# Efficient Planning of Ambulance Services

van Buuren, M.

2018

# SUMMARY

Mathematical models can contribute substantially to improve the service quality of ambulance care. This Ph.D. thesis entails various new models, which are not only interesting from a scientific point of view, but which have also widely found their way into practice.

The *dispatch center model*, as presented in Chapter 2, is a simulation model that describes the processes of the dispatch center agents in detail. Although the literature on general dispatch centers is diverse and extensive, it appears that specific models for ambulance dispatch centers are scarce. Characteristic for dispatch centers is that there are very strict requirements on the response times and a large diversity of tasks. Unique in our approach is that we have included various feedback moments, respect the priorities of the incoming call type, and that we can simulate the multiple dispatch centers staffing policies. More specifically, our model allows for both the function differentiation and the only generalist policies. In (complete) *function differentiation* there is a group of specialized centralists called the *call takers* who answer incoming emergency calls, and who perform the *triage* procedure. Another group of specialists are responsible for the coordination of the fleet and the communication to third-parties such as hospitals; these are the so-called *dispatchers*. Generalists are flexible because they can do both the call taking and dispatching, but they are more expensive. The mixed model allows for a function differentiation policy that is supplemented with generalists. The dispatch center model, in all its forms, predicts the waiting and sojourn times, given the number of operators of each specialism. Based on the model results a cost–benefit analysis can be carried out. Using meta-data of telephone conversations by the dispatch center in Utrecht, we have been able to make good parameter estimations. We conclude that depending on the workload of the dispatch center, the optimal staffing policy changes: for low arrival rate, generalists are the most cost-efficient, but as this rate increases, function differentiation pays off. On top of the function differentiation policy, it is beneficial to have one generalist who can assist when the workload peaks, either in the call taking or the dispatching.

Chapters 3 and 4 present the so-called *adjusted queuing* framework. Primarily, ambulance service providers are assessed on the fraction of in-time ambulance arrivals for (potentially) life-threatening situations. For example, a fraction of 94% of the high urgency calls of an ambulance region has a response time of at most 15 minutes. In this calculation, the annual performance of the entire

ambulance region is aggregated into a single number, i.e., into one point in space and time. The disadvantage of this approach is that local interests are barely represented; after all, it pays off to concentrate all ambulances around the urban areas that have the majority of the demand, and as a consequence, take the relatively small amount of late ambulance arrivals in the rural areas for granted. *Minimal-reliability* (min-rel) models approach the performance differently. They calculate the minimal number of ambulances required, such that a given response time performance can be achieved at any subarea in the region. The related *maximal availability* (max-av) models try to place a limited number of ambulances, such that as many potential patients as possible are served within the given performance target. Existing min-rel and max-av models have a number of drawbacks. Many models have a system-wide busy fraction that is the equal for every ambulance, which is unrealistic as ambulances in rural areas are more at the base locations than urban ambulances. In previously published models, the arrival rate, the mean service time and the minimal required reliability threshold may not differ too much between neighboring (demand) points in the ambulance region; if that happens, it results in a higher number of required ambulances. Our approach, on the other hand, allows the aforementioned three parameters to differ for every demand point, and on top of that, in our approach the busy fraction can have a different value for each ambulance. In Chapter 3 we explain why the earlier models give an overestimation, and we show how these models have to be adapted such that they can be applied to the mixed regions that have both urban and rural demand. This results in the so-called *adjusted queuing framework*.

Chapter 4 introduces two new min-rel models that use the adjusted queuing framework from the previous chapter. The first model formulates a *mixed integer program* that finds the absolute optimum for the minimum number of ambulances, such that the required performance is achieved at every point in the ambulance region. A drawback of this model is that it is only applicable for relatively small model instances, because of the complexity and the calculation times. The second model is a heuristic that can provide solutions for larger model instances. We evaluate these models on the basis of real ambulance regions that contain a mixture of both rural and urban subareas. Numerical calculations show that (and how) we can meet the given reliability requirements for each point with fewer ambulances than in the previous models. The results of Chapter 4 are in line with the numbers that are used in practice. Consequently, these are the first min-rel models that can be used in practice.

Chapter 5 focuses on an operational issue, the so-called *dynamic routing*. In addition to achieving the response time of as many emergency calls as possible within the 15 minutes threshold, ambulance services also want each municipality to score well: if a municipality scores very badly, ambulance service providers have to give a good explanation to the regional politicians.

During a relocation, an available ambulance is requested to move to a specific location for the benefit of providing an improved coverage. Various models for dynamic ambulance management (DAM) already exist to determine what vehicle should drive to what location. According to what route this ambulance has to drive to its destination such that it can provide an *as good as possible* coverage while driving, is still an open question in the literature. This might contribute to the performance of (parts of) the region, because the ambulance generates coverage during the driving to the relocation destination station, in particular over the areas it passes. By requiring an ambulance to drive a given route, it may take a little longer for the ambulance to arrive at its destination, but in the meantime it provides a better coverage of (subareas of) the region. In this chapter a method is developed to calculate good alternative routes for ambulances. Then we show to assign a coverage value to a route. By calculating alternative routes for each route to the destination and giving each route a coverage value, an efficient route to the destination can be given to the ambulance. We show that timeliness is distributed more fairly across the region, while the number of journeys that arrive on time does not change. Most local performance gain is achieved in places that are difficult to reach from a base location. Dynamic routing is much faster to implement and cheaper to realize as a solution than building a new base location.

Chapter 6 describes a *pilot study*, which we performed to evaluate DAM policies in practice. In collaboration with ambulance service provider GGD Flevoland, two policies that have been developed within the REPRO research project have been evaluated over a period of twelve weeks. For this study both policies have been adjusted to be able to apply the policies in practice. Furthermore, software has been created and made available for operational use in the dispatch center. During the pilot the fraction of late ambulance arrivals was decreased by a third, making it the first year that the ambulance region reached the national standard of responding in 95% of the high urgency calls within 15 minutes. We also observed qualitative advantages that we had not previously predicted: centralists take faster relocation decisions (especially when there is a high workload in the dispatch center), and the software enables the dispatchers to have a better overview of the available ambulances.

The *testing interface for ambulance research* software, TIFAR, is a software framework that focuses on calculating the implications of decisions in the logistic processes within ambulance care—though simulation and optimization. In Chapter 7 we give an overview of the structure of TIFAR. This can be used to evaluate the consequences of policy changes, such as moving base locations, adjustments in personnel planning, or a replacement of the relocation policy. TIFAR contains classes that can accurately address the behavior of ambulances and the dispatch center, and it has easy-to-use connectors to various databases that contain information about the ambulance region, including information about schedules and historical call record data. Through a link with the Coin-OR optimization software, mathematical models can also be implemented in TIFAR that calculate the optimal base locations. The results of Chapters 2, 4 and 5 were obtained using TIFAR's simulation engine or calculated using its classes and database infrastructure. Furthermore, this framework was used to implement the software of the pilot study as described in Chapter 6.