orhc.2015.01.001

Jagtenberg, C. J., Bhulai, S., & van der Mei, R. D. (2017). Dynamic ambulance dispatching: Is the closest-idle policy always optimal? *Health Care Management Science*, *20*(4), 517–531. https://doi.org/10.1007/s10729-016-9368-0

Jagtenberg, C. J., & Mason, A. (2020a). Improving fairness in ambulance planning by time sharing. *European Journal of Operational Research*, *280*(3), 1095–1107. https://doi.org/10.1016/j.ejor.2019.08.003

Jagtenberg, C. J., & Mason, A. (2020b). Fairness in the ambulance location problem: Maximizing the Bernoulli-Nash social welfare. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3536707

Jagtenberg, C. J., van den Berg, P. L., & van der Mei, R. D. (2017). Benchmarking online dispatch algorithms for emergency medical services. *European Journal of Operational Research*, *258*(2), 715–725. https://doi.org/10.1016/j.ejor.2016.08.061

Jagtenberg, C. J., Vollebergh, M. A. J., Uleberg, O., & Røislien, J. (2021). Introducing fairness in Norwegian air ambulance base location planning. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *29*(1), 50–60. https://doi.org/10.1186/s13049-021-00842-0

Jánošíková, Ľ., Jankovič, P., Kvet, M., & Zajacová, F. (2021). Coverage versus response time objectives in ambulance location. *International Journal of Health Geographics*, *20*(1), 32–48. https://doi.org/10.1186/s12942-021-00285-x

Jánošíková, Ľ., Kvet, M., Jankovič, P., & Gábrišová, L. (2019). An optimization and simulation approach to emergency stations relocation. *Central European Journal of Operations Research*, *27*(3), 737–758. https://doi.org/10.1007/s10100-019-00612-5

Jarvis, J. P. (1981). Optimal assignments in a markovian queueing system. *Computers and Operations Research*, *8*(1), 17–23. https://doi.org/10.1016/0305-0548(81)90028-9

Ji, S., Zheng, Y., Wang, W., & Li, T. (2020). Real-time ambulance redeployment: A data-driven approach. *IEEE Transactions on Knowledge and Data Engineering*, *32*(11), 2213–2226. https://doi.org/10.1109/TKDE.2019.2914206

Jones, J., & Hunter, D. (1995). Consensus methods for medical and health services research. *BMJ*, *311*(7001), 376–380. https://doi.org/10.1136/bmj.311.7001.376

Kergosien, Y., Bélanger, V., Soriano, P., Gendreau, M., & Ruiz, A. (2015). A generic and flexible simulation-based analysis tool for EMS management. *International Journal of Production Research*, *53*(24), 7299–7316. https://doi.org/10.1080/00207543.2015.1037405

Knight, V. A., Harper, P. R., & Smith, L. (2012). Ambulance allocation for maximal survival with heterogeneous outcome measures. *Omega*, *40*(6), 918–926. https://doi.org/10.1016/j.omega.2012.02.003

Koch, O., & Weigl, H. (2003). Modeling ambulance service of the Austrian red cross. *Proceedings of the 2003 Winter Simulation Conference*, 1701–1706. https://doi.org/10.1109/WSC.2003.1261622

Krueger, U., & Schimmelpfeng, K. (2013). Characteristics of service requests and service processes of fire and rescue service dispatch centers: Analysis of real world data and the underlying probability distributions. *Health Care Management Science*, *16*(1), 1–13. https://doi.org/10.1007/s10729-012-9207-x

Kubiak, T., Dudziński, Ł., Kasperczyk, R., & Czyżewski, Ł. (2024). Emergency services response to eCall system alerts: Observations from 2016–2022 in the national fire and rescue system. *Traffic Injury Prevention*, *25*(6), 819–824. https://doi.org/10.1080/15389588.2024.2348041

Larsen, M. P., Eisenberg, M. S., Cummins, R. O., & Hallstrom, A. P. (1993). Predicting survival from out-of-hospital cardiac arrest: A graphic model. *Annals of Emergency Medicine*, *22*(11), 1652–1658. https://doi.org/10.1016/S0196-0644(05)81302-2

Lavoie, G., Bélanger, V., De Montigny, L., & Lahrichi, N. (2020). A realistic simulation model of Montreal emergency medical services. In V. Bélanger, N. Lahrichi, E. Lanzarone, & S. Yalçındağ (Eds.), *Health care systems engineering* (pp. 67–78, Vol. 316). Springer International Publishing. https://doi.org/10.1007/978-3-030-39694-7_6

Law, A. M. (2015). *Simulation modeling and analysis* (5th ed.). McGraw-Hill US Higher Ed USE Legacy.

Lee, S. (2011). The role of preparedness in ambulance dispatching. *Journal of the Operational Research Society*, *62*(10), 1888–1897. https://doi.org/10.1057/jors.2010.145

Lee, S. (2017). A new preparedness policy for EMS logistics. *Health Care Management Science*, *20*(1), 105–114. https://doi.org/10.1007/s10729-015-9340-4

Lee, T., Cho, S.-H., Jang, H., & Turner, J. G. (2012). A simulation-based iterative method for a trauma center — air ambulance location problem. *Proceedings of the 2012 Winter Simulation Conference (WSC)*, 1–12. https://doi.org/10.1109/WSC.2012.6465042

Leknes, H., Aartun, E. S., Andersson, H., Christiansen, M., & Granberg, T. A. (2017). Strategic ambulance location for heterogeneous regions. *European Journal of Operational Research*, *260*(1), 122–133. https://doi.org/10.1016/j.ejor.2016.12.020

Li, H., & Chai, W. (2015). Research of resources allocation model in emergency decision of incidents. *Proceedings - 2015 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015*, 148–151. https://doi.org/10.1109/WI-IAT.2015.259

Li, X., Zhao, Z., Zhu, X., & Wyatt, T. (2011). Covering models and optimization techniques for emergency response facility location and planning: A review. *Mathematical Methods of Operations Research*, *74*(3), 281–310. https://doi.org/10.1007/s00186-011-0363-4

Lim, C. S., Mamat, R., & Braunl, T. (2011). Impact of ambulance dispatch policies on performance of emergency medical services. *IEEE Transactions on Intelligent Transportation Systems*, *12*(2), 624–632. https://doi.org/10.1109/TITS.2010.2101063

Liu, K., Li, X., Zou, C. C., Huang, H., & Fu, Y. (2020). Ambulance dispatch via deep reinforcement learning. *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 123–126. https://doi.org/10.1145/3397536.3422204

Liu, M.-S., & Lee, J.-T. (1988). A simulation of a hospital emergency call system using SLAMII. *Simulation*, *51*(6), 216–221. https://doi.org/10.1177/003754978805100602

Lopez, B., Innocenti, B., & Busquets, D. (2008). A multiagent system for coordinating ambulances for emergency medical services. *IEEE Intelligent Systems*, *23*(5), 50–57. https://doi.org/10.1109/MIS.2008.76

Lubicz, M., & Mielczarek, B. (1987). Simulation modelling of emergency medical services. *European Journal of Operational Research*, *29*(2), 178–185. https://doi.org/10.1016/0377-2217(87)90107-X

Luxem, J., Runggaldier, K., Karutz, H., & Flake, F. (Eds.). (2016). *Notfallsanitäter heute* (6. Auflage). Elsevier, Urban et Fischer.

