UNIVERSITY OF TWENTE.

Rijksinstituut voor Volksgezondheid en Milieu
Ministerie van Volksgezondheid, Welzijn en Sport

Optimizing ambulance allocations for A0 emergency calls in The Netherlands

Bachelor Thesis

Industrial Engineering and Management Science

Faculty of Behavioral, Management, and Social Sciences (BMS)

Date
December 13, 2024

Author
Lucas Cuervo Redel
l.cuervoredel@student.utwente.nl

Supervisors
Dr. Ir. Eduardo Lalla (University of Twente)
Dr. Derya Demirtas (University of Twente)
Ir. Geert Jan Kommer (RIVM)

University of Twente
Industrial Engineering and Management
P.O. Box 217
7500AE Enschede, The Netherlands
Tel. +31 (0) 53 489 9111

RIVM
P.O. Box 1
3720BA Bilthoven, The Netherlands
Tel. +31 (0) 88 689 8989

Preface

Dear reader,

In the following pages you will find the result of my bachelor's thesis, which was carried out in collaboration with the RIVM. I have enjoyed tremendously working on such an interesting subject as is EMS management, and I look forward to learning more about it in the future, as I believe that in the uncertainty of the future, the optimization of our resources will always be crucial.

I would like to thank my supervisors, Eduardo Lalla-Ruiz and Geert Jan Kommer, for their support and guidance throughout the thesis period. Your high expectations, good energy, and expertise have pushed me and motivated me to find a path through the noise and uncertainty during the past few months. Additionally, I'd like to thank my second supervisor, Derya Demirtas for taking the time to read my thesis and provide much needed feedback.

I would also like to thank my family, for supporting me throughout these past years as a student, and for never giving up on me. I would not have reached this point without your love. I am grateful to every one of you.

Thank you, Ivanna, for motivating me and showing me what real discipline is, and for being by my side for most of this bumpy journey.

Lastly, I thank all the friends that have been with me throughout these years. Life is about responsibility, but it is also about fun.

I dedicate this work to human knowledge, perseverance, and ingenuity. Above all, I dedicate it to God, as without Him I would not be where I am today.

"Per aspera, ad astra."

Lucas Cuervo Redel

December 2024

Management Summary

In the Dutch EMS system, emergency calls are triaged into several categories based on urgency: A1 (urgent with a target response time of 15 minutes), A2 (non-urgent with a target response time of 30 minutes), and B (primarily for patient transport). Recently, a new category has been introduced, namely the **A0** category, which represents the most critical cases requiring response times of 6 minutes or less. Meeting this response time in 95% of cases is crucial to maintaining acceptable service levels. That is why in this thesis we address the following research question: ***"How can the allocation of ambulances to A0 emergency calls in the Netherlands be optimized to improve service levels?"***

In particular, we focus on high-urgency emergency calls, denoted as **A0**, taking place in the region of Groningen, The Netherlands. The data for this region was provided by the Dutch National Institute for Public Health and the Environment (RIVM). This thesis was conducted in partnership with the RIVM, who sought to comprehend the effect of adding the new **A0** emergency category on the Dutch EMS system.

To answer this research question, the thesis objectives included:

1. Analyzing the existing ambulance demand distribution.
2. Reviewing EMS allocation models in the literature
3. Developing an optimized allocation solution using the Erlang Loss model.

The Erlang Loss model was chosen for its ability to calculate the probability of loss, or unserved calls, based on ambulance demand as well as capacity. An extension to the Erlang Loss model was developed in this thesis to include a prioritization methodology, where a portion of ambulances at a base would be reserved specifically for **A0** calls. The objective of using this solution approach is to assess how the new EMS system can function with limited resources and extreme scenarios, so that better and more robust policies can be developed.

The research methodology consisted of several key phases: synthetic data generation, data analysis, baseline scenario development, solution implementation, and sensitivity analysis. Initial data analysis identified the distinction between high-demand and rural cities within the Groningen region. The baseline scenario (E1) confirmed this distinction by analyzing the current system's ability to respond to A1 and A2 calls without the inclusion of **A0** calls. Then, in the second experiment (E2), **A0** calls were introduced to evaluate the impact of this additional call category on the Dutch EMS system. By reclassifying 6% of A1 calls as **A0**, and maintaining the existing ambulance allocations, the study assessed the strain on service levels through the calculation of loss probabilities for each city and time of the day. The results for this phase of the thesis showed that the introduction of **A0** calls caused a significant increase in loss probabilities for A1 and A2 calls, as ambulances were now prioritized for the urgent **A0** cases. This effect was particularly pronounced in high-demand cities such as Groningen and Veendam, or rural areas such as Uithuizen and Sappemeer. This indicated that these locations require additional ambulances to meet acceptable service levels across all call categories. Lastly, the third experiment (E3) explored the effectiveness of ambulance allocations through three scenarios. The first scenario tested how varying demand levels influenced loss probabilities across emergency call categories. This revealed that especially for A1 and A2 calls, the increase in the demand of EMS lead to higher loss probabilities. The second scenario assessed the impact of reduced ambulance availability, simulating breakdowns or operational disruptions. This analysis identified cities most vulnerable to ambulance shortages and emphasizes the need for a buffer in high-demand areas to

sustain stable service levels. The final scenario focused on varying the number of ambulances reserved for **A0** calls at each base. For instance, reserving one to three ambulances for **A0** calls in Groningen and Veendam was shown to significantly enhance **A0** response rates while keeping A1 and A2 losses within acceptable limits. The following are the overview of results from our experiments:

- Reserving a small portion of ambulances for **A0** calls improves the speed and reliability of high-urgency emergencies by 70% in small cities, and by 20% in big cities.
- Cities with limited ambulance availability are at risk of service delays for lower-priority calls (A1 and A2) when ambulances are reserved for high-priority **A0** emergencies.
- In the event of demand for EMS doubling, assuming that **A0** calls are prioritized, the service level decreases by 50% in small cities, and by 20% in big cities.
- The time of day has a large effect in the demand for EMS, and ambulance service providers should reassess their allocations of ambulances especially for high-demand hours in the midday (between 8am and 4pm).
- Prioritization for **A0** calls can be relaxed during low-demand periods without compromising service quality, allowing more flexibility for handling less urgent cases.
- For the region of Groningen, the following ambulances allocations should be made to improve the success of the Dutch EMS after the implementations of **A0** calls in the country:

City	Total Ambulances Allocated	Reserved for A0
Appingedam	7	2
Groningen	?	9
Leens	3	1
Niebert	3	1
Sappemeer	6	2
Stadskanaal	9	3
Ter Apel	2	1
Uithuizen	6	2
Veendam	10	3
Vlagtwedde	3	1
Winschoten	10	3
Winsum	2	1

The results collectively indicate that the Erlang Loss model, combined with a prioritization-based allocation strategy, enhances the response efficacy for **A0** emergencies while allowing for flexibility in handling A1 and A2 calls. Based on these findings, it is recommended that the RIVM adopt the adjusted allocation strategy for the selected cities, as well as to carry out additional tests on different Dutch cities using our methodology and actual real data from ambulance demand, so that the entire Dutch EMS will be ready to handle the demand with the new **A0** category. Additionally, future research could explore dynamic dispatching strategies that integrate this static allocation model with real-time dispatch adjustments to maximize system efficiency across different regions in the Netherlands.

Table of Contents

1. Introduction

This chapter explains the context of the problem faced by the RIVM (Rijksinstituut voor Volksgezondheid en Milieu). In Section 1.1, a description of the RIVM is given. Section 1.2 introduces the main problem in ambulance allocations for the different emergency call categories in The Netherlands. Finally, Section 1.3 outlines the main research question and sub-questions, as well as the scope and limitations of this thesis.

1.1 Company Description

The RIVM is a research institute of the Kingdom of The Netherlands, which was originally formed in 1909 to study the spread of infectious diseases and develop vaccines [1]. In addition to studying infectious diseases, since 1984 the institute has also been focusing on environmental safety as well as public health. As a national institute, the RIVM falls under the support of the Dutch Ministry of Health, Welfare, and Sport, however, they produce their own independent scientific research which aims at supporting the Dutch society.

The institute carries out various kinds of research which are then used to create recommendations for the Dutch Ministry of Health, Welfare and Sport, as well as external clients and the general public. Any new developments in healthcare, or changes in the status quo of the Dutch medical and environmental systems are analyzed and assessed by the RIVM before being recommended to the general public. For instance, the breakdown of the COVID-19 global pandemic greatly affected The Netherlands, and the RIVM played a crucial role in developing vaccines and studying effective anti-contagion strategies and methodologies [1].

