EFFICIENCY AND FAIRNESS IN AMBULANCE PLANNING

Caroline Jagtenberg
In 1792, Napoleon Bonaparte decided to have his injured soldiers dragged off the battlefield by horse-drawn carriages. At the time, these so-called *flying ambulances* were a novel idea, and they proved a complete success: not only did they increase the chances of survival of the wounded soldiers, they also lifted the morale and the confidence of the French troops [105]. Nowadays, ambulances are commonplace: in the Netherlands alone there are more than one million ambulance trips per year [89]. Emergency medical services (EMS) have evolved into a complex system of interacting ambulances, dispatch centers and hospitals, providing us with a challenge to model and optimize their dynamics. A timely response can literally be a matter of life or death, so naturally research is focused on reducing response times. One solution to improve response times is simply to drive faster. While Top Gear has looked into this option [10], this dissertation takes a different approach. We introduce mathematical models for various planning stages in the EMS process, aiming to reduce response times by a more efficient use of resources.

We continue this chapter by describing the events and processes that occur in an EMS system, and the typical planning questions an ambulance provider might face. Additionally, we give an overview of the literature in this field. Since most of the case studies in this thesis involve the Dutch EMS system, we also include a brief description of ambulance care in the Netherlands. We finish with an outline of the remaining chapters of this thesis.

### 1.1 Background and motivation

Emergency medical services deal with urgent requests for medical care and/or patient transport. A typical response process is as follows. The EMS provider learns about a request when a call arrives at the dispatch centre. The call is answered by a dispatcher, who starts the triage: a process to determine the location of the patient and the urgency of the request. If an ambulance is needed, the dispatcher decides which vehicle to send to the scene of the incident. Almost always, this will be the closest idle ambulance - except if a special vehicle is needed due to the specifics of the incident. The ambulance drives to the scene of the incident, where the paramedics spend a certain amount of time with the
patient. Then, it is decided whether or not the patient needs to be transported to a hospital. If not, the ambulance becomes idle at the scene of the incident. Otherwise, there is a travel time to the hospital, followed by a drop-off time during which the crew transfers the patient to the emergency department. If an ambulance becomes idle, it returns to one of the predefined waiting sites or bases. This response process is depicted in Figure 1.1.

In case a call arrives while all ambulances are busy, the dispatcher places the request in a queue, in order for it to be served as soon as a vehicle becomes idle. This situation, sometimes called *code red*, is quite rare in the Netherlands, but it appears to be more common in other places (e.g., Edmonton, Canada [58]).

Although there can be some differences between countries, the main goal of ambulance providers world wide is the same: provide good health care at a reasonable cost. Naturally, a tradeoff arises, and this warrants research for efficient operations. Medical decisions aside, in order to obtain efficiency there are many logistic aspects worth considering.

When ambulance providers face questions regarding their planning, geographical aspects and service level agreements are often involved. Typical questions are, for example: ‘Can we improve performance by placing bases in different locations?’ or ‘If we were to purchase one extra vehicle, would we be able to serve 95% of all calls on time?’ Other questions might be staff-related, such as whether to hire more paramedics, or reconsidering the roster. Perhaps an EMS provider is thinking about merging with a neighbouring provider, and wants to know how this will affect response times. Furthermore, EMS managers may anticipate to new scenarios, due to changing circumstances that they want to evaluate in theory before it occurs in practice. These what if scenarios could for example be: ‘What if this hospital closes their emergency department?’ or ‘What if the demand for ambulances increases by 5%?’

The effects of EMS related changes or decisions can be difficult to oversee due to the stochastic nature of incidents. Furthermore, the decisions involved are often interrelated. This creates challenging mathematical problems, which - combined with the importance of high-quality EMS operations to society - have led Operations Research practitioners to pay much attention to EMS systems.
Over the years, numerous mathematical models have been developed that deal with planning and efficiency questions.

The decisions involved in planning EMS operations can be divided in three different planning stages: (1) at the strategic level, long-term decisions are made such as the opening and closing of base locations, the purchase of vehicles and the hiring and firing of staff; (2) the tactical level deals with the medium-long term, which may include decisions like how many vehicles to position at each base and how to design a staff roster, and (3) operational planning involves day to day or even real-time decisions. The latter includes decisions regarding the dispatch policy, which hospital to choose and where to send idle vehicles.