MacLachlan, J., Mei, Y., Zhang, F., & Zhang, M. (2022). Genetic programming for vehicle subset selection in ambulance dispatching. *2022 IEEE Congress on Evolutionary Computation, CEC 2022 - Conference Proceedings*, 1–8. https://doi.org/10.1109/CEC55065.2022.9870323

MacLachlan, J., Mei, Y., Zhang, F., Zhang, M., & Signal, J. (2023). Learning emergency medical dispatch policies via genetic programming. *GECCO 2023 - Proceedings of the 2023 Genetic and Evolutionary Computation Conference*, 1409–1417. https://doi.org/10.1145/3583131.3590434

Maio, V. J., Stiell, I. G., Wells, G. A., & Spaite, D. W. (2003). Optimal defibrillation response intervals for maximum out-of-hospital cardiac arrest survival rates. *Annals of Emergency Medicine*, *42*(2), 242–250. https://doi.org/10.1067/mem.2003.266

Marianov, V., & ReVelle, C. (1994). The queuing probabilistic location set covering problem and some extensions. *Socio-Economic Planning Sciences*, *28*(3), 167–178. https://doi.org/10.1016/0038-0121(94)90003-5

Maxwell, M. S., Henderson, S. G., & Topaloglu, H. (2009). Ambulance redeployment: An approximate dynamic programming approach. *Proceedings of the 2009 Winter Simulation Conference (WSC)*, 1850–1860. https://doi.org/10.1109/WSC.2009.5429196

Mayorga, M. E., Bandara, D., & Mclay, L. A. (2013). Districting and dispatching policies for emergency medical service systems to improve patient survival. *IIE Transactions on Healthcare Systems Engineering*, *3*(1), 39–56. https://doi.org/10.1080/19488300.2012.762437

McCormack, R., & Coates, G. (2015). A simulation model to enable the optimization of ambulance fleet allocation and base station location for increased patient survival. *European Journal of Operational Research*, *247*(1), 294–309. https://doi.org/10.1016/j.ejor.2015.05.040

McLay, L. A. (2009). A maximum expected covering location model with two types of servers. *IIE Transactions*, *41*(8), 730–741. https://doi.org/10.1080/07408170802702138

McLay, L. A., & Mayorga, M. E. (2010). Evaluating emergency medical service performance measures. *Health Care Management Science*, *13*(2), 124–136. https://doi.org/10.1007/s10729-009-9115-x

Mclay, L. A., & Mayorga, M. E. (2013a). A dispatching model for server-to-customer systems that balances efficiency and equity. *Manufacturing and Service Operations Management*, *15*(2), 205–220. https://doi.org/10.1287/msom.1120.0411

Mclay, L. A., & Mayorga, M. E. (2013b). A model for optimally dispatching ambulances to emergency calls with classification errors in patient priorities. *IIE Transactions*, *45*(1, SI), 1–24. https://doi.org/10.1080/0740817X.2012.665200

Meinzer, N., & Storandt, S. (2014). Decision support in emergency medical systems: New strategies for dynamic ambulance allocation. *AAAI Workshop - Technical Report*, *WS-14-14*, 10–13. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84974808156&partnerID=40&md5=4a457945cacbf4294ce9dd56e52c1972

Metelmann, C., Metelmann, B., Kohnen, D., Brinkrolf, P., Andelius, L., Böttiger, B. W., Burkart, R., Hahnenkamp, K., Krammel, M., Marks, T., Müller, M. P., Prasse, S., Stieglis, R., Strickmann, B., & Thies, K. C. (2021). Smartphone-based dispatch of community first responders to out-of-hospital cardiac arrest - statements from an international consensus conference. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *29*(1), 29–38. https://doi.org/10.1186/s13049-021-00841-1

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & for the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*, *339*, Article b2535. https://doi.org/10.1136/bmj.b2535

Moore, L. (1999). Measuring quality and effectiveness of prehospital ems. *Prehospital Emergency Care*, *3*(4), 325–331. https://doi.org/10.1080/10903129908958963

Moré, J. J. (1978). The Levenberg-Marquardt algorithm: Implementation and theory. In G. A. Watson (Ed.), *Numerical analysis* (pp. 105–116, Vol. 630). Springer. https://doi.org/10.1007/BFb0067700

Mukhopadhyay, A., Pettet, G., Samal, C., Dubey, A., & Vorobeychik, Y. (2019). An online decision-theoretic pipeline for responder dispatch. *ICCPS 2019 - Proceedings of the*

2019 ACM/IEEE International Conference on Cyber-Physical Systems, 185–196. https://doi.org/10.1145/3302509.3311055

Mukhopadhyay, A., Pettet, G., Vazirizade, S. M., Lu, D., Jaimes, A., Said, S. E., Baroud, H., Vorobeychik, Y., Kochenderfer, M., & Dubey, A. (2022). A review of incident prediction, resource allocation, and dispatch models for emergency management. *Accident Analysis and Prevention*, *165*, Article 106501. https://doi.org/10.1016/j.aap.2021.106501

Müller, M. P., Ganter, J., Busch, H.-J., Trummer, G., Sahlmann, J., Brettner, F., Reden, M., Elschenbroich, D., Preusch, M., Rusnak, J., Katzenschlager, S., Nauheimer, D., Wunderlich, R., & Pooth, J.-S. (2024). Out-of-hospital cardiac arrest & SmartphonE RespOndErS trial (HEROES trial): Methodology and study protocol of a pre-post-design trial of the effect of implementing a smartphone alerting system on survival in out-of-hospital cardiac arrest. *Resuscitation Plus*, *17*, Article 100564. https://doi.org/10.1016/j.resplu.2024.100564

Murphy, A., Wakai, A., Walsh, C., Cummins, F., & O'Sullivan, R. (2016). Development of key performance indicators for prehospital emergency care. *Emergency Medicine Journal*, *33*(4), 286–292. https://doi.org/10.1136/emermed-2015-204793

Myers, J. B., Slovis, C. M., Eckstein, M., Goodloe, J. M., Isaacs, S. M., Loflin, J. R., Mechem, C. C., Richmond, N. J., & Pepe, P. E. (2008). Evidence-based performance measures for emergency medical services systems: A model for expanded EMS benchmarking. *Prehospital Emergency Care*, *12*(2), 141–151. https://doi.org/10.1080/10903120801903793

Nair, R., & Miller-Hooks, E. (2009). Evaluation of relocation strategies for emergency medical service vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, *2137*(1), 63–73. https://doi.org/10.3141/2137-08

Nasrollahzadeh, A. A., Khademi, A., & Mayorga, M. E. (2018). Real-time ambulance dispatching and relocation. *Manufacturing and Service Operations Management*, *20*(3), 467–480. https://doi.org/10.1287/msom.2017.0649

Neira-Rodado, D., Escobar-Velasquez, J. W., & McClean, S. (2022). Ambulances deployment problems: Categorization, evolution and dynamic problems review. *ISPRS International Journal of Geo-Information*, *11*(2), 109–146. https://doi.org/10.3390/ijgi11020109