1.2 The Problem

Recently, The Netherlands introduced a new dispatch category for urgent life-threatening situations, denoted as **A0**. This new dispatch category will be used to dispatch ambulances in emergency situations requiring an arrival time of 6 minutes or less [2]. The RIVM has been working on the implementation and supervision of this new category and has asked us to look into possible ways to optimize the allocation of ambulances into this emergency call category.

Before the implementation of this new category, the Dutch Emergency Medical Services (EMS) had three emergency dispatch categories for ambulances:

- A1 calls: life-threatening situations which require an arrival time of at most 15 minutes.
- A2 calls: non-life-threatening situations that require emergency care; an arrival time of most 30 minutes is required.
- B calls: patient transports from one location to another (e.g. from hospital to another).

Based on this description, it may seem as though there is no need for a new category, however, in recent years, the Dutch ambulance system has been facing more deployments and longer response times, leaving patients dissatisfied [2]. This means that resources need to be optimally allocated in order to fulfill the demand for emergency services adequately. That is why the new category **A0** was introduced, so that the Dutch EMS can reach the patients faster when it really matters.

The challenge is now to understand in what way can the system with a shortage in healthcare professionals [2] as well as a limited amount of ambulances [3] can reach patients in less than 6 minutes in 95% of the cases.

1.2.1 Core Problem

The core problem is defined by the RIVM as "the sub-optimal allocation of ambulances to each emergency call category". It is important to first consider why this is a real issue, as many resources and time will be spent on solving this.

When analyzing the issue of low service levels in the Dutch EMS, one may consider the cause to be delays due to traffic, delays due to breakdowns in the ambulances, restocking of medicinal resources on ambulances, unavailability of proper ambulance technology, or misunderstanding of the emergency from the side of the EMS. From all of these possible causes, the one that can be altered, and which can be measured and analyzed is the availability of ambulances.

For instance, if we were to select "inexperienced staff" as a core problem, this is not something which we can affect systematically, or which will have definite effects on the overall service level.

Therefore, we conclude that the most natural core problem for this thesis research is the following:

"Non-Optimal allocation of ambulances to A0 calls"

The norm which the RIVM would like to implement in the Dutch EMS is to have enough ambulances for each emergency call category so that in 95% of the cases, the patients are attended in time. However, currently A1 and A2 calls do not meet this time threshold requirement, which is why the new **A0** category was introduced. Optimizing the allocation to this new category is vital to the future functioning, quality, and reliability of the Dutch EMS.

1.3 Research Design

1.3.1 Main Research Question

After considering the problem context in Section 1.2 and having identified the core problem, we are now able to establish and define the main research question. As previously mentioned, the RIVM is concerned with the increased number of urgent life-threatening calls and how ambulances can improve their service levels through better allocation of vehicles. Therefore, we define the following research question:

"What is the optimal allocation strategy of ambulances to urgent life-threatening A0 calls to maximize the overall service levels of EMS in The Netherlands?"

Attempting to answer the main research question directly is a daunting task. To begin doing so, we must first identify the different components that constitute this question: research sub-questions. By answering these sub-questions, we will be able to have a deeper understanding of what are the important metrics and points of information that affect our core problem as well as our possible solutions.

1.3.2 Research sub-questions

Context Analysis

In order to answer our main research question, we must first understand how the EMS operate in The Netherlands. Once we are informed with the current constraints and variables that are considered for allocating ambulances, we will have a better understanding of the important Key Performance Indicators (KPIs) that are currently being used by the Dutch EMS to allocate ambulances. This information will be used as a baseline for generating a solution at a later stage of our thesis. These are the questions which will be answered in this chapter:

1. How is the dispatch of ambulances currently structured in The Netherlands?
    - 1.1 What is the current arrangement of ambulance dispatch centers in The Netherlands?
    - 1.2 How do ambulance dispatches work in practice?
    - 1.3 What are the KPIs for the ambulance dispatch centers in The Netherlands?

Literature Review

Having gathered information regarding the current functioning of the Dutch EMS with regards to the allocation and dispatching of ambulances, we now want to see which alternative models and strategies exist in the literature. While the initial approach presented by the RIVM was to develop a Linear Programming (LP) model, we consider as well different solution methods for ambulance management in emergency situations. The following questions will be considered:

2. What are the most relevant models and strategies in the literature for managing EMS?
    - 2.1 Which is the solution approach that is most aligned with the Dutch EMS system?

Solution Design

After having selected a solution approach from the literature, we will now work on implementing this approach to optimize the allocation of ambulances in The Netherlands. We should also consider the underlying assumptions and limitations of the approach. These are the questions that will be addressed:

3. How should the solution approach for allocation of ambulances in the different dispatch call categories be modeled?
    - 3.1 Which variables and constraints need to be defined?
    - 3.2 What are the assumptions and limitations of the solution approach?

Solution Evaluation

The fourth part of this thesis is to evaluate the selected model based on the ambulance data provided to us by the RIVM. The goal of this chapter is to assess how the selected model performs on the real-life data. Various experiments with different scenarios will be carried out to evaluate the performance of the model based on the selected KPIs. The following will be answered:

4. What is the performance of the selected solution approach for the optimization of service levels in The Netherlands?
    - 4.1 Which available data will be used to evaluate the resource allocation solution approach?
    - 4.2 Which experiments will be performed to determine the optimality of the resource allocation solution approach?
    - 4.3 How are the selected KPIs improved or changed by the new solution approach?

Conclusion and Recommendations

In the final chapter of this thesis, we draw conclusions and recommendations for the RIVM. The recommendations that we give should be assessed based on metrics of plausibility, as well as usefulness. The following questions are answered:

5. What conclusions and recommendations can be drawn?
    - 5.1 What are the outcomes of the performed experiments?
    - 5.2 What recommendations can be proposed to policymakers and EMS agencies in The Netherlands for implementing the optimized ambulance allocation strategy?
    - 5.3 What future research could be conducted with regards to the allocation of ambulances to emergency calls in The Netherlands?

1.3.4 Scope and Limitations

This thesis aims to determine the optimal allocation of ambulances for **A0** calls in The Netherlands. The research period took place between July 2024 to September 2024. The goal of this thesis paper is to improve the service levels of ambulances assigned to **A0** calls in The Netherlands so that in 95% of the cases, ambulances arrive on time to emergencies. To do this, a resource allocation model is developed. While there are possibly other ways to improve the service level of ambulances (e.g. facility location, better hiring, better equipment, etc.), the scope of this thesis narrows down on a mathematical solution to the problem.

The RIVM expects to receive a model with which they can make more informed decisions on how many ambulances are required for **A0** calls in order to maintain a sufficiently high service level across all emergencies. Additionally, RIVM supervisors mentioned that certain extensions on the mathematical model could be implemented, yet these were not initially required or expected. For instance, one possible implementation is to design a dispatching strategy for the newly allocated ambulances. However, due to the constrained time given for the thesis, this was not undertaken.

For generating the mathematical model, different constraints need to be considered among which are geographical constraints such as distances covered by the ambulances from dispatch centers to emergency locations, or from emergency locations to hospitals. Additionally, the time of the day and day of the week must be considered, since these factors affect both the distribution of demand as well as the supply of EMS calls.

Lastly, initial conversations with the RIVM also suggested that due to the duration of the thesis period, it would be perhaps more efficient to focus on a specific region of The Netherlands or a specific day of the week rather than on the whole country. This change in focus would still allow for useful results while allowing for the timely completion of the thesis. The insights and recommendations from this thesis can then be adapted by other researchers, by generalizing or extrapolating the results to different regions and applications.

2. Context Analysis

In this chapter, the following question will be answered: "*How is the dispatch of ambulances currently structured in The Netherlands?"*. This question is meant to provide a better understanding on the functioning of Dutch ambulance services. Firstly, Section 2.1 gives the background information of how ambulance dispatches function. In Section 2.2, an outline of the Key Performance Indicators (KPIs) considered for evaluating the performance of Dutch ambulances is given. Then, Section 2.3 provides an overview of the capacity model currently used by the RIVM to allocate ambulances in different regions.

2.1 Ambulance Dispatch Process

2.1.1 Call Centers

When the emergency number 112 is dialed, a call center picks up the phone and locates the patient in need. In simple terms, the role of call centers is to help hospitals handle the demand for EMS. Specifically, the call centers play a crucial role in pre-hospital care by managing the incoming calls and directing the correct resources to the location where they are most needed. As can be seen in Figure 3, The Netherlands has a network of 'control centers' or LMS (in Dutch: Landelijke Meldkamers) that take calls for emergency situations [5].