In practice, the same ambulance providers that serve emergency requests also handle non-urgent patient transport. These transports are often ordered and scheduled in advance, which make them intrinsically different from the urgent requests. The transports can be planned, and consequently have led to a separate set of models in literature. While some decisions may involve both the urgent and non-urgent ambulance operations - for example, when they are executed by the same fleet - this thesis focuses on the urgent requests only.

1.2 Literature review

This section discusses the literature on ambulance planning and gives a short overview of the various techniques used. We focus on models that can be solved analytically, which we divide in two types: (1) static models, which deal with problems at the strategic and tactical level, and (2) dynamic models, which are concerned with daily or even real-time planning.

1.2.1 Static planning

When it comes to ambulance planning, strategic and tactical problems are often solved simultaneously. At this planning stage the problems are often emergency facility location problems. They deal with two types of decisions: ‘Which bases should be opened?’ and ‘How many vehicles should be placed at each base?’

At this point, static models are often used to describe the problem. Here, ‘static’ means that each ambulance is assigned to a base location, and after serving an incident the ambulance is assumed to return its own home base. Typically there is a limited number of vehicles that need to be distributed over a set of possible base locations. These static models often use integer linear programming (ILP) to solve the problem.

Numerous objectives exist in literature, inspired by either the local EMS rules or a researcher’s belief of what is a relevant measure. This section gives a brief overview of the literature; for a more elaborate discussion, see [71] and [22].

A common way to measure the performance of an EMS provider is in terms of the fraction of late arrivals, i.e., the fraction of all calls for which the response time is larger than a certain response time threshold (RTT). This is probably the
most widely used objective - and certainly the most relevant for the work in this thesis.

Early research in ambulance planning focused on deterministic location problems. These formulations ignore the stochastic aspects of an EMS system, for example by assuming that one vehicle is enough to cover a demand point. This is done, e.g., in the Location Set Covering Problem (LSCP) [110], which searches for the minimal number of bases to cover a region, and in the Maximal Covering Location Problem (MCLP) [34], which searches for the best possible locations for a given number of bases. Slightly more advanced models such as [51] recognize that one vehicle per base is most likely not enough to cover the demand; they include backup coverage by requiring a constant number of vehicles within reach of each demand point.

Later, research turned to probabilistic models: these explicitly model the probability that a vehicle is busy (due to serving other patients). A well-known example is the Maximum Expected Covering Location Problem formulation (MEXCLP) [36]. The MEXCLP model is particularly relevant for this thesis, as the underlying idea of MEXCLP is used throughout several chapters. Therefore, we next recap the full model as it was originally published.

**The MEXCLP Model.** In this formulation there is a set of ambulances, denoted $A$, that needs to be distributed over a set of possible base locations $W$. Each ambulance is modeled to be unavailable with a pre-determined probability $q$, called the *busy fraction*. Consider a node $i \in V$ that is within range of $k$ ambulances. The travel times $\tau_{ij}, i, j \in V$ are assumed to be deterministic, which allow us to straightforwardly determine this number $k$. If we let $d_i$ be the demand at node $i$, the expected covered demand of this vertex is

$$E_k = d_i(1 - q^k).$$

(1.1)

The marginal contribution of the $k$th ambulance to this expected value is $E_k - E_{k-1} = d_i(1 - q)q^{k-1}$. We introduce a binary variable $y_{ik}$ that is equal to 1 if and only if vertex $i \in V$ is within range of at least $k$ ambulances. The variables $x_j$ (for $j \in W$) represent the number of vehicles at each base. Let $W_i$ denote the set of bases that are within range of demand point $i$, that is: $W_i = \{j \in W : \tau_{ij} \leq T\}$, then we can formulate the MEXCLP model as:

Maximize \( \sum_{i \in V} \sum_{k=1}^{p} d_i(1 - q)q^{k-1} y_{ik} \)

subject to

\[
\begin{align*}
\sum_{j \in W_i} x_j &\geq \sum_{k=1}^{p} y_{ik}, & i \in V, \\
\sum_{j \in W} x_j &\leq |A|,
\end{align*}
\]
1.2. Literature review

\[ x_j \in \mathbb{N}, \quad j \in W, \]
\[ y_{ik} \in \{0, 1\}, \quad i \in V, k = 1, \ldots, p. \]

Note that there is no need to add the constraint \( y_{ih} \leq y_{ik} \) for \( h \leq k \). This will always hold for an optimal solution, since \( E_k - E_{k-1} \) is decreasing in \( k \).