Neukamm, J., Gräsner, J.-T., Schewe, J.-C., Breil, M., Bahr, J., Heister, U., Wnent, J., Bohn, A., Heller, G., Strickmann, B., Fischer, H., Kill, C., Messelken, M., Bein, B., Lukas, R., Meybohm, P., Scholz, J., & Fischer, M. (2011). The impact of response time reliability on CPR incidence and resuscitation success: A benchmark study from the German resuscitation registry. *Critical Care*, *15*(6), Article R282. https://doi.org/10.1186/cc10566 van den Berg, P. L., Henderson, S. G., Jagtenberg, C. J., & Li, H. (2025). Modeling the impact of community first responders. *Management Science*, *71*(2), 992–1008. https://doi.org/10.1287/mnsc.2022.04024

van den Berg, P. L., Kommer, G., & Zuzáková, B. (2016). Linear formulation for the maximum expected coverage location model with fractional coverage. *Operations Research for Health Care*, *8*, 33–41. https://doi.org/10.1016/j.orhc.2015.08.001

van den Berg, P. L., & van Essen, J. T. (2019). Scheduling non-urgent patient transportation while maximizing emergency coverage. *Transportation Science*, *53*(2), 492–509. https://doi.org/10.1287/trsc.2018.0823

van Buuren, M., Kommer, G. J., van der Mei, R., & Bhulai, S. (2017). EMS call center models with and without function differentiation: A comparison. *Operations Research for Health Care*, *12*, 16–28. https://doi.org/10.1016/j.orhc.2016.12.001

van Buuren, M., van der Mei, R., Aardal, K., & Post, H. (2012). Evaluating dynamic dispatch strategies for emergency medical services: TIFAR simulation tool. *Proceedings of the 2012 Winter Simulation Conference (WSC)*, 1–12. https://doi.org/10.1109/WSC.2012.6465214

van Buuren, M., van der Mei, R., & Bhulai, S. (2018). Demand-point constrained EMS vehicle allocation problems for regions with both urban and rural areas. *Operations Research for Health Care*, *18*, 65–83. https://doi.org/10.1016/j.orhc.2017.03.001

Olave-Rojas, D., & Nickel, S. (2021). Modeling a pre-hospital emergency medical service using hybrid simulation and a machine learning approach. *Simulation Modelling Practice and Theory*, *109*, Article 102302. https://doi.org/10.1016/j.simpat.2021.102302

Olivos, C., & Caceres, H. (2022). Multi-objective optimization of ambulance location in Antofagasta, Chile. *Transport*, *37*(3), 177–189. https://doi.org/10.3846/transport.2022.17073

OpenStreetMap contributors. (2025). Planet dump retrieved from https://planet.osm.org. https://www.openstreetmap.org

Pap, R., Lockwood, C., Stephenson, M., & Simpson, P. (2018). Indicators to measure prehospital care quality: A scoping review. *JBI Database of Systematic Reviews and Implementation Reports*, *16*(11), 2192–2223. https://doi.org/10.11124/JBISRIR-2017-003742

Pap, R., Lockwood, C., Stephenson, M., & Simpson, P. (2020). Development and testing of Australian prehospital care quality indicators: Study protocol. *BMJ OPEN*, *10*(7), Article e038310. https://doi.org/10.1136/bmjopen-2020-038310

Park, S. H., & Lee, Y. H. (2019). Two-tiered ambulance dispatch and redeployment considering patient severity classification errors. *Journal of Healthcare Engineering*, *2019*, 1–14. https://doi.org/10.1155/2019/6031789

Peleg, K., & Pliskin, J. S. (2004). A geographic information system simulation model of EMS: Reducing ambulance response time. *The American Journal of Emergency Medicine*, *22*(3), 164–170. https://doi.org/10.1016/j.ajem.2004.02.003

Pinto, L. R., Silva, P. M. S., & Young, T. (2015). A generic method to develop simulation models for ambulance systems. *Simulation Modelling Practice and Theory*, *51*, 170–183. https://doi.org/10.1016/j.simpat.2014.12.001

Pons, P. T., Haukoos, J. S., Bludworth, W., Cribley, T., Pons, K. A., & Markovchick, V. J. (2005). Paramedic response time: Does it affect patient survival? *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, *12*(7), 594–600. https://doi.org/10.1197/j.aem.2005.02.013

Pons, P. T., & Markovchick, V. J. (2002). Eight minutes or less: Does the ambulance response time guideline impact trauma patient outcome? *The Journal of Emergency Medicine*, *23*(1), 43–48. https://doi.org/10.1016/S0736-4679(02)00460-2

R Core Team. (2023). *R: A language and environment for statistical computing*. Manual. Vienna, Austria. https://www.R-project.org/

Repede, J. F., & Bernardo, J. J. (1994). Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *European Journal of Operational Research*, *75*(3), 567–581. https://doi.org/10.1016/0377-2217(94)90297-6

Restrepo, M., Henderson, S. G., & Topaloglu, H. (2009). Erlang loss models for the static deployment of ambulances. *Health Care Management Science*, *12*(1), 67–79. https://doi.org/10.1007/s10729-008-9077-4

Reuter-Oppermann, M., van den Berg, P. L., & Vile, J. L. (2017). Logistics for emergency medical service systems. *Health Systems*, *6*(3), 187–208. https://doi.org/10.1057/s41306-017-0023-x

Revelle, C., & Hogan, K. (1989). The maximum reliability location problem and α-reliable p-center problem: Derivatives of the probabilistic location set covering problem. *Annals of Operations Research*, *18*(1), 155–173. https://doi.org/10.1007/BF02097801

Reza Hosseini, S., Maleki, M., Gorji, H., Khorasani-Zavareh, D., & Roudbari, M. (2017). Challenges to use response time standard in assessing emergency medical services in Iran: A systematic review. *Archives of Trauma Research*, *6*(4), 55–63. https://doi.org/10.4103/atr.atr_29_17

Ridler, S., Mason, A. J., & Raith, A. (2022). A simulation and optimisation package for emergency medical services. *European Journal of Operational Research*, *298*(3), 1101–1113. https://doi.org/10.1016/j.ejor.2021.07.038

Roa, J. C. P., Escobar, J. W., & Moreno, C. A. M. (2020). An online real-time matheuristic algorithm for dispatch and relocation of ambulances. *International Journal of Industrial Engineering Computations*, *11*(3), 443–468. https://doi.org/10.5267/j.ijiec.2019.11.003

Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation*, *7*(1), 12–24. https://doi.org/10.1057/jos.2012.20

Savas, E. S. (1969). Simulation and cost-effectiveness analysis of New York's emergency ambulance service. *Management Science*, *15*(12), 608–627. https://doi.org/10.1287/mnsc.15.12.B608

Schehadat, M. S., Groneberg, D. A., Bauer, J., & Bendels, M. H. K. (2017). Hilfsfristen des Rettungsdienstes in den deutschen Bundesländern. *Zentralblatt Für Arbeitsmedizin, Arbeitsschutz Und Ergonomie*, *67*(5), 255–260. https://doi.org/10.1007/s40664-017-0203-3

Schlechtriemen, T., Burghofer, K., Stolpe, E., Altemeyer, K. H., & Lackner, C. K. (2005). Der Münchner NACA-score. *Notfall + Rettungsmedizin*, *8*(2), 109–111. https://doi.org/10.1007/s10049-005-0719-5

Schmid, V. (2012). Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research*, *219*(3), 611–621. https://doi.org/10.1016/j.ejor.2011.10.043

Schüttig, W., Flemming, R., Mosler, C. H., Leve, V., Reddemann, O., Schultz, A., Brua, E., Brittner, M., Meyer, F., Pollmanns, J., Martin, J., Czihal, T., Von Stillfried, D., Wilm, S., & Sundmacher, L. (2022). Development of indicators to assess quality and patient pathways in interdisciplinary care for patients with 14 ambulatory-care-sensitive conditions in Germany. *BMC Health Services Research*, *22*(1), Article 1015. https://doi.org/10.1186/s12913-022-08327-1

Shariat-Mohaymany, A., Babaei, M., Moadi, S., & Amiripour, S. M. (2012). Linear upper-bound unavailability set covering models for locating ambulances: Application to Tehran rural roads. *European Journal of Operational Research*, *221*(1), 263–272. https://doi.org/10.1016/j.ejor.2012.03.015