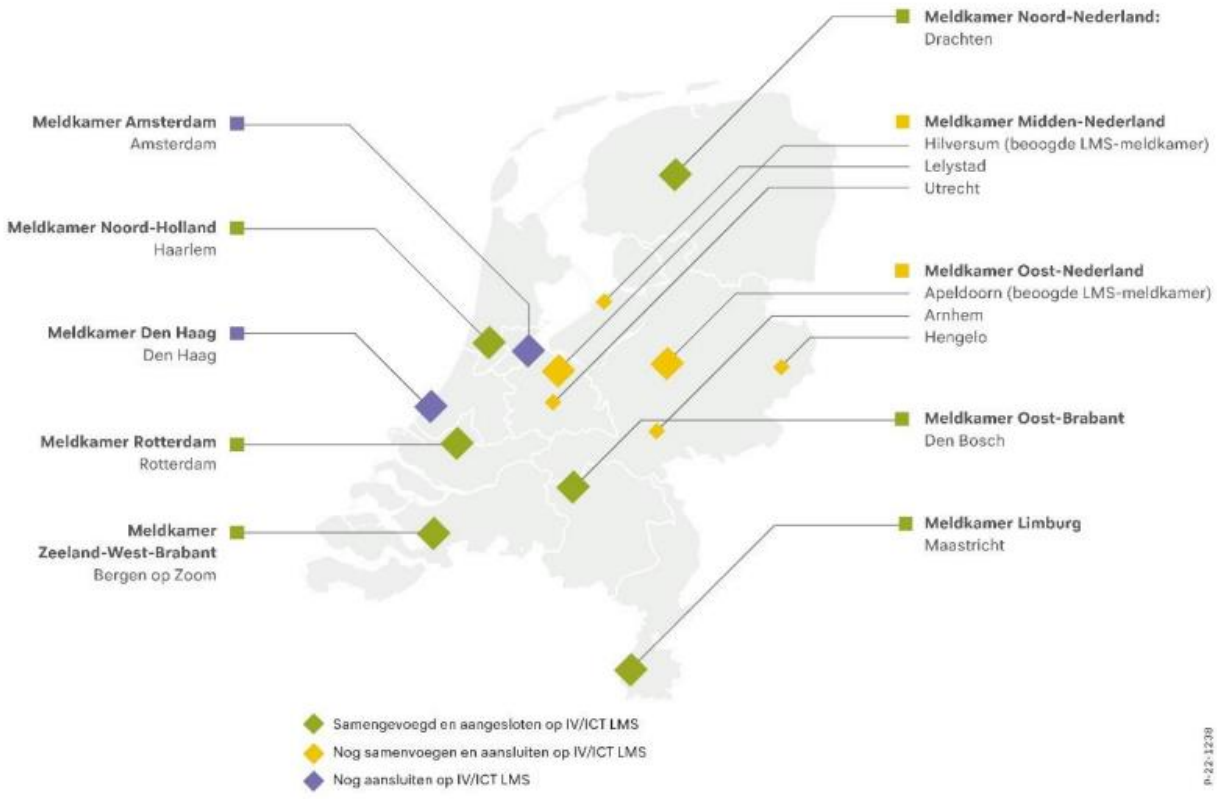


Figure 1: Locations of National Control Rooms (2023) [5]

The current goal of the Dutch ministry of Justice and Security is to allow all the 10 LMS to communicate with one another and take over each other's calls if necessary. For this, the LMS need to be properly informed, having access to the appropriate data at all times. This goal will be finalized in 2027, once the Amsterdam LMS will be connected to the central system [5].

2.1.2 Emergency Regions and Dispatch Centers

If an LMS receives an emergency call which requires an ambulance dispatch, then the local 'regional ambulance service provider' or RAV (in Dutch: Regionale Ambulancevoorziening) is contacted. As seen in Figure 4, The Netherlands has 25 RAVs which are responsible for providing and coordinating ambulance care within their respective regions [9].

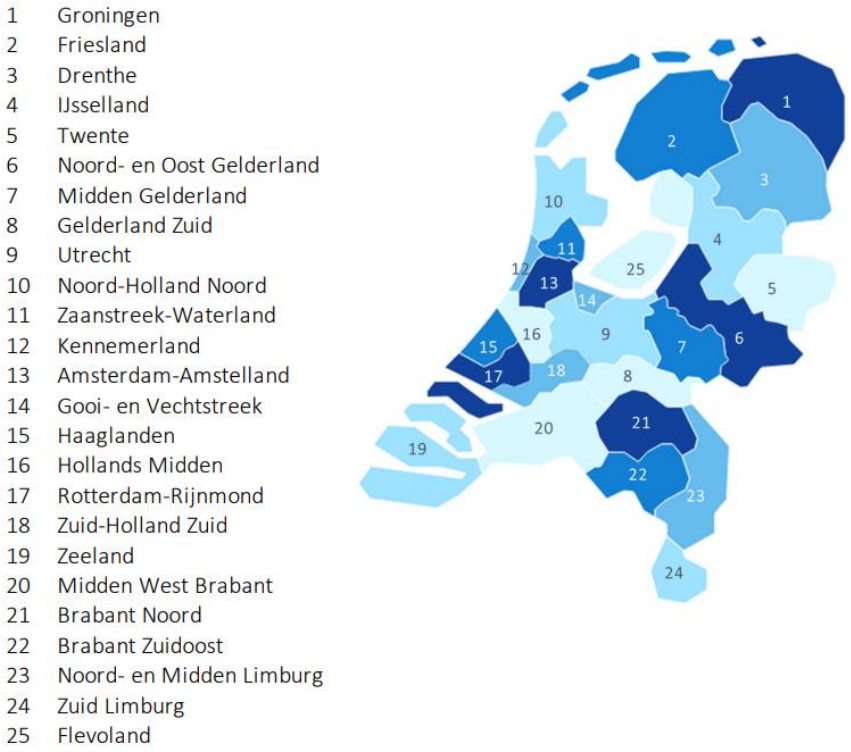


Figure 2: RAV regions of The Netherlands (dark blue: high quality service; light blue: low quality service) [8]

Within each of these RAVs, several ambulance dispatch locations exist (seen in Figure 3). Depending on the location of the emergency, the RAVs can locate the nearest ambulance to arrive as fast as possible to the patient.

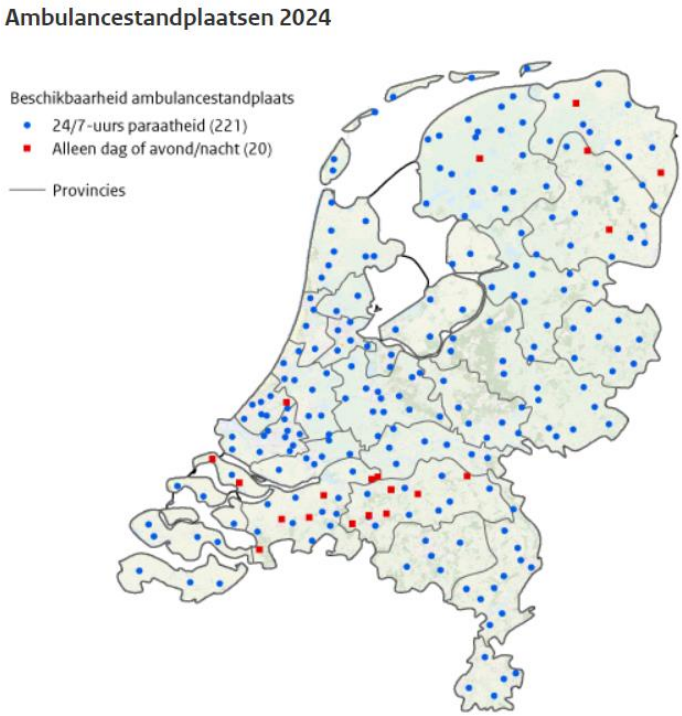


Figure 3: Ambulance dispatch centers 2024 (blue: working 24/7; red: working only during day/night) [7]

2.1.3 Ambulance Dispatch Process

Once the dispatch center has received the emergency information, the dispatch begins. Hence, if the dispatch centers miscalculate the type of emergency, the EMS team will not be able to prepare sufficiently to aid the patient in time. Timing is everything. Figure 4 shows an overview of the different timestamps of the EMS process, from the arrival of a call, to the return of the ambulance to the dispatch center. Additionally, the Service and Response times are specified.