Using a busy fraction makes the MEXCLP model elegant in its simplicity, but the underlying assumptions are quite strong. For example, in the definition of \( E_k \) in Equation (1.1) the underlying assumption is that vehicles operate independently. Furthermore, the busy fraction is the same for all vehicles, regardless of their position with respect to the demand and the other vehicles.

Despite these assumptions, the MEXCLP model has several upsides. First of all, the simplicity of the model ensures it is scalable. Second, it is a suitable base for many extensions. For example, there are extensions with stochastic travel times \([17, 54]\), and a time-dependent version that divides the time horizon in a set of time periods \([15]\).

A slightly different approach is taken in the Maximum Available Location Problem (MALP). MALP also uses a busy fraction \( q \), but maximizes the population that will find a vehicle available within a time standard with a certain (fixed) reliability \([97]\).

Some of the strong assumptions in MEXCLP - independent vehicles all having the same busy probability - are relaxed in the Hypercube Queuing Models (HQM) \([69]\), providing a more accurate representation of real systems. However, it should be noted that while restrictive assumptions limit a model’s applicability, improving the modelling of the system performance makes the problem increasingly complicated and correspondingly more difficult to optimize.

There are other models that consider more than just response time thresholds. We leave the definition of other performance indicators open, but examples include an average response time, or a probability of survival. To compute such performance measures, it is useful to condition on the incident location and the base location of the responding ambulance. These models implicitly or explicitly assume that the closest idle vehicle is sent to an incident.

1.2.2 Dynamic planning

Dynamic models are used in the operational planning. They concern on-the-fly decisions, based on real-time information such as the current position and status of vehicles. Note that this stands in contrast to the static models described earlier. Dynamic solutions often outperform static solutions; however, optimality can usually not be guaranteed. This section briefly summarizes the literature on dynamic models; a more elaborate overview can be found in \([11]\).

Most dynamic models concern redeployment: they aim to find good (re)distributions of vehicles when a number of ambulances is busy responding to incidents. This is sometimes referred to as repositioning, dynamic ambulance management or move-up. Over the last few years, redeployment has become increasingly popular in practice. Surveys of North American EMS operators
showed that the percentage of operators who used a dynamic strategy increased from 23% in 2001 [28] to 37% in 2009 [117] (see also [2]). This indicates that the EMS community is becoming more aware that a dynamic policy can help to perform better without increasing capacity.

Dynamic models usually do not search for good base locations, but instead consider the bases as a given, fixed set. The point of issue is to make decisions based on real-time information on the state of all vehicles and incidents. This makes for a complex problem, and systems quickly become intractable when the number of vehicles grows.

Perhaps it is due to the difficulty of the problem, that dynamic models attract a wide range of solution methods. For example, there are approaches using dynamic programming [18], Integer Linear Programming (ILP) [46], stochastic programming [87], simulation-based optimization [20] and approximate dynamic programming [77]. The redeployment policies that have been published so far are roughly dividable in two subclasses, which we will refer to as compliance tables and real-time optimization.

**Compliance tables** are essentially lookup tables describing the desired configuration for each number of available ambulances. In order to obtain such a table, the system’s state is defined as the number of available ambulances, and a model is formulated to find an optimal configuration for each state. Typically, such a model is an ILP that maximizes some objective for all possible system states. Constraints may be added to control the number of vehicles relocated between states. The model is solved once (a priori), and the result is stored in a compliance table, to be looked up and applied when needed. Examples of such models are [46, 86].

In general, it is hard to give a reasonable estimate for the performance of a lookup table policy without simulating the system. However, [2] introduces a Markov chain model that provides a good approximation to several performance measures. This model can thus be used to identify near-optimal lookup tables.

A redeployment policy in the form of a lookup table has advantages and disadvantages. On the upside, a lookup table is easy to explain, and many EMS providers are familiar with this type of policy. Note that the job of steering the set of available vehicles towards the prescribed configuration is usually left to the dispatchers. This brings us to a downside: a poorly executed redeployment can devaluate even the most crafty lookup table. Furthermore, a lookup table is in general not able to suggest the most effective move-ups, because the amount of information used is limited. Another type of policy - that also uses the current locations of vehicles - may therefore perform better or require fewer move-ups, or both. Moreover, note that in busy regions, where the number of idle ambulances changes rapidly, the system will not be in compliance with the lookup table for most of the time.