Sieber, F., Kotulla, R., Urban, B., Groß, S., & Prückner, S. (2020). Development of dispatch numbers and spectrum of emergency medical services in Germany; [Entwicklung der Frequenz und des Spektrums von Rettungsdiensteinsätzen in Deutschland]. *Notfall + Rettungsmedizin*, *23*(7), 490–496. https://doi.org/10.1007/s10049-020-00752-1

Silva, P. M. S., & Pinto, L. R. (2010). Emergency medical systems analysis by simulation and optimization. *Proceedings of the 2010 Winter Simulation Conference*, 2422–2432. https://doi.org/10.1109/WSC.2010.5678938

Snyder, M., & Smucker, B. J. (2022). Metamodel optimization of a complex, rural–urban emergency medical services system. *Simulation Modelling Practice and Theory*, *118*, Article 102526. https://doi.org/10.1016/j.simpat.2022.102526

Song, J., Li, X., & Mango, J. (2020). Location optimization of urban emergency medical service stations: A hierarchical multi-objective model with a new encoding method of genetic algorithm solution. In S. Di Martino, Z. Fang, & K.-J. Li (Eds.), *Web and wireless geographical information systems* (pp. 68–82, Vol. 12473). Springer International Publishing. https://doi.org/10.1007/978-3-030-60952-8_7

SQR-BW. (2016). Einsatzstichwortkatalog. https://www.sqrbw.de/fileadmin/SQRBW/Leitstellen/Einsatzstichwortkatalog_V1.1.pdf

SQR-BW. (2023, July). *Qualitätsbericht: Berichtsjahr 2022* (tech. rep.). Retrieved February 2, 2026, from https://www.sqrbw.de/fileadmin/SQRBW/Downloads/Qualitaetsberichte/SQRBW_Qualitaetsbericht_2022.pdf

SQR-BW. (2024, July). *Qualitätsbericht: Berichtsjahr 2023* (tech. rep.). Retrieved February 2, 2026, from https://www.sqrbw.de/fileadmin/SQRBW/Downloads/Qualitaetsberichte/SQRBW_Qualitaetsbericht_2023_web_Lesezeichen.pdf

SQR-BW. (2025, July). *Qualitätsbericht: Berichtsjahr 2024* (tech. rep.). Retrieved February 2, 2026, from https://www.sqrbw.de/fileadmin/SQRBW/Downloads/Qualitaetsberichte/SQRBW_Qualitaetsbericht_2024.pdf

Stein, C., Wallis, L., & Adetunji, O. (2015). The effect of the emergency medical services vehicle location and response strategy on response times. *South African Journal of Industrial Engineering*, *26*(2), 26–41. https://doi.org/10.7166/26-2-1078

Stelfox, H. T., Bobranska-Artiuch, B., Nathens, A., & Straus, S. E. (2010). A systematic review of quality indicators for evaluating pediatric trauma care: *Critical Care Medicine*, *38*(4), 1187–1196. https://doi.org/10.1097/CCM.0b013e3181d455fe

Stelfox, H. T., & Straus, S. E. (2013). Measuring quality of care: Considering conceptual approaches to quality indicator development and evaluation. *Journal of Clinical Epidemiology*, *66*(12), 1328–1337. https://doi.org/10.1016/j.jclinepi.2013.05.017

Strauß, N., Berrendorf, M., Haider, T., & Schubert, M. (2022). A comparison of ambulance redeployment systems on real-world data. *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 1–8. https://doi.org/10.1109/ICDMW58026.2022.00010

Strauss, C., Bildstein, G., Efe, J., Flacher, T., Hofmann, K., Huggler, M., Stämpfli, A., Schmid, M., Schmid, E., Gehring, C., Häske, D., Prückner, S., Stock, J. P., & Trentzsch, H. (2021). Optimizing emergency medical service structures using a rule-based discrete event simulation—a practitioner's point of view. *International Journal of Environmental Research and Public Health*, *18*(5), 1–22. https://doi.org/10.3390/ijerph18052649

Strohmandl, J., Tomek, M., Molnár, V., Jakubčeková, J. M., Fedorko, G., & Maláková, S. (2020). Using travel times for optimization numbers of medical rescue service points—case study from Slovakia. *Sustainability*, *13*(1), 207–219. https://doi.org/10.3390/su13010207

Su, S., & Shih, C.-L. (2002). Resource reallocation in an emergency medical service system using computer simulation. *The American Journal of Emergency Medicine*, *20*(7), 627–634. https://doi.org/10.1053/ajem.2002.35453

Su, S., & Shih, C.-L. (2003). Modeling an emergency medical services system using computer simulation. *International Journal of Medical Informatics*, *72*(1-3), 57–72. https://doi.org/10.1016/j.ijmedinf.2003.08.003

Sudtachat, K., Mayorga, M. E., & Mclay, L. A. (2014). Recommendations for dispatching emergency vehicles under multitiered response via simulation. *International Transactions in Operational Research*, *21*(4), 581–617. https://doi.org/10.1111/itor.12083

Swoveland, C., Uyeno, D. H., Vertinsky, I. B., & Vickson, R. G. (1973). A simulation-based methodology for optimization of ambulance service policies. *Socio-Economic Planning Sciences*, *7*(6), 697–703. https://doi.org/10.1016/0038-0121(73)90033-5

Theeuwes, N., Van Houtum, G.-J., & Zhang, Y. (2021). Improving ambulance dispatching with machine learning and simulation. In Y. Dong, N. Kourtellis, B. Hammer, & J. A. Lozano (Eds.), *Machine learning and knowledge discovery in databases. Applied data science track* (pp. 302–318, Vol. 12978). Springer International Publishing. https://doi.org/10.1007/978-3-030-86514-6_19

Toregas, C., Swain, R., ReVelle, C., & Bergman, L. (1971). The location of emergency service facilities. *Operations Research*, *19*(6), 1363–1373. https://doi.org/10.1287/opre.19.6.1363

Toro-Díaz, H., Mayorga, M. E., Chanta, S., & McLay, L. A. (2013). Joint location and dispatching decisions for emergency medical services. *Computers & Industrial Engineering*, *64*(4), 917–928. https://doi.org/10.1016/j.cie.2013.01.002

Toro-Díaz, H., Mayorga, M. E., Mclay, L. A., Rajagopalan, H. K., & Saydam, C. (2015). Reducing disparities in large-scale emergency medical service systems. *Journal of the Operational Research Society*, *66*(7), 1169–1181. https://doi.org/10.1057/jors.2014.83

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, *169*(7), 467–473. https://doi.org/10.7326/M18-0850

Trudeau, P., Rousseau, J.-M., Ferland, J. A., & Choquette, J. (1989). An operations research approach for the planning and operation of an ambulance service. *INFOR: Information Systems and Operational Research*, *27*(1), 95–113. https://doi.org/10.1080/03155986.1989.11732084

Ünlüyurt, T., & Tunçer, Y. (2016). Estimating the performance of emergency medical service location models via discrete event simulation. *Computers & Industrial Engineering*, *102*, 467–475. https://doi.org/10.1016/j.cie.2016.03.029

Uyeno, D. H., & Seeberg, C. (1984). A practical methodology for ambulance location. *Simulation*, *43*(2), 79–87. https://doi.org/10.1177/003754978404300202

Valenzuela, T. D., Roe, D. J., Cretin, S., Spaite, D. W., & Larsen, M. P. (1997). Estimating effectiveness of cardiac arrest interventions: A logistic regression survival model. *Circulation*, *96*(10), 3308–3313. https://doi.org/10.1161/01.CIR.96.10.3308