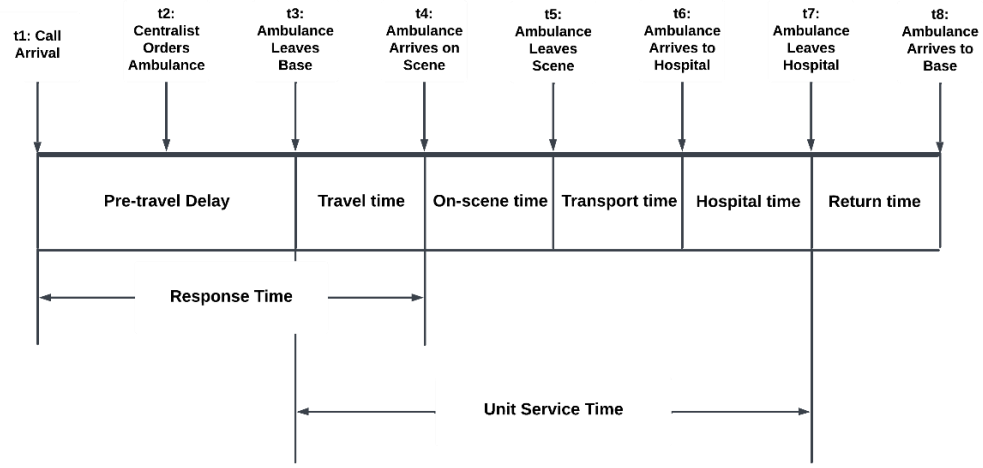


Figure 4: Events and timestamps for an EMS call.

2.2 Key Performance Indicators

Appropriate timing is at the core of EMS. It is exactly for this reason that the RIVM and the Dutch ambulance service providers decided to implement a new emergency call category **A0**, in order to have better control of the timing of emergency vehicles. By having all the resources prepared and ready to deploy before an emergency takes place, the EMS have a higher chance of reaching the patient in time, and therefore saving more lives.

While saving lives is a good metric for success of EMS, talks with the RIVM have informed us that the main KPI that ambulance service providers consider is the time of arrival to an emergency. This is a measurable metric that the EMS staff can control hand affect. Specifically, the general rule in The Netherlands among EMS providers is that ambulances must arrive to the scene of an emergency in 95% of the cases within the predetermined time. Table 1 portrays the time limits for each category.

Table 1: Amount of time for which 95% of cases must be met per emergency category.

Emergency category	Time (min.)
A0	6
A1	15
A2	30

It is worth noting, however, that to generate a solution to our core problem, other variables need to be considered in addition to the main KPI. Such variables are availability of vehicles in a given geographical area, distance to emergency, availability of EMS staff, distance to nearest hospital, demand distribution, or ambulance travel time.

2.3 Conclusion

This Chapter aimed to answer the question *How are ambulances currently allocated to the dispatch centers in The Netherlands?* It was found that The Netherlands is divided in 25 dispatch regions, each with their own ambulance service provider. Additionally, there exist 241 dispatch locations within these 25 regions. The main KPI for the service level of ambulance dispatches has also been defined. Lastly, information has been gathered on the current methods used by the RIVM to select the number of ambulances to assign to each dispatch location.

With this information, it is now clear that the problem of assigning ambulances to different dispatch locations is a complex issue that has many influencing factors. By adding the new emergency call category **A0**, this complexity will increase. The solution approach that is developed in this paper must be able to simplify the process of assigning ambulances, while taking into account this complexity.

3. Literature Review

In this chapter, we address the question: *What are the most relevant models and strategies in the literature for managing emergency medical services (EMS)?* We are interested in assessing which models and solution approaches from the literature are best suited for solving the ambulance allocation problem of the Dutch EMS system. Section 3.1 gives an overview of commonly used models within the EMS management field. Then, Section 3.2 introduces the models which are best suited to solve our problem. Lastly, Section 3.3 provides an overview of our findings for this chapter.

3.1 Models for EMS Management

As previously mentioned, EMS require a coordinated allocation of resources so that ambulances are available across demand regions to minimize response times and maximize coverage. In this section, we introduce the different types of models that exist in the literature for solving common EMS problems. Particularly, models that solve issues in the EMS pipeline such as where the ambulances should be placed, how many ambulances to locate at each location, which ambulance should be dispatched, when to dispatch each ambulance, etc. The models we explore, are categorized into the following types:

- Static and Dynamic models: In static models, ambulances are allocated once and remain fixed to the same location; this means that the ambulance returns to the assigned location after completing service. As such, the uncertainty of real-time conditions is not taken into account. Dynamic models allow ambulances to relocate throughout the day and attend different emergencies in one trip, which optimizes the coverage of the demand at any given period.
- Deterministic and Stochastic models: Deterministic models assume that input parameters are known and fixed, and that once a vehicle is dispatched it is no longer available for other calls. Stochastic models, on the other hand, consider the randomness of call arrivals, fluctuating travel times, or ambulance availability. By using queueing theory, stochastic models represent the real world as a series of queues, with arrival rates, waiting times, and service times. They are able to handle more complex systems where ambulances may not be available all the time, and they help to adapt hourly, daily, or weekly operations.
- Heuristic models: Heuristics are rules that iteratively adjust parameters until a satisfactory solution is found (not necessarily optimal solution). These models give us a balance between solution quality and computational efficiency. Heuristic models are especially useful in situations where exact optimization is either impractical or unnecessary.

3.1.1 Static and Dynamic models

One important aspect of EMS management is the allocation of resources to different locations. What is meant by static and dynamic, is merely the type of deployment of these resources. For instance, in static deployments the vehicles return to their base after completing an emergency call. On the other hand, a dynamic deployment of ambulances is one where the ambulances may fulfill different emergencies one after another, without necessarily arriving to the original dispatch location after completing the calls [11].

Static Models

As discussed by Daskin (1983), static models focus on the optimal placement of ambulances at fixed locations to cover demand points effectively. These models are primarily concerned with maximizing coverage under the assumption that ambulances remain at their assigned locations until dispatched. The previously presented MEXCLP is itself a widely referenced static model that optimizes the placement of ambulances to maximize expected coverage while considering the probability of ambulances being busy. Additionally, another example of static model is the one presented by Restrepo (2008), which considers each deployment base as an 'island' that does not interact in any way with other islands or bases. Of course, these sort of static models are not able to account for much of the uncertainty present in the real world, which is why they are often paired with other more dynamic or stochastic approaches. Nevertheless, as presented by Belanger et al. (2019), many of the existing static models are focused on covering a certain percentage of demand in a given area rather than specific descriptions of how to deploy the ambulances (see also Brotcorne et al., 2003).

Dynamic Models

On the other hand, dynamic models consider the real-time movement and repositioning of ambulances based on data for the current demand and available resources. Gendreau et al. (1997) introduced a dynamic double standard model (DDSM), which optimizes ambulance deployment by dynamically relocating available ambulances to maximize coverage in response to fluctuating demand. Their findings show that this approach is particularly effective in urban environments where demand is variable and unpredictable. However, as explained by Becker et al. (2023), such dynamic models are not perfect either, as they can cause logistic problems by relocating one ambulance to another location while leaving the original point at a higher risk of not meeting the demand. Additionally, such dynamic models become more complex as in the real world there is also a need to consider the scheduling of ambulance staff and adding this to the deployment schedule (Becker et al., 2023). Lastly, as Gendreau et al. (2001) explain, the dynamic dispatching of ambulances must also consider the change in priority of already dispatched vehicles. This means when an ambulance is already deployed for a less urgent emergency, it may be relocated to assist a higher-priority emergency if it is in closer proximity to the given emergency [13]. Overall, it is clear from the literature that a dynamic component is needed in optimizing EMS systems, as this accounts for real scenarios better.

3.1.2 Deterministic and Stochastic models

When we talk about deterministic and stochastic models, we are referring to Facility Location Problems (FLPs). The goal of an FLP is to find the optimal placement of ambulance dispatch centers which would allow for the highest coverage of the demand. While the literature contains several types of FLPs, for instance focusing on survival rather than coverage [see Erkut et al. (2007)) or adding uncertainty in ambulance availability and response times [see Ingolfsson et al. (2008)], we will focus on deterministic models and stochastic models only, as these are most important to understand EMS operations.

Deterministic Models

One of the foundational approaches to resource allocation in EMS are deterministic location problems, which focus on finding the optimal placement of facilities (such as ambulance dispatch centers) to maximize coverage of demand points. What makes these models deterministic is the assumption that all the information necessary for running the models is known in advance, as well as the assumption that there is no uncertainty in the input variables. One such model is the Set Covering Location Problem (SCLP) [10]. The objective is to minimize the number of facilities required to ensure that all demand points are within a specified service range so that they can be reached in the appropriate time limit. The approach assumes that all demand points must be covered, which is why it is a relevant model for EMS providers in high emergency situations such as **A0** calls.