**Real-time optimization** models calculate the ‘best’ ambulance movement in real time. The first of such models is known as the Dynamic Double Standard.
1.2. Literature review

Model (DDSM), published in 2001 [45]. This is an extension of the static Double Standard Model [44]. The model is an integer program (IP) that maximizes the demand covered twice, subject to two coverage constraints with different thresholds. The suggested moves are balanced with a certain cost that takes into account ambulance activity history: this reduces the number of moves that are undesirable from the crew’s perspective (such as round trips).

Various papers model the randomness in the system explicitly, for example, by formulating the problem as a Markov decision process. When the model has only a few ambulances, one can solve it using exact dynamic programming (e.g., [121]). However, when the state space grows - for example due to the number of vehicles considered - the problem quickly becomes intractable. This is known as the curse of dimensionality [93]. Hence, in order to compute results for realistically-sized EMS regions, one needs to turn to alternative solution methods.

One way to deal with this curse of dimensionality is by looking only one time-step ahead. This is done in [6], which classifies possible redeployment actions by constructing several scenarios that may occur one time-step later and evaluates each feasible action under these scenarios. Another example of a myopic approach is [118], which determines redeployments of idle ambulances from a greedy algorithm that attempts to minimize a weighted sum of expected late and lost calls, as evaluated through simulations.

Other authors overcome the curse of dimensionality by applying Approximate Dynamic Programming (ADP). ADP is a powerful tool for solving stochastic and dynamic problems, and scales well to high-dimensional applications. There are multiple ways to apply ADP. In [102] the authors use a combination of aggregation and the post-decision state. The original problem is aggregated by placing a spatial grid over the geographic area, and dividing the time horizon in sub-intervals. The value function is then approximated by computing estimates for the aggregated states. The post-decision state describes the state of the system immediately after making the decision but before any new information arrives. Approximating the value function around the post-decision state removes the stochasticity at this point. For an elaborate discussion of the post-decision state, see [93]. In [77] ADP is applied in a different way. The value function is approximated by a linear combination of so-called base functions: well-chosen functions that each use limited state information, and are considered to hold explanatory power over the value of a state. The parameters that define the importance of each base function are tuned using simulated cost trajectories of the system. The mechanism to tune parameters to the use case is described in more detail in subsequent work [78]. This approach is a novel one, but it is time consuming to both implement and execute: for a large city the tuning process can take a year, which is reduced to twelve hours by using the post-decision state. It remains possible to calculate the repositioning decision in real time, because these heavy computations are done in a preparatory phase. The performance of the method

---

1 In fact, [93] mentions three curses of dimensionality: the state space, the outcome space and the action space.
is highly dependent on the choice of base functions. In [78], base functions are essentially Erlang loss functions: the city is decoupled into smaller, independent regions each containing only a single ambulance base, and each region is modeled as an Erlang loss system. Note that this includes the implicit assumption that an incident is likely to be served late if there are no idle vehicles present at the nearest base.

In [120, Chapter 8] the authors introduce an IP model that they claim can be viewed as an extension of the ADP model in [77]. The model is extended in multiple ways, including the use of at-hospital ambulances and adding a cost for moving an ambulance to a base. Furthermore, the tuning process is updated, although the general idea that simulations are employed for function evaluations remains the same. The article mentions that future research may be directed towards rewards collected on the road during ambulance moves, which is relatively unexplored in current literature.

Although it appears that the majority of the dynamic models has not been implemented in practice, [75] is an exception. It describes an IP model that has been implemented in a commercial software package called Optima Live [75]. The method is a real-time optimization system that maximizes the total value from user-defined coverage reward functions minus redeployment costs.

In order to evaluate and validate move-up models, researchers typically use simulation. This makes it possible to get realistic estimates of the performance of an EMS system. Simulation is also useful stand-alone, to evaluate and compare scenarios. This is done for example in [53], which estimates the impact if all ambulances in Edmonton were to begin and end their shifts at the same location. Finally, simulation is used in so-called simulation optimization approaches (e.g., [122]). An overview of computer simulation models used for the analysis and improvement of EMS can be found in [1].