Vianen, N. J., Campfens, J. R., Brouwer-Bergsma, M., Van Ditshuizen, J. C., Giannakopoulos, G. F., Hoogerwerf, N., Den Hartog, D., Van Lieshout, E. M. M., Maissan, I. M., Schober, P., Venema, L., Verhofstad, M. H. J., & Van Vledder, M. G. (2025). Establishing outcome parameters for helicopter emergency medical services research in the Netherlands: Results of a mixed-methods delphi consensus study. *Prehospital Emergency Care*, *29*(7), 862–875. https://doi.org/10.1080/10903127.2024.2413038

Wang, W., Wu, S., Wang, S., Zhen, L., & Qu, X. (2021). Emergency facility location problems in logistics: Status and perspectives. *Transportation Research Part E: Logistics and Transportation Review*, *154*, Article 102465. https://doi.org/10.1016/j.tre.2021.102465

Wang, Y., Luangkesorn, K. L., & Shuman, L. (2012). Modeling emergency medical response to a mass casualty incident using agent based simulation. *Socio-Economic Planning Sciences*, *46*(4), 281–290. https://doi.org/10.1016/j.seps.2012.07.002

Watzinger, S. (2020, July). Hybrid simulation for quality assurance in emergency medical services in Baden-Württemberg (master thesis).

Watzinger, S., Nießner, C., Schutz, C., Groß, D., Schmitz, D., Stock, J.-P., Fabrizio, M., Frey, P., Böhm, R., Sebold, S., Ade, T., & Nickel, S. (2024). Patientenorientierte Planungskriterien für die Logistik in der Notfallrettung. *Notfall + Rettungsmedizin*. https://doi.org/10.1007/s10049-023-01254-6

Wears, R., & Winton, C. (1993). Simulation modeling of prehospital trauma care. *Proceedings of 1993 Winter Simulation Conference - (WSC '93)*, 1216–1224. https://doi.org/10.1109/WSC.1993.718382

WHO. (2006). *Quality of care: A process for making strategic choices in health systems* (R. Kawar, P. Key, S. Leatherman, & R. Massoud, Eds.). World Health Organization. https://iris.who.int/handle/10665/43470

Yang, S., Hamedi, M., & Haghani, A. (2005). Online dispatching and routing model for emergency vehicles with area coverage constraints. *Transportation Research Record: Journal of the Transportation Research Board*, *1923*(1), 1–8. https://doi.org/10.1177/0361198105192300101

Yoon, S., & Albert, L. A. (2018). An expected coverage model with a cutoff priority queue. *Health Care Management Science*, *21*(4), 517–533. https://doi.org/10.1007/s10729-017-9409-3

Yoon, S., & Albert, L. A. (2020). A dynamic ambulance routing model with multiple response. *Transportation Research Part E: Logistics and Transportation Review*, *133*, Article 101807. https://doi.org/10.1016/j.tre.2019.11.001

Yoon, S., & Albert, L. A. (2021). Dynamic dispatch policies for emergency response with multiple types of vehicles. *Transportation Research Part E: Logistics and Transportation Review*, *152*, Article 102405. https://doi.org/10.1016/j.tre.2021.102405

Zaffar, M. A., Rajagopalan, H. K., Saydam, C., Mayorga, M., & Sharer, E. (2016). Coverage, survivability or response time: A comparative study of performance statistics used in ambulance location models via simulation–optimization. *Operations Research for Health Care*, *11*, 1–12. https://doi.org/10.1016/j.orhc.2016.08.001

Appendices

A Reviewing Quality Indicators for EMS Logistics

A.1 Overview of Quality Measures and Statistics

Table A.1: Process quality indicators with patient focus

	call wait time	dispatch time				response time								on scene time				service time	scene to hospital time	prehospital time						alarm to patient handover time	survival probability		
	mean	mean	sd	min	max	% < T	mean	sd	min	max	mean $\mid t > T$	$x_{0.95}$	ns	% < T	mean	max	ns	mean	ns	% < T	mean	sd	min	max	ns	ns	mean	$x_{0.25}$	$x_{0.75}$
Aboueljinane et al. (2014)						x																							
Lavoie et al. (2020)							x																						
Bélanger et al. (2020)							x																						
Amorim et al. (2018)						x	x																				x		
Jánošíková et al. (2019)						x	x			x																			
Kergosien et al. (2015)						x	x																						
McCormack and Coates (2015)						x																					x		
Pinto et al. (2015)							x	x																					
Stein et al. (2015)						x	x																						
Ünlüyurt and Tunçer (2016)						x																							
Zaffar et al. (2016)						x	x																				x		
van den Berg and van Essen (2019)						x	x			x		x																	
Jagtenberg and Mason (2020a)						x																							
Aringhieri et al. (2016)						x																							
Lim et al. (2011)						x	x																						
Coster et al. (2018)						x	x																		x				
Dantas et al. (2015)							x																						
Haugland et al. (2017)							x																		x				
Haugland et al. (2019)													x				x		x							x			
McLay and Mayorga (2010)						x																					x		
Reza Hosseini et al. (2017)						x	x																						
Chen et al. (2015)						x																							
Stelfox et al. (2010)						x								x															

Table A.1: Process quality indicators with patient focus

	call wait time	dispatch time				response time								on scene time				service time	scene to hospital time	prehospital time						alarm to patient handover time	survival probability		
	mean	mean	sd	min	max	% < T	mean	sd	min	max	mean | t > T	$x_{0.95}$	ns	% < T	mean	max	ns	mean	ns	% < T	mean	sd	min	max	ns	ns	mean	$x_{0.25}$	$x_{0.75}$
Ji et al. (2020)						x	x																						
Yoon and Albert (2021)						x																							
Jagtenberg et al. (2015)						x																							
Nasrollahzadeh et al. (2018)						x	x				x																		
Savas (1969)						x	x																						
Fitzsimmons (1971)		x	x	x	x		x	x	x	x											x	x	x	x					
Fitzsimmons (1973)						x	x					x						x											
Swoveland et al. (1973)							x	x																					
Berlin and Liebman (1974)						x	x																						
Uyeno and Seeberg (1984)							x																						
Fujiwara et al. (1987)							x											x											
Lubicz and Mielczarek (1987)						x	x			x					x	x					x			x					
M.-S. Liu and Lee (1988)																					x		x	x					
Iskander (1989)		x					x								x						x								
Trudeau et al. (1989)						x	x																						
Goldberg et al. (1990)						x																							
Wears and Winton (1993)		x					x																				x		
Repede and Bernardo (1994)						x	x	x																					
Harewood (2002)						x	x																						
Su and Shih (2002)							x	x													x	x							
Su and Shih (2003)		x	x			x	x	x													x	x							
Ingolfsson et al. (2003)						x	x																						
Koch and Weigl (2003)		x					x														x								

Table A.1: Process quality indicators with patient focus

	call wait time	dispatch time				response time								on scene time				service time	scene to hospital time	prehospital time						alarm to patient handover time	survival probability		
	mean	mean	sd	min	max	% < T	mean	sd	min	max	mean | t > T	$x_{0.95}$	ns	% < T	mean	max	ns	mean	ns	% < T	mean	sd	min	max	ns	ns	mean	$x_{0.25}$	$x_{0.75}$
Peleg and Pliskin (2004)						x																							
Gunes and Szechtman (2005)							x																						
Henderson and Mason (2004)						x																							
Aringhieri et al. (2007)						x																							
Maxwell et al. (2009)						x																							
Inakawa et al. (2010)							x																						
Silva and Pinto (2010)	x					x	x																						
T. Lee et al. (2012)																				x									
van Buuren et al. (2012)						x																							
Y. Wang et al. (2012)																											x	x	x
Jánošíková et al. (2021)						x	x																						
Carvalho et al. (2020)						x	x				x																		
T. Andersson and Värbrand (2007)						x																							
S. Lee (2017)							x																						
Restrepo et al. (2009)						x																							
Bandara et al. (2014)																											x		
Leknes et al. (2017)						x																					x		
Borrás and Pastor (2002)						x																							