The SCLP was first introduced by Toregas et al. (1971) in the context of fire station facilities. The model's objective is to find the minimum number of facilities needed to cover all demand points within a certain distance. This is in contrast with the classic Maximum Covering Location Problem (MCLP), which aims at maximizing the total demand that is covered within a specific distance [20]. However, one limitation of the SCLP is that it does not consider the dynamic interactions of the real world, specifically it assumes that having a certain amount of facilities will always be enough to cover the demand, but since the demand can differ from the expected value, and the availability of vehicles may change, this solution is not truly realistic. That is where the solution by Daskin (1983) can help us to arrive at a more robust solution. In his paper, Daskin proposes the Maximum Expected Covering Location Problem (MEXCLP), which expands on the SCLP by considering the probability that a facility might be busy when a call comes in. Another interesting change with regards to the SCLP is the consideration of not only the probability of a facility to be busy, but the amount of facilities that are required to fulfill the demand at a certain geographical point. This follows from the fact that the probability of busyness of a location may be too high to cover the demand point, and therefore, there is a need to include a second facility that has a lower busyness to be able to cover the demand.

Stochastic Models

As we have seen, there are scenarios where uncertainty in demand and travel times, as well as in availability of resources, can significantly affect the performance of EMS. Stochastic models provide a more robust approach to solve these situations. Larson (1974) proposed the Hypercube Queueing Model, where using elements of queuing theory and location models, he was able to account for the randomness in call arrivals and ambulance travel times. This model has been extended and applied in various contexts to optimize the dispatching or ambulance systems (see [19] and [21]).

By modeling the ambulance dispatch and relocation system as a network of queues, it is then possible to predict and optimize the performance of the EMS system under different demand conditions. For instance, Marianov and ReVelle (1996), utilized queuing theory to develop the Queueing Maximal Availability Location Problem (Q-MALP) model that minimizes the fraction of lost calls due to ambulances being unavailable by taking into account a system of queues for each ambulance. This model combines a location problem with a queueing approach, which allows them to show the dependence between different servers (ambulances) and how in reality when one cannot arrive to an emergency, another ambulance or EMS provider handles the call. The dispatching of emergency vehicles is then done on a basis of availability of

such vehicles, rather than solely on the geographical demand where the vehicle is located. The Q-MALP is an improved version of the previous MALP and before that the SCLP. It is an improvement because it allows EMS systems to adapt dynamically to real-time conditions, such as fluctuating call volumes or delays due to traffic congestion. By incorporating stochastic elements like call arrival rates and ambulance busy probabilities, it arrives at more efficient dispatching policies.

3.1.3 Heuristic models

As introduced by Gendreau et al. (2001), heuristic methods such as the parallel tabu search can also be used to solve location and allocation problems for a fleet of ambulances. The tabu search algorithm tests different parameters in a system, going from one possible solution to one of the neighboring solutions even when the neighboring solution is not optimal (a neighboring solution being a solution similar to the present one, but with slightly different parameters). This procedure is repeated until making any more changes to a solution does not lead to improvements in the objective function. Church & ReVelle (1975) explain that heuristic solutions can also be used to solve their MCLP model. They mention the Greedy Adding (GA) algorithm, which adds optimal locations one at a time to a 'basket' of locations, until the 'basket' is full. Lastly, Jagtenberg et al. (2015) show that a hybrid static-heuristic solution is effective at generating a solution that keeps time-to-arrival at a minimum while having a set number of ambulances and locations. The benefit of heuristic approaches is their ability to reach a solution in a computationally efficient manner. When the complexity of the problem is excessive, heuristic methods improve a solution stepwise until it reaches optimality.

3.2 Models for Ambulance Allocation

As we have seen, much of the applications for EMS management focuses on locating a facility to be able to care for as many people as possible. In these sorts of solutions, the natural next step is to find a dispatching or relocation of emergency vehicles. However, more importantly is first the determination of how many resources should be located at each base in order to optimally dispatch and relocate the ambulances once the emergency calls start to arrive.

One such model to allocate ambulances is the previously presented Q-MALP by Marianov and ReVelle (1996). This model recognizes that ambulances can be busy, and that increased demand may result in waiting times, which then form queues in the system. For this reason, Q-MALP utilizes M/G/c/c queues, where c represents the number of ambulances at a dispatch center. The objective of the model is to maximize the coverage of demand points while considering the probability that an ambulance will be busy when a call arrives. This is achieved through the busy probability p_j for each facility j. The objective function is then to minimize the expected loss in coverage due to busy servers. Q-MALP is especially useful for regions which have unexpected demands and where some amount of waiting or queuing is accepted. Due to this last point, it is not truly effective for high-urgency emergencies.

The Erlang Loss model is another effective method to allocate ambulances, particularly in situations where delays cannot be tolerated (Restrepo et al., 2009). In this model, the assumption is that if all ambulances at a given dispatch center are busy, any additional calls are "lost" and cannot be served by that dispatch center.

This makes the Erlang Loss model highly applicable to high-priority emergency situations where immediate response is critical. The model uses the Erlang-B formula (1), which calculates the probability of a lost call $E(c, \lambda_b/\mu_b)$, where λ_b is the arrival rate, μ_b is the service rate, and c is the number of ambulances stationed at a dispatch center.

$$E(c, \lambda_b/\mu_b) = \frac{(\lambda/\mu)^c/c!}{\sum_{k=0}^{c}(\lambda/\mu)^k/k!} \qquad (1)$$

The objective of this model is to minimize $E(c, \lambda_b/\mu_b)$ by adjusting the number of ambulances c. Therefore, the model helps determine how many ambulances are needed to ensure that the likelihood of unserved calls remains below an acceptable threshold, making it particularly useful for high-urgency cases like **A0** emergency calls in the Dutch EMS.

3.3 Conclusion

In this Chapter, we explored the question *What are the most relevant resource allocation models and strategies in the literature for managing emergency medical services (EMS)?* The main models for locating emergency facilities as well as for allocating and dispatching emergency vehicles have been discussed. Among the methods that were explored in this chapter, the Erlang Loss model stood out for its ability to consider several variables and constraints of a stochastic or random nature, while at the same time making sure that the service level for high-urgency emergencies remains high.

Given the complexity of the Dutch EMS system, where an increased amount of high-urgency emergencies has led to the creation of the **A0** call category, the demand for such calls will remain unpredictable, and an effective solution must not only handle this randomness but also allow rapid response times under varying conditions. However, before we can tackle this stochastic environment, we must first understand the dynamics of allocating ambulances statically in the Dutch dispatch centers. Since the **A0** category is new, there is no data or understanding of the consequences that this new category will have on the overall EMS. That is why in this thesis we focus on creating the baseline for understanding how static ambulance allocations for the newly formed category (as well as the previous A1 and A2) affects the overall emergency system. Based on the literature review we just conducted, the Erlang Loss model is the most suited to carry out the static allocation of ambulances to bases. We will not use any of the other models such as stochastic or dynamic models since these mostly focus on facility location and dispatching strategies, which are not of use to us for solving the problem of the Dutch EMS. However, the algorithm that we developed in this thesis has heuristic as well as simulation components. A detailed description of the model is given in the next chapter.

4. Solution Design

Having determined the best methods for solving our resource allocation problems in the context of Dutch EMS, we are now ready to explore the following question: *How should the solution approach for allocation of ambulances in the different dispatch call categories be modeled?* In this chapter we explore the Erlang Loss model, which was selected in the previous chapter as the most suitable model. Section 4.1 gives a description of the model's background and how we reached this approach. Then, Section 4.2 gives a detailed overview of the solution model, regarding variables, constraints, as well as objectives. Lastly, Section 4.3 provides an overview of the findings.

4.1 Model Approach

The Erlang Loss model (M/G/c/c) was originally developed in 1917 to study telephone call congestion as a system of queues. Since then, it has been used to optimize the allocation of resources for situations where demand fluctuates, and resources are limited [see 24 and 25].

The Erlang Loss model is a probabilistic queueing model, which can also be used for the problem of the Dutch EMS as it is useful in systems that have a limited capacity of c servers, and no queueing allowed (as only c users are allowed to be in the system at any given time). In this model, each dispatch center b is treated as a separate unit with a fixed number of ambulances c (servers), and the arrival of emergency calls to the dispatch center are considered to follow a Poisson distribution with rate λ_b, while the service times of ambulances from the dispatch center respond to emergencies according to the general distribution with rate μ_b. The key objective is to minimize the loss probability [see formula (1)], which represents the likelihood that all ambulances are busy at the same time and a call cannot be served (and is therefore 'lost').