1.2.3 Model features

Several choices can be made regarding the modelling of EMS processes. In general, it is safe to say that George E. P. Box was right when he said: “All models are wrong, but some are useful.” [21, p.424]. This section discusses the most important model features, motivates the choices in this thesis and summarizes alternative approaches in literature.

Response times
A response time is defined as the time between the receipt of a call until an ambulance arrives on-scene (see also Figure 1.1). It consists of several components. First of all, the triage process takes place. Then, the dispatcher decides which ambulance to send. Subsequently, the crew makes their way to its ambulance. Together, these three events constitute the pre-trip delay. The total response time is given by the pre-trip delay plus the actual driving time.

In literature we see several ways to incorporate this pre-trip delay. It is not uncommon to simply add the average delay to the travel time. However, in [54] it
is argued that the duration of the pre-trip delay is highly variable, and that hence such a deterministic approach is insufficient to accurately predict performance. This article includes a small case study showing that it may lead to either an under- or overestimation, but being only a couple of percent off, the magnitude of these mistakes seems small.

In this dissertation we take the same approach as Dutch ambulance providers: three minutes are ‘reserved’ for the pre-trip delay. This leaves at most twelve minutes of driving time in order to reach the incident within the prescribed fifteen minutes.

It is also debatable whether the driving times should be modeled as deterministic or stochastic. In literature many examples can be found taking either approach. Although the stochastic approach seems realistic, authors do not seem to agree on which distribution to use. For example, [54] suggests a lognormal distribution, whereas [17] proposes a normal distribution. Differences may depend on many things, including the country where the case study took place. Most chapters in this thesis assume driving times to be deterministic. This is perhaps the biggest simplification done in our models. Although they could be extended to stochastic driving times, this would make the notation more cumbersome - and solutions harder to compute.

Other literature includes approaches based on the distance between two points as the crow flies [3] and using Google Maps data [62] (perhaps multiplying the result with a factor to correct for the fact that EMS vehicles usually drive faster than regular cars).

Vehicle and patient types
Recall from Section 1.3 that an EMS provider may use several types of vehicles. Each vehicle type has its own characteristics, such as travel speed or the ability to reach a certain area. Vehicles and the corresponding crew may also differ in their capabilities regarding the handling of patients. For example, less equipped vehicles may not be able to help the most severely injured patients. Other vehicles may serve a patient at the incident scene, but do not provide patient transportation. Not only the vehicles, but also the patients may fall into different categories: the nature of the request may cause a need for a specifically equipped vehicle, for example with an incubator or a psychiatric nurse. Also, the patient’s urgency may dictate the use of a different response time target.

The EMS system is rather complex, and to accurately capture it one also needs a complex model. Incorporating one or several of the features above would make the problem even less tractable. Furthermore, such a model would lead to a solution that is highly tailored to the specific situation. Instead, this dissertation focuses on a single type of patients, all equally urgent, and one type of vehicle which is capable of serving any patient.

Variations over time
Several aspects of the EMS process may vary throughout the day or week. Some
models in literature anticipate time-dependent fluctuations, generally by considering the redeployment of ambulances to be pre-planned. These models are extensively surveyed in [71] and [22]. The rest of this section explores variations in demand and travel times in more detail.

It is commonly assumed that EMS call volumes follow a Poisson process. However, call volumes may vary by month, day of the week, and hour of the day (see for example [52]). If this is the case, the arrival process may be modelled as a time-varying Poisson process. To predict the arrival intensity in the future, there are several methods available. Successful approaches in an ambulance context include classical time series models [30] and Singular Spectrum Analysis (SSA) [114].

The travel times, and the corresponding coverage, may also vary over time. Some papers consider this explicitly (e.g., [103]). However, we point out that emergency services do not always experience the impact of the time of day on their response velocities. For example, empirical evidence shows only a minor impact for fire fighters in New York [64] and ambulances in Calgary [24]. Furthermore, even if one is certain that the time of day is relevant for the response velocities, the task remains to estimate the different velocities accurately. Care has to be taken of how to handle the data, for example, there is a risk of overfitting due to the data containing only a small number of trips from \( i \) to \( j \) in each time segment.