Table A.2: Process quality indicators with resource utilization focus

	round trip time	mileage	resource utilization rate						ambulances simultaneously utilized	time for which no ambulance is busy	unused vehicles	overtime for paramedics		calls per ambulance	calls per base
	mean	sum	mean	sd	cv	min	max	ns	max	prop	count	count	mean	count	count
Aboueljinane et al. (2014)			x												
Jánošíková et al. (2019)			x		x										
Kergosien et al. (2015)			x									x			
McCormack and Coates (2015)			x												
Pinto et al. (2015)			x	x											
Zaffar et al. (2016)					x										
Savas (1969)	x		x												
Fitzsimmons (1971)			x												
Swoveland et al. (1973)								x							
Berlin and Liebman (1974)	x		x						x						
Fujiwara et al. (1987)	x		x							x				x	
Lubicz and Mielczarek (1987)			x												
Iskander (1989)			x								x				
Trudeau et al. (1989)			x												
Goldberg et al. (1990)			x			x	x								
Wears and Winton (1993)			x												
Harewood (2002)			x							x					
Su and Shih (2002)								x							
Su and Shih (2003)	x		x												
Ingolfsson et al. (2003)													x		
Koch and Weigl (2003)		x													
Gunes and Szechtman (2005)			x	x											x
Henderson and Mason (2004)								x							
Jánošíková et al. (2021)			x		x										
Restrepo et al. (2009)			x												

Table A.2: Process quality indicators with resource utilization focus

	round trip time	**mileage**	**resource utilization rate**						**ambulances simultaneously utilized**	**time for which no ambulance is busy**	**unused vehicles**	**overtime for paramedics**		**calls per ambulance**	**calls per base**
	mean	sum	mean	sd	cv	min	max	ns	max	prop	count	count	mean	count	count
Borrás and Pastor (2002)			x												

Table A.3: Process quality indicators with resource availability focus

	ambulance available at call			**dispatch from a depot which was not the closest**	**call queue length**		**system preparedness**
	ns	prop	count	count	mean	max	mean
Haugland et al. (2017)	x						
Berlin and Liebman (1974)				x			
Fujiwara et al. (1987)			x	x			
Wears and Winton (1993)		x					
Su and Shih (2003)		x					
Inakawa et al. (2010)		x					
Silva and Pinto (2010)					x	x	
Carvalho et al. (2020)							x

Table A.4: Structure quality indicators based on a graph

	coverage				**response time**							**driving distance to nearest base**
	single	double	triple	expected	mean	expected mean	max	$x_{0.95}$	utilitarian social welfare	bernoulli-nash social welfare	iso-elastic social welfare	mean
van den Berg and van Essen (2019)	x	x	x	x	x	x	x	x				
Song et al. (2020)												x
Jagtenberg and Mason (2020a)									x	x		
Jagtenberg and Mason (2020b)									x	x		
Jagtenberg et al. (2021)									x		x	

Table A.4: Structure quality indicators based on a graph

	coverage				**response time**							**driving distance to nearest base**
	single	double	triple	expected	mean	expected mean	max	$x_{0.95}$	utilitarian social welfare	bernoulli-nash social welfare	iso-elastic social welfare	mean
Strohmandl et al. (2020)	x	x										
Degel et al. (2015)	x	x										
Fujiwara et al. (1987)	x											

Table A.5: Structure quality indicators not based on a graph

	ems units per 100.00 inhabitants count	**ems units per km^2** count	**bases** count	**cost of stations and ambulances** sum
Haugland et al. (2017)	x	x		
Strohmandl et al. (2020)			x	
Savas (1969)				x

B Comparing Dispatching Strategies for Heterogeneous Patients

B.1 Initial Comparison of Strategies

Table B.1: Medians of utilities with 95% confidence intervals of simulation experiments for 2021

	1		2	
Strategy	KA	WT	KA	WT
Closest Available	0.141 (0.137-0.145)	0.204 (0.19-0.213)	0.244 (0.241-0.247)	0.188 (0.177-0.197)
Lee	0.142 (0.138-0.145)	0.204 (0.191-0.212)	0.245 (0.241-0.248)	0.187 (0.176-0.197)
Andersson	0.144 (0.14-0.148)	0.205 (0.197-0.217)	0.245 (0.242-0.249)	0.187 (0.176-0.196)
Amorim	0.141 (0.138-0.145)	0.204 (0.191-0.213)	0.245 (0.241-0.248)	0.187 (0.176-0.197)
Roa	0.132 (0.127-0.136)	0.194 (0.18-0.205)	0.243 (0.239-0.247)	0.183 (0.175-0.195)
Li	0.144 (0.139-0.148)	0.211 (0.199-0.219)	0.245 (0.242-0.249)	0.186 (0.176-0.196)
Bandara	0.148 (0.143-0.152)	0.207 (0.197-0.217)	0.247 (0.243-0.25)	0.187 (0.177-0.197)
Jagtenberg	0.146 (0.142-0.15)	0.208 (0.197-0.217)	0.246 (0.243-0.25)	0.187 (0.176-0.197)
Aringhieri	0.148 (0.144-0.152)	0.205 (0.196-0.216)	0.246 (0.242-0.25)	0.187 (0.176-0.197)

Table B.2: 5th percentiles of utilities with 95% confidence intervals of simulation experiments for 2021

	1		2	
Strategy	KA	WT	KA	WT
Closest Available	0.004 (0.004-0.005)	0.035 (0.022-0.044)	0.078 (0.074-0.081)	0.041 (0.034-0.046)
Lee	0.006 (0.005-0.006)	0.035 (0.022-0.044)	0.08 (0.076-0.083)	0.041 (0.034-0.046)
Andersson	0.006 (0.005-0.007)	0.037 (0.023-0.047)	0.081 (0.076-0.083)	0.04 (0.034-0.046)
Amorim	0.005 (0.005-0.006)	0.035 (0.023-0.044)	0.08 (0.076-0.083)	0.041 (0.034-0.046)
Roa	0.004 (0.004-0.005)	0.031 (0.023-0.038)	0.077 (0.073-0.08)	0.039 (0.034-0.046)
Li	0.005 (0.005-0.006)	0.037 (0.025-0.046)	0.079 (0.075-0.082)	0.041 (0.034-0.046)
Bandara	0.006 (0.005-0.007)	0.037 (0.029-0.047)	0.081 (0.077-0.084)	0.039 (0.034-0.046)
Jagtenberg	0.006 (0.005-0.007)	0.039 (0.025-0.047)	0.081 (0.078-0.083)	0.039 (0.034-0.046)
Aringhieri	0.006 (0.005-0.007)	0.034 (0.023-0.046)	0.081 (0.077-0.083)	0.039 (0.034-0.046)

Table B.3: Medians of driving times with 95% confidence intervals of simulation experiments for 2021

	A	
Strategy	KA	WT
Closest Available	6.96 (6.91-7.01)	6.26 (6.21-6.37)
Lee	6.94 (6.88-6.98)	6.26 (6.21-6.36)
Andersson	6.87 (6.8-6.94)	6.45 (6.37-6.6)
Amorim	6.95 (6.89-6.99)	6.26 (6.21-6.37)
Roa	7.04 (6.98-7.11)	6.42 (6.31-6.56)
Li	6.7 (6.64-6.78)	6.21 (6.16-6.29)
Bandara	6.58 (6.53-6.64)	6.21 (6.04-6.21)
Jagtenberg	6.67 (6.61-6.75)	6.21 (6.08-6.26)
Aringhieri	6.64 (6.6-6.7)	6.21 (6.11-6.28)