4.1.1 Model Assumptions

The primary assumption of the model is that the number of ambulances in the system is finite, and that once all ambulances are occupied, any additional emergency calls are lost. This is the reason why the Erlang loss model is particularly useful for determining how many ambulances should be stationed at each dispatch center for high-urgency **A0** emergency calls. That is because if no ambulance is available, and no dispatch happens for that given **A0** emergency call, it is assumed that the patient is 'lost' (dead). In reality, however, it could happen that due to the high urgency of the call, if no ambulances are available at the nearest dispatch center, or no ambulance is dispatched nearby, then a different ambulance agency or emergency service may take over the call.

Then we have the assumptions which are related to queueing theory: first, that **A0** emergency calls arrive according to a Poisson distribution; second, that ambulance service times for successive **A0** calls are independent of one another. The first of these assumptions says that arrivals of **A0** emergency calls are randomly distributed over a time interval while at the same time having a constant average rate of arrivals. This assumption is needed to form the *memoryless property,* which states that the probability of the next event occurring is independent of events that took place before it. In the context of high-urgency **A0** emergencies (and EMS in general), this property is needed, as the probability of an emergency taking place is not dependent on previous emergencies having taken place. The second of these assumptions says that

the service of an emergency should not depend on the time taken to serve previous emergencies. While in practice this may not always be the case (e.g. ambulance may take longer to reach the next call if the previous dispatch call was taking place far away), we make this assumption because we do not consider dispatching or reassignments of ambulances. Since the assignments we make are static, this means that ambulances will return to the base, and the service time of a previous call should not affect the next call.

Another assumption we make is that **A0** emergency calls are only serviced by ambulance vehicles, and not by other types of vehicles such as motorcycles, cars, or helicopters. It is assumed that only one ambulance vehicle is required per call. Additionally, only **A0**, A1, and A2 emergency calls will be considered, as it is assumed that determining the amount of ambulances needed for B calls is trivial.

Lastly, we assume that the proportion of demand of **A0** emergencies that is covered by each dispatch center is known in advance. In reality, the dispatching of the ambulances affects how much of the total demand is covered by each base (e.g. ambulances may not be available at a certain base, and then that base would have a different coverage of the total demand as a result).

4.2 Solution Model

We base ourselves on the work of Restrepo et al. (2009), where they apply the Erlang Loss model to solve an ambulance allocation and dispatching problem within the Alberta EMS in Canada. Our contribution in this thesis is a new prioritization methodology within the Erlang Loss model, which allows for more precise allocations not only based on demand, but also on the type of emergency. While in most EMS around the world the dispatching policy is what determines which ambulance gets sent to a given emergency, by 'reserving' certain vehicles for **A0** ambulances, the EMS system can be better prepared to deal with the randomness of these types of calls. To understand our solution method, consider this: an ambulance dispatched for an A2 call is closer to an **A0** call; the dispatching policy decides to redirect the ambulance assigned to the A2 call, and send it to the **A0** call; assuming that using our model, another ambulance vehicle was 'reserved' for the **A0** call, it can now be sent to the A2 call directly. The description of the model begins with the relevant parameters, followed by the decision variables, the objective function, and lastly the constraints.

4.2.1 Parameters

The following are the most important sets and variables of our problem.

- B: The set of ambulance bases.
- $\lambda_{b,c}$: The arrival rate of calls of type c to base b (calls per unit time).
- $\mu_{b,c}$: The service rate of ambulances from base b assigned to emergency calls of type c (calls per unit time).
- c_b: The maximum capacity of ambulances at base b.
- $E(x_{b,c}, \lambda_b / \mu_b)$: The Erlang loss probability, representing the probability of lost calls.

4.2.2 Decision variables

The decision variables are the unknown information that we aim to determine. For our problem, the following are the two decision variables of our model:

- n_b: The number of ambulances allocated to base b.
- $x_{b,c}$: The number of ambulances at base b reserved for calls of type c.

When compared to the original ambulance allocation model by Restrepo et al. (2009), we added the variable x_b which represents the percentage of ambulances that are reserved for **A0** calls in a given base. As our goal is to make allocations that consider prioritizations of specific vehicles to **A0** calls, this variable helps us to understand the effect that different allocations reserved for **A0** calls have on the overall system.

4.2.3 Model Objective

The objective of this function is to minimize the total expected number of **A0** emergency calls that are lost across all dispatch bases. Each term in the summation represents the expected lost calls for a specific base b, weighted by the arrival rate of calls λ_b at that base. Increasing n_b reduces the probability that all ambulances are busy, which then decreased the expected number of lost calls. Conversely, allocating fewer ambulances increases the chance that an arriving call will not be served due to ambulance unavailability.

$$Minimize \sum_{b \in B} \lambda_b \cdot E(n_b, \lambda_b/\mu_b) \qquad (2.1)$$

4.2.4 Model Requirements

If we were to only focus on minimizing the loss probability, we would simply add as many ambulances as possible to each location (among other things). However, this is far from realistic. We must consider certain real-world constraints which must be met in order for our solution to be viable and useful. The following are the constraints of our solution model, which will help us find a realistic allocation of ambulances to **A0** calls.

- $\sum_{b \in B} n_b = N$ (2.2)
- $n_b \leq c_b \quad \forall b \in B$ (2.3)
- $x_b \leq n_b \quad \forall b \in B$ (2.4)
- $n_b \geq 0 \quad \forall b \in B$ (2.5)
- $n_b \; integer \quad \forall b \in B$ (2.6)

The constraint (2.2) represents the total capacity of the system; for all bases b, the sum of their allocated ambulances must be equal to N. Since there is a limited number of ambulances in The Netherlands, it is necessary to take this into account. Then, constraint (2.3) indicates that the amount of ambulances assigned to base b, must be lower than the maximum capacity of that base. The constraint (2.4) was added by us in this thesis to account for the decision variable x_b, representing the number of allocated ambulances reserved for **A0** calls at a base b. In particular, this constraint says that the amount of ambulances reserved for **A0**, must be equal or lower than the number of ambulances assigned to that base. Lastly, constraint (2.5) indicates the non-negativity of the ambulance assignments, and constraint (2.6) indicates the integrality of the ambulance assignments, meaning that only integer values can be used as variables in our algorithm. However, based on the literature [see 28 and 29], this last constraint can be relaxed to make the results analytically more accurate.

4.3 Example Application of Erlang Loss Model

In order to have a more intuitive understanding of how our Erlang Loss model works, we apply it to a simpler ambulance allocation problem within a city's EMS. This example also serves as an introduction to fundamental concepts such as arrival rates, busy probabilities, and service time, which will later be extended to our problem within the Dutch EMS.

4.3.1 Background Information

A city has two ambulance bases (A and B), and a total of 10 ambulances. The city has asked us to help them deal with the increased influx of high-urgency emergencies. Like our real-life situation, the city's emergency guidelines say that for 95% of the calls, ambulances must arrive to the scene of an emergency in time. Therefore, we must decide the number of ambulances to assign at each base to ensure that this requirement is met. We are given the expected demand (arrival rate) for **A0** calls at each base, and the service rate at which each ambulance can respond to emergencies.

Assumptions

- Arrivals of **A0** emergency calls follow a Poisson distribution.
- Ambulance service times for **A0** calls are exponentially distributed.
- If an ambulance is already responding a call, it is not available for another call.
- Ambulances must be available at least in 95% of the cases for **A0** calls.

Key Variables

- n_b: Number of ambulances assigned to base b. $(b \in \{A, B\})$
- x_b: Number of ambulances reserved for **A0** calls at base b. $(b \in \{A, B\})$

While we have determined above the main variables for the model, we must include certain variables from queueing theory, namely: arrival rate, service rate, and traffic intensity.

- λ_b: Arrival rate of **A0** calls at base b (calls/hour). $(b \in \{A, B\})$
- μ_b: Service rate of **A0** calls at base b (calls/hour). $(b \in \{A, B\})$
- ρ_b: The traffic intensity for base b. $(b \in \{A, B\})$

In queuing theory, the traffic intensity represents the ratio between the arrival rate and the service rate. In the case of this example application, the traffic intensity can be calculated with the following equation:

$$\rho_i = \frac{\lambda_i}{\mu_i}$$

Erlang Loss Probability

The Loss probability formula calculates the chances that all ambulances are unavailable when an **A0** emergency arises:

$$E(x_b, \rho_b) = \frac{\frac{\rho_b^{x_b}}{x_b!}}{\sum_{k=0}^{x_b} \frac{\rho_b^k}{k!}}$$

4.3.2 Calculating Loss Probabilities

In order to optimize how the ambulances are being assigned, we must first assess the current capabilities of the small city's EMS with the current assignment. At the moment, each base is assigned 5 ambulances, as the previous regulators didn't take into account demand patterns. Additionally, we know that the arrival rate of **A0** calls to base A is 2 calls per hour, and the arrival rate of **A0** calls to base B is 1.5 calls per hour.