The methods and models introduced in this thesis assume fixed values for the demand and travel times. This assumption simplifies the notation and discussion, and allows the reader to focus on the core ideas of the proposed methods. When used in practice, the correct usage of models should be discussed with EMS managers. For example, one can simply use the peak demand (this is done, e.g., in [76]), or the week may be split up into different time blocks, using different parameters for each block.

### 1.3 Ambulance care in the Netherlands

The Netherlands is divided in 24 EMS regions, called Regionale Ambulancevoorzieningen (RAVs), depicted in Figure 1.2. The RAVs operate independently, although occasionally a neighboring RAV may be contacted for help.

In the Netherlands, ambulance providers distinguish four different call priorities. The most urgent calls are labeled A1, and require an ambulance to be at the scene within 15 minutes in 95% of the cases. A1 calls would include, for example, heart attacks or strokes. A response to A1 calls generally includes lights and sirens. Non-life-threatening yet urgent calls get an A2 label, which corresponds to a response time threshold of 30 minutes. Although lights and sirens are usually omitted for A2 calls, an ambulance is dispatched immediately. Both A1 and A2 calls have to be served by Advanced Life Support (ALS) ambulances. Additionally, there are B calls, which are non-urgent patient transports. These transports are often ordered in advance, and consequently there is no time
1.3. Ambulance care in the Netherlands

Currently, the Netherlands has 24 EMS regions. The B calls are further subdivided: B1 patients need to be transported in an ALS ambulance, while for B2 patients a cheaper Basic Life Support (BLS) ambulance would also be sufficient. All vehicles can transport at most one patient at a time. Unlike some other countries, the Netherlands uses the same standards for urban and rural areas\(^2\).

The association *Ambulancezorg Nederland* (AZN) reports on the numbers of production and performance of RAVs on an annual basis. In 2014, 1,190,320 incidents were served in the Netherlands. Roughly 49% of these were A1 calls, 24% were A2 calls and 27% B calls. They were served using 231 bases and 755 vehicles nationwide. For 93.4% of the A1 requests an ambulance arrived within the prescribed fifteen minutes. The average A1 response time, however, was a lot smaller: 6 minutes and 41 seconds [89].

The Dutch National Institute for Public Health and the Environment (RIVM) [98] distributes the national budget for ambulance care among the different RAVs. This is done using several mathematical models, which are updated and published every few years (see, e.g., [65, 67]). The RAVs are free to spend their budget whichever way they choose, for example by investing in different vehicle types that they deem appropriate. Currently, there exist configurations ranging from a paramedic on a bike to a mobile intensive care unit (micu) - pretty much an operating room on wheels - with three staff members on board. Additionally, some RAVs use helicopters and boats to extend their service.

When it comes to modelling Dutch EMS systems, there are a few standard approaches. For example, demand is typically aggregated by using the first four digits of postal codes. This leads to regions of moderate sizes, with 40 to 456 demand points. The demand per point can then be estimated in a few different ways. One may use the observed demand in recent years, albeit at the risk of overfitting. Alternatively, one may assume that demand is roughly proportional to the number of inhabitants per demand point. This assumption may not be completely correct, but on the upside the number of inhabitants is known with

\(^2\)although one might argue that the Netherlands does not have truly rural areas
great accuracy.

The RIVM estimated ambulance driving times between any two postal codes in the Netherlands [66, Chapter 3]. For this purpose, the RIVM used historical data from ambulances that drove from a base location to an incident. These measurements were further differentiated by time of day and type of region (urban or not). The travel times were then predicted by distinguishing twelve different road types, and estimating the travel speed at each road type.

1.4 Thesis outline

This thesis contains research that focuses both on theoretical results and practical applications. The content of this thesis is organized in six chapters.

Chapter 2 deals with the dispatch process. Most literature assumes that the closest idle ambulance is always sent, but this is not necessarily optimal. We provide two alternatives for the ‘closest idle’ method: one method is obtained by modelling the dispatch process as a Markov decision process, the other method is a heuristic. The optimal dispatch policy, however, remains unknown.

In Chapter 3 we bound the performance of an optimal dispatch policy. We do this by computing the optimum for the offline version of the dispatch problem. In the offline dispatch problem, the time and location of incidents are known in advance, which allows us to get better solutions than for the online problem. We analyze the problem from both a worst-case as well as the average-case point of view. By benchmarking the offline optimum against online policies, we give the first quantification of the ‘performance gap’ between online and offline dispatch policies.