Table B.4: 95th percentiles of driving times with 95% confidence intervals of simulation experiments for 2021

	A	
Strategy	KA	WT
Closest Available	15.09 (14.85-15.27)	15.67 (15.02-16.04)
Lee	14.78 (14.51-15.06)	15.59 (14.96-15.94)
Andersson	14.2 (14.01-14.4)	16.97 (16.27-17.79)
Amorim	14.93 (14.67-15.1)	15.6 (15.02-16.04)
Roa	17.65 (17.42-17.99)	16.03 (15.67-16.24)
Li	15.19 (15.04-15.36)	15.89 (15.51-16.57)
Bandara	14.06 (13.91-14.3)	15.36 (15-15.78)
Jagtenberg	13.97 (13.78-14.18)	15.27 (14.93-15.71)
Aringhieri	13.85 (13.69-14.02)	15.54 (15.02-15.85)

B.2 Analysis of Effective Strategies

Table B.5: Medians of utilities with 95% confidence intervals of simulation experiments for 2023

	1		2	
Strategy	North	South	North	South
Closest Available	0.195 (0.193-0.197)	0.245 (0.241-0.25)	0.254 (0.252-0.256)	0.261 (0.258-0.263)
Bandara	0.203 (0.2-0.205)	0.251 (0.248-0.255)	0.256 (0.255-0.258)	0.262 (0.259-0.265)
Jagtenberg	0.2 (0.198-0.202)	0.251 (0.247-0.254)	0.256 (0.255-0.258)	0.262 (0.259-0.265)
Aringhieri	0.201 (0.198-0.203)	0.245 (0.241-0.25)	0.256 (0.254-0.257)	0.259 (0.256-0.261)

Table B.6: 5th percentiles of utilities with 95% confidence intervals of simulation experiments for 2023

	1		2	
Strategy	North	South	North	South
Closest Available	0.014 (0.013-0.015)	0.046 (0.044-0.05)	0.084 (0.082-0.086)	0.078 (0.074-0.08)
Bandara	0.018 (0.017-0.019)	0.054 (0.05-0.057)	0.086 (0.084-0.089)	0.078 (0.075-0.081)
Jagtenberg	0.018 (0.017-0.02)	0.052 (0.049-0.056)	0.087 (0.085-0.089)	0.078 (0.074-0.081)
Aringhieri	0.017 (0.016-0.018)	0.043 (0.04-0.046)	0.086 (0.084-0.088)	0.077 (0.073-0.08)

Table B.7: Medians of driving times with 95% confidence intervals of simulation experiments for 2023

	A	
Strategy	North	South
Closest Available	6.87 (6.84-6.9)	5.92 (5.89-5.98)
Bandara	6.49 (6.46-6.51)	5.73 (5.68-5.78)
Jagtenberg	6.57 (6.55-6.6)	5.74 (5.68-5.79)
Aringhieri	6.55 (6.53-6.58)	5.9 (5.86-5.94)

Table B.8: 95th percentiles of driving times with 95% confidence intervals of simulation experiments for 2023

	A	
Strategy	North	South
Closest Available	14.54 (14.44-14.66)	14.73 (14.52-14.99)
Bandara	13.54 (13.45-13.62)	13.95 (13.82-14.09)
Jagtenberg	13.47 (13.38-13.56)	13.86 (13.74-14.01)
Aringhieri	14.31 (14.2-14.45)	15.75 (15.55-15.95)

Table B.9: Medians of utilities with 95% confidence intervals of simulation experiments for 2023

	3		4		5		6	
Strategy	North	South	North	South	North	South	North	South
Closest Available	0.729 (0.727-0.73)	0.766 (0.763-0.768)	0.7 (0.698-0.701)	0.722 (0.72-0.724)	0.956 (0.955-0.956)	0.975 (0.974-0.976)	0.912 (0.91-0.914)	0.935 (0.933-0.938)
Bandara	0.613 (0.611-0.615)	0.669 (0.665-0.672)	0.659 (0.658-0.66)	0.689 (0.687-0.691)	0.714 (0.711-0.716)	0.794 (0.787-0.8)	0.7 (0.692-0.707)	0.813 (0.802-0.823)
Jagtenberg	0.611 (0.609-0.613)	0.656 (0.651-0.66)	0.659 (0.658-0.661)	0.683 (0.681-0.686)	0.715 (0.713-0.718)	0.795 (0.788-0.802)	0.701 (0.693-0.709)	0.813 (0.803-0.821)
Aringhieri	0.585 (0.583-0.587)	0.637 (0.633-0.641)	0.65 (0.649-0.651)	0.675 (0.673-0.678)	0.73 (0.728-0.733)	0.794 (0.787-0.801)	0.685 (0.676-0.694)	0.779 (0.761-0.793)

Table B.10: 5th percentiles of utilities with 95% confidence intervals of patient categories in simulation experiments for 2023.

	3		4		5		6	
Strategy	North	South	North	South	North	South	North	South
Closest Available	0.44 (0.434-0.447)	0.55 (0.542-0.556)	0.374 (0.371-0.378)	0.372 (0.366-0.379)	0.779 (0.775-0.782)	0.875 (0.871-0.878)	0.643 (0.629-0.654)	0.788 (0.772-0.801)
Bandara	0.394 (0.388-0.398)	0.478 (0.475-0.482)	0.337 (0.334-0.341)	0.337 (0.331-0.342)	0.474 (0.471-0.476)	0.516 (0.512-0.519)	0.355 (0.345-0.365)	0.48 (0.472-0.489)
Jagtenberg	0.395 (0.389-0.4)	0.469 (0.466-0.472)	0.338 (0.334-0.341)	0.332 (0.325-0.338)	0.476 (0.474-0.478)	0.515 (0.512-0.519)	0.37 (0.355-0.382)	0.474 (0.467-0.482)
Aringhieri	0.378 (0.373-0.383)	0.461 (0.457-0.464)	0.329 (0.326-0.332)	0.323 (0.317-0.329)	0.466 (0.464-0.469)	0.502 (0.499-0.505)	0.355 (0.342-0.366)	0.454 (0.445-0.462)

Table B.11: Medians of driving times with 95% confidence intervals of dispatching categories for simulation experiments for 2023

	B		C		D	
Strategy	North	South	North	South	North	South
Closest Available	7.25 (7.23-7.27)	6.12 (6.07-6.17)	7.06 (6.97-7.14)	5.86 (5.75-5.99)	7.32 (7.28-7.37)	6.19 (6.08-6.28)
Bandara	13.68 (13.64-13.72)	12.01 (11.95-12.06)	21.79 (21.5-22.09)	19.41 (18.95-19.9)	31.34 (31.18-31.49)	31.1 (30.65-31.44)
Jagtenberg	13.74 (13.71-13.78)	13.11 (13.04-13.18)	21.58 (21.28-21.85)	19.75 (19.31-20.35)	31.45 (31.31-31.6)	31.86 (31.42-32.26)
Aringhieri	15.64 (15.62-15.67)	14.65 (14.58-14.74)	24.62 (24.34-24.89)	23.54 (23.18-24.15)	31.95 (31.78-32.14)	32.21 (31.92-32.56)

Table B.12: 95th percentiles of driving times with 95% confidence intervals of dispatching categories in simulation experiments for 2023