Baseline input data

Base b	Allocated ambulances, n_b	Ambulances reserved for A0, c_b	Arrival rate, λ_b
A	5	2	2
B	5	2	1.5

If we assume that the service rate μ_b is equivalent to 1 for both bases (meaning that one emergency can be taken care of each hour by each ambulance of base b), then the traffic intensity for each base is equivalent to the arrival rate. Therefore, we have $\rho_A = 2$, $\rho_B = 1.5$.

With the values for the traffic intensity, we can now calculate the loss probability for each base based on the current allocation of ambulances at each base.

- Base A: $E(2,2) = \frac{2}{5} \approx 0.4$
- Base B: $E(2, 1.5) = \frac{1.125}{3.625} \approx 0.31$

These values indicate that with the current allocation of ambulances, both base A and base B would not be able to service **A0** calls in time, for 95% of the cases. We therefore repeat the experiment but without setting a value for how many ambulances to reserve for **A0**. For this new calculation, we assume that the arrival rate and service rate remain the same for each base, we simply focus on finding the optimal number of ambulances such that the Loss probability would be ≤ 0.05. The pseudo-code below illustrates how the model finds the correct allocation quantity:

```
begin
    Set target_loss_prob = 0.05 // Target loss probability threshold
    Set c = 1 // Initial number of ambulances to test
    Calculate ρ for each of the bases // Traffic intensity based on arrival rate lambda and service rate mu
    while True do:
    begin
        Calculate loss_prob with E(c, ρ) // Calculate the Erlang Loss probability for c ambulances
        if loss_prob < target_loss_prob then:
            Output "Optimal number of ambulances is {c}"
            Output "Loss probability: {loss_prob}"
            Exit loop // Desired service level achieved
        elseif c > max_ambulances:
            Output "Maximum ambulances reached. Minimum achievable loss probability: {loss_prob}"
            Exit loop // Maximum constraint reached
        increase variable c: c = c + 1 // Test with one more ambulance
    end
end
```

After applying the algorithm with the parameters $\rho_A = 2$, $\rho_B = 1.5$, and $E(x_b, \rho_b)$ set to 0.05, we get the following results for our example application:

- Base A: Reserve 5 ambulances for **A0** emergencies.
- Base B: Reserve 4 ambulances for **A0** emergencies.

4.3.3 Results from the Example Application

The example application we just presented shows the fundamentals of the Erlang Loss model in a simpler and more intuitive context. We have also shown that the Erlang loss model does give us insight into the allocation of resources, and therefore is a proven way to manage the capacity of scarce resources in times of emergencies. In Chapter 5 of this thesis, we will take a better look at the model and whether it is also an efficient solution approach for the Dutch EMS.

The main limitation of this example is that the service rate is assumed to be the same for all ambulances. In practice, high-urgency **A0** emergencies may have higher service rates per unit of time. Additionally, the model assumes that ambulance availability and the number of **A0** calls are independent of each other, this may be relaxed when applied to the real world.

4.4 Conclusion

Through this chapter, we have concretized the solution approach that we will use to improve the allocation of ambulances in The Netherlands. In particular, our aim was to answer the following research question: *How should the solution approach for the allocation of ambulances in different dispatch call categories be modeled?* We defined the Erlang loss model, as it provides an effective way for minimizing the expected number of lost calls for high-urgency **A0** emergencies.

As part of forming a solution, we were able to identify and define the most relevant variables and constraints. We introduced the new variable x_b —which indicates the proportion of ambulances at a given base that are reserved for **A0** calls— and constraints that ensure the model is based on realistic conditions.

It is worth noting that the Erlang loss model makes certain assumptions and limitations, including the assumption of exponential service times and that each call is independent, with no queuing allowed. Additionally, one of the main limitations of our model is that it does not account for real-time dynamic factors like demand variability, since the demand around a given dispatch location is assumed to be known in advance. As portrayed by the example application, it is nevertheless a good method from a tactical point of view for static allocation of resources in emergency situations. Further experiments will have to prove whether this is also the case with real ambulance data.

5. Evaluation

We have now defined our solution approach for solving the allocation problem in the Dutch EMS presented by the RIVM. However, we first have to assess whether this solution is indeed useful to the real world, or whether it just "looks good on paper". In this chapter we address the following research question: *What is the performance of the selected solution approach for the optimization of service levels in The Netherlands?* In Section 5.1, we give an overview of the data used for testing our model. Section 5.2 deals with the experiments that we undertake. Lastly, Section 5.3 provides a summary of our findings.

5.1 Data Analysis

5.1.1 Data Context

In order to test our model's performance, data was provided by the RIVM on the ambulance dispatching in the RAV 1: Groningen. As seen in Figure 7, the region of Groningen, managed by Ambulancezorg Groningen, contains 12 dispatch centers and uses the call center in the town of Drachten (region Friesland) to coordinate all its emergency calls. Currently, the region operates 35 ambulances for urgent emergencies, which are equipped with both Advanced Life Support (ALS) and Basic Life Support (BLS) [see 31]. ALS and BLS are terms used to refer to the type of equipment present in the ambulance vehicle, which in turn determines the types of emergencies that the EMS team aboard the ambulance can handle.

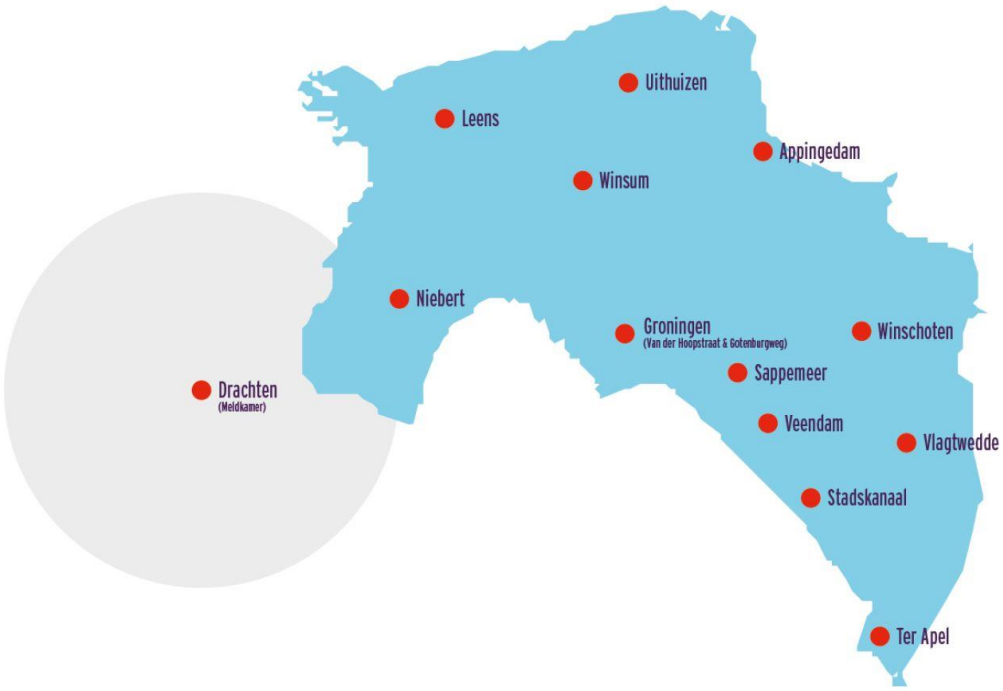


Figure 7: The 13 dispatch centers in RAV 1: Groningen (call center in Drachten). [30]

5.1.2 Data Understanding

Due to data privacy concerns, the RIVM was only able to share a sample of the ambulance data from the region of Groningen. The sampled data only contains 25 calls from 11 cities/towns in the region of Groningen. This amount of data is not significant enough to generate generalizable results, as it is likely that bias or skewness can take place. Therefore, a synthetic dataset has to be generated based on the population density of the region of Groningen. With the data from AllChartsInfo (2024), the daily arrival rates of urgent emergencies can be calculated for each city.

Table 2: Overview of population and arrival of urgent emergencies in Groningen dispatch locations.

City	Population	Daily arrival rate (A1 & A2 calls)
Groningen	211,120	43.7
Veendam	21,615	4.47
Stadskanaal	19,005	3.93
Winschoten	18,750	3.88
Appingedam	10,895	2.25
Ter Apel	9,847	2.04
Sappemeer	8,330	1.72
Winsum	7,445	1.54
Uithuizen	5,470	1.13
Vlagtvedde	3,185	0.66
Leens	1,680	0.35
Niebert	638	0.13

As seen in Table 2, the only 12 cities and towns considered are the ones where ambulance dispatch centers are located. While this decision constrains the validity of the results (as it does not consider the demand in the whole region), it allows us to reach more relevant outcomes for **A0** calls, as realistically only ambulances which are stationed 6 minutes away or less to an emergency are able to attend to it in time. For rural areas, a different type of solution is needed to be able to reach patients of **A0** calls in time (e.g. adding dispatch locations or using different types of vehicles).