Chapter 4 introduces an algorithm for proactive ambulance redeployment. Unlike many other redeployment algorithms in literature, our proposed solution is a polynomial-time heuristic that is easy to implement. We evaluate its performance in a simulation model of (EMS) operations and compare it to static solutions. The practical relevance of this chapter is demonstrated by the implementation of our heuristic in practice.

In Chapters 5 and 6 we focus on fairness in ambulance planning. Most models in ambulance planning maximize the number of people served, regardless of where they are living. This approach benefits people living in cities, at the expense of people living in remote areas. While most alternative models tend to aim for equity (providing the same service to people in every location), we seek for a compromise between these two options. This is done by viewing the ambulance location problem from a social welfare perspective: we show that maximizing the so-called Bernoulli-Nash social welfare results in a solution that we consider fair. Chapter 5 and 6 approach fairness in different ways: Chapter 5 introduces a facility location problem: we compute where to locate vehicles such that the Bernoulli-Nash social welfare is maximized. This requires the use of a non-linear model, which we approximate with piecewise linear functions and solve using a Mixed-Integer Linear Programming (MILP) solver. Chapter 6, on the
other hand, proposes to improve fairness by time-sharing several static ambulance configurations. The individual configurations are evaluated by simulation, and the optimal mix between configurations is then computed using an Interior Point optimizer.

Finally, Chapter 7 deals with stochastic scheduling: the scheduling of jobs with a stochastic processing time, on parallel, identical machines. In particular, we focus on Smith’s rule - scheduling jobs according to ratios weight over processing time - for minimizing the weighted sum of completion times. For jobs with deterministic processing times, Smith’s rule is known to have a tight performance guarantee of \((1 + \sqrt{2})/2\). We recap the instance that proves this performance bound is tight, and analyze its stochastic counterpart with exponentially distributed processing times. Our analysis allows us to derive new qualitative insights, and sheds light on previously unknown phenomena in stochastic scheduling.

The work that resulted in this thesis was part of a larger project, called ‘From REactive to PROactive planning of ambulance services’, shortly REPRO. The REPRO project was focused on several areas, including (1) the development of relocation algorithms for dynamic ambulance management, using for example compliance tables [5] or taking into account different vehicle types [8], (2) the development of facility location models, incorporating aspects such as fractional coverage [17] and time dependency [15], and (3) the development of capacity models for EMS call centers [25]. A key aspect of REPRO is that the research not only led to a range of academic contributions, but that the tool implementations of the models were also successfully applied in real life, supporting the operational processes of several ambulance service providers in the Netherlands [27] and in Norway [100].


[90] NHS. Ambulance quality indicator. [http://www.ambulancestats.co.uk](http://www.ambulancestats.co.uk).


[112] Personal communication with ambulance provider Utrecht (RAVU).


Caroline Jagtenberg was born in Grubbenvorst, the Netherlands, in 1987. She started her studies in 2005, combining two Bachelor’s programs: Mathematics and Physics & Astronomy at Utrecht University. She continued with a Master’s degree in Mathematical Sciences, also at Utrecht University, including a semester abroad at Lund University, Sweden. After finishing her Master’s program cum laude, she spent another semester at Monash University, Australia, studying various topics in computer Science.

In 2011, she started working as a software engineer at ORTEC in Gouda. There, she developed the back end of a routing application, and later switched focus to workforce planning. During the time of this research, Caroline remained employed at ORTEC for two days a week.

In 2012, Caroline started her PhD at the Stochastics group at CWI (the Dutch national research institute for mathematics and computer science) in Amsterdam. Her research was part of a larger research project called REPRO: from Reactive to Proactive planning of ambulance services. In the last year of her PhD program, Caroline spent four months at the Engineering Science department of the University of Auckland in New Zealand.
List of publications

Peer reviewed publications


M. van Buuren, C.J. Jagtenberg, T.C. van Barneveld, R.D. van der Mei and S. Bhulai. Ambulance dispatch center pilots proactive relocation policies to enhance effectiveness. *Submitted*.


C.J. Jagtenberg and A.J. Mason. Improving fairness in ambulance planning by time-sharing. *In preparation*. 
Professional publications