	B		C		D	
Strategy	North	South	North	South	North	South
Closest Available	15.44 (15.37-15.5)	15.09 (14.94-15.21)	15.19 (14.95-15.45)	15.13 (14.75-15.63)	15.66 (15.57-15.74)	15.37 (15.13-15.6)
Bandara	20.83 (20.81-20.86)	20.5 (20.44-20.55)	35.93 (35.63-36.1)	36.01 (35.66-36.45)	55.41 (55.24-55.6)	55.19 (54.96-55.53)
Jagtenberg	20.76 (20.74-20.79)	20.79 (20.75-20.82)	35.3 (34.98-35.53)	36.03 (35.77-36.41)	55.27 (55.08-55.43)	55.21 (54.91-55.52)
Aringhieri	21.51 (21.5-21.53)	21.36 (21.33-21.38)	37.62 (37.49-37.78)	37.44 (37.29-37.78)	57.67 (57.5-57.85)	57.79 (57.49-58.05)

Table B.13: Medians of driving times with 95% confidence intervals of simulation experiments for 2023

	A	
Strategy	North	South
Closest Available	6.92 (6.88-6.94)	5.97 (5.92-6.03)
Bandara_100	6.52 (6.49-6.54)	5.77 (5.72-5.82)
Bandara_97.5	6.6 (6.57-6.63)	5.85 (5.8-5.88)
Bandara_95	6.69 (6.66-6.73)	5.9 (5.87-5.94)
Bandara_92.5	6.83 (6.79-6.85)	5.97 (5.91-6.03)
Bandara_90	6.92 (6.89-6.94)	6.07 (6.01-6.15)

C Locating and Sizing Ambulance Stations for Heterogeneous Patients

Table C.1: Prescribed resources and simulated coverage for the QRLSCP and QRUBUL

	Stations		Ambulances		Driving time coverage		$\Delta^{max}_{driving\_coverage}$	
ID	QRLSCP	QRUBUL	QRLSCP	QRUBUL	QRLSCP	QRUBUL	QRLSCP	QRUBUL
1	17	**16**	44	**34**	93.53%	**92.92%**	0%pt.	**-0.61%pt.**
2	19	**18**	53	**42**	94.37%	**93.57%**	0%pt.	**-0.8%pt.**
3	**17**	**17**	38	**30**	91.08%	**90.33%**	0%pt.	**-0.75%pt.**
4	**22**	**22**	45	**37**	**92.8%**	93.61%	**-0.81%pt.**	0%pt.
5	**16**	**16**	46	**36**	**91.49%**	91.83%	**-0.34%pt.**	0%pt.
6	**20**	**20**	45	**35**	90.6%	**88.9%**	0%pt.	**-1.7%pt.**
7	**25**	**25**	46	**38**	**95.41%**	95.46%	**-0.05%pt.**	0%pt.
8	33	**32**	90	**72**	89.26%	**87.39%**	0%pt.	**-1.87%pt.**
9	**24**	**24**	65	**50**	**90.44%**	91.59%	**-1.15%pt.**	0%pt.
10	29	**28**	84	**67**	**89.56%**	91.09%	**-1.53%pt.**	0%pt.
11	**40**	41	90	**69**	96.15%	**95.72%**	0%pt.	**-0.42%pt.**
12	**43**	44	133	**101**	91.51%	**91.31%**	0%pt.	**-0.2%pt.**
13	**84**	85	180	**142**	96.43%	**96.28%**	0%pt.	**-0.15%pt.**
14	**84**	**84**	236	**187**	90.75%	**90.32%**	0%pt.	**-0.43%pt.**

Table B.14: 95th percentiles of driving times with 95% confidence intervals of simulation experiments for 2023

	A	
Strategy	North	South
Closest Available	14.7 (14.57-14.81)	14.98 (14.73-15.18)
Bandara_100	13.64 (13.57-13.75)	14.09 (13.95-14.24)
Bandara_97.5	14.61 (14.47-14.74)	14.75 (14.56-14.97)
Bandara_95	15.69 (15.59-15.8)	15.48 (15.31-15.68)
Bandara_92.5	16.41 (16.25-16.57)	16.36 (16.11-16.54)
Bandara_90	17.49 (17.37-17.62)	16.81 (16.65-17)

Table C.2: Comparison of optimization and simulation results for BRLSCP, QRLSCP, UBUL, and QRUBUL

	Stations				Ambulances				Driving time coverage				$\Delta^{max}_{driving\_coverage}$			
ID	BRLSCP	QRLSCP	UBUL	QRUBUL	BRLSCP	QRLSCP	UBUL	QRUBUL	BRLSCP	QRLSCP	UBUL	QRUBUL	BRLSCP	QRLSCP	UBUL	QRUBUL
1	**16**	17	**16**	**16**	41	44	**34**	**34**	93.78%	93.53%	**92.07%**	92.92%	0%pt.	-0.26%pt.	**-1.71%pt.**	-0.86%pt.
2	**18**	19	**18**	**18**	51	53	**38**	42	**92.82%**	94.37%	93.47%	93.57%	**-1.55%pt.**	0%pt.	-0.9%pt.	-0.8%pt.
3	**17**	**17**	**17**	**17**	37	38	34	**30**	91.42%	91.08%	**88.33%**	90.33%	0%pt.	-0.34%pt.	**-3.08%pt.**	-1.09%pt.
4	**22**	**22**	**22**	**22**	43	45	44	**37**	**92.78%**	92.8%	93.63%	93.61%	**-0.85%pt.**	-0.83%pt.	0%pt.	-0.02%pt.
5	**16**	**16**	18	**16**	44	46	**32**	36	91.29%	91.49%	**90.54%**	91.83%	-0.54%pt.	-0.34%pt.	**-1.29%pt.**	0%pt.
6	**20**	**20**	**20**	**20**	43	45	42	**35**	89.76%	90.6%	90.5%	**88.9%**	-0.84%pt.	0%pt.	-0.09%pt.	**-1.7%pt.**
7	**24**	25	25	25	46	46	47	**38**	95.83%	95.41%	**95.19%**	95.46%	0%pt.	-0.42%pt.	**-0.64%pt.**	-0.37%pt.
8	33	33	33	**32**	84	90	**71**	72	87.9%	89.26%	88.26%	**87.39%**	-1.36%pt.	0%pt.	-1%pt.	**-1.87%pt.**
9	**24**	**24**	**24**	**24**	60	65	51	**50**	90.6%	**90.44%**	90.74%	91.59%	-0.99%pt.	**-1.15%pt.**	-0.85%pt.	0%pt.
10	30	29	29	**28**	77	84	**67**	**67**	**89.48%**	89.56%	89.97%	91.09%	**-1.61%pt.**	-1.53%pt.	-1.13%pt.	0%pt.
11	**40**	**40**	41	41	87	90	85	**69**	96.56%	96.15%	96.66%	**95.72%**	-0.1%pt.	-0.52%pt.	0%pt.	**-0.94%pt.**
12	**43**	**43**	46	44	122	133	**101**	**101**	91.8%	91.51%	92.38%	**91.31%**	-0.58%pt.	-0.87%pt.	0%pt.	**-1.07%pt.**
13	85	**84**	86	85	173	180	170	**142**	96.31%	96.43%	**96.24%**	96.28%	-0.12%pt.	0%pt.	**-0.19%pt.**	-0.15%pt.
14	**83**	84	89	84	220	236	**184**	187	**89.86%**	90.75%	90.74%	90.32%	**-0.89%pt.**	0%pt.	-0.01%pt.	-0.43%pt.