The arrival rates are calculated for each city based on the annual per capita arrival rate in the Groningen region: 0.0755. This value indicates that each person in Groningen accounts for about 8% of the urgent emergency calls in a year. Then, to get the daily values for each city, we multiply it by the respective population and divide by 365 days. Based on the current arrival rates, the dataset assigns A1 and A2 calls to the different cities. After simulating a scenario with the given data, the results are portrayed in Figure 8. Here, the number of ambulance dispatches that take place at each city on a given day are represented (for A1 and A2 calls, as we do not yet want to introduce the **A0** calls). It is important to note that Figure 8 is an example, and therefore running the simulation again may lead to slightly different results. Additionally, Figure 8 is meant to showcase the degree to which the demand for EMS at the different cities differs.

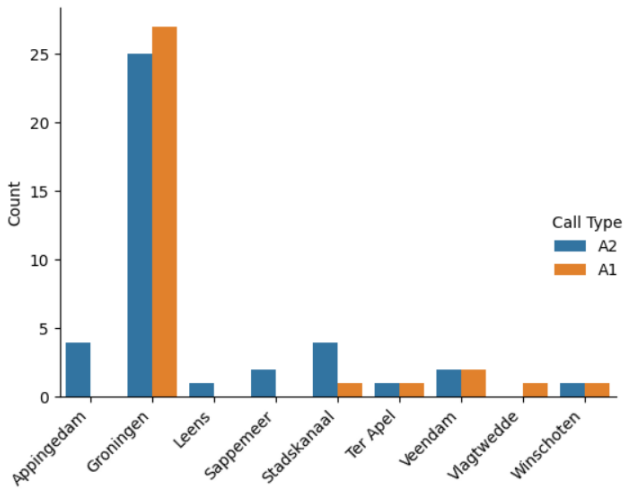


Figure 8: Example of simulation results.

While running this simulation may lead to slightly different results, Figure 8 indicates that on this given day Niebert, Leens, Uithuizen, Winsum, and Vlagtwedde have too few calls arriving per day, meaning that there is a chance that in these locations no calls will arrive during a given day for either or both of the call categories. Aside from this observation, we see that Groningen has much of the calls in the region, due to the large population density. Lastly, in terms of total calls, there are more A1 calls taking place than A2 calls. The proportion of A1 and A2 ambulance calls assigned to each city is determined from the Reference Framework 2023 [32]. Out of the 45,004 urgent emergencies that took place in the Groningen region in the year 2023, 24,831 were A1 and 20,173 were A2. Therefore 55.2% of emergencies in Groningen are A1 and 44.8% are A2. Lastly, we take a look at the distribution of the 'time to arrival', the main KPI of most EMS providers.

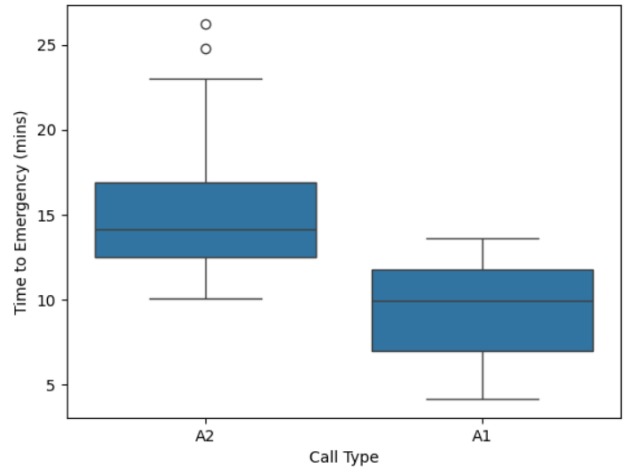


Figure 9: Distribution of 'time to emergency' for A1 and A2 calls.

As seen in Figure 9, the generated data of the current scenario is adhering to the guidelines for each type of call, namely that A1 calls should be serviced within 15 minutes, and that A2 calls should be serviced within 30 minutes. These values were determined based on results mentioned in the Reference Framework (2023), where it is mentioned that in recent years, the amount of short and urgent calls increased, while at the same time mentioning that the average time for an entire ambulance journey is about 70 minutes. The journey

time includes the entire duration of an ambulance's travel for a single incident, beginning from the moment it departs for the emergency scene until it returns to the station. Lastly, for the generation of the 'time to emergency' data, we also consider the fact that 99% of people can be serviced within 12 minutes in the region of Groningen [32]. For the aforementioned reasons, we take as the mean for the A1 and A2 calls, values of 10 and 15 minutes respectively.

For generating the synthetic data on the arrival of emergency calls across the cities, the Poisson distribution is used. As it was mentioned previously, since each call is an independent event and the likelihood of multiple calls happening simultaneously is low, the Poisson distribution can help us to generate real-world fluctuations in demand without requiring additional complexity. While other distributions such as the Normal or Uniform could be used, they are not suitable because they assume a regularity in call arrivals, which does not conform with the reality of EMS in The Netherlands. It is worth noting that to make the results repeatable, the same seed is used for generating the data.

5.2 Numerical Experiments

To assess the effectiveness of our model at solving the problem faced by the Dutch EMS, we carry out several experiments using the data we have generated. The goal of these experiments is not only to find the best ambulance allocation at each base, but to also carry out a sensitivity analysis of the possible scenarios present in daily operations of EMS. First, an overview of our experiments is given, followed by an explanation of what is done in each of them, how they are set-up, and what the results are.

5.2.1 Experimental Framework

As seen in Table 3, each individual experiment has a different goal, and together, they are meant to evaluate the model's performance in different scenarios. Particularly, we focus on two main experimental variables: amount of ambulances allocated to each base, and amount of ambulances reserved for **A0** at each base. To assess these changes, we compare how effective they are at improving the baseline situation in terms of the loss probability.

Table 3: Experimental Design.

Experiment ID	Title	Goal
E1	Baseline	Establish a baseline service level by calculating the arrival rates, required ambulances, and loss probabilities for A1 and A2 calls only.
E2	Addition of A0	Evaluate the impact of adding **A0** calls to the data, by assessing loss probabilities and allocations.
E3	Sensitivity Analysis	Perform sensitivity analysis by simulating various real-world scenarios (increased demand, ambulance availability, changes in of **A0** reservations) to observe their impact on model's performance and ambulance allocations.

The first experiment, E1, looks at the baseline service level (loss probability) in the region of Groningen. For this, **A0** calls are not included in the data, as the objective is to evaluate the current system's performance, and currently **A0** calls are not present. With this initial experiment we are able to assess how call arrivals are handled currently. A total of 30 ambulances are considered in the system, as recommended by the Reference Framework for Ambulance care [32]. Additionally, within the Groningen region, only a selected 12 cities are taken for the experiments, in particular the cities containing a dispatch center. This decision is taken as any other location would not realistically be able to reach patients in the required 6 minutes (for **A0** calls).

In experiment E2, **A0** calls are introduced to the system in order to evaluate the effect that this has on the service levels for all emergency categories (**A0**, A1, and A2). The same procedure as in E1 is followed, however, 6% of A1 calls are reclassified as **A0** calls. This percentage is an estimation given by the RIVM on their current understanding of the system. Additionally, this percentage only applies to A1 calls, as these are the calls most likely to be considered **A0** in the current system. By comparing the results from this experiment with the Baseline experiment, we can assess how the addition of **A0** calls in the system influences the demand on the ambulance system and whether there is a need to adjust the number of ambulances allocated to each base.

Lastly, in experiment E3, several sensitivity analyses are carried out to examine the system under different conditions. For this, various factors are assessed, such as ambulance availability, or demand fluctuations. The following sub-experiments are carried out:

- Reduced Ambulance Availability: We decrease the total number of available ambulances for A0 calls, to represent possible shortages (e.g., due to breakdowns or maintenance) and examine how this affects service levels.
- Increased Demand Scenario: We artificially increase the number of calls to represent a surge in demand, potentially due to seasonal or event-based fluctuations, and assess how the system handles higher volumes.

Through these sensitivity analyses, we aim to understand the flexibility of the new EMS system with the addition of **A0** calls. These insights provide a foundation for making effective policy recommendations.

5.2.2 E1: Baseline

In our first experiment, we calculate the baseline service level of the EMS to have a benchmark to which we can compare later results. As we mentioned previously, for this first experiment we do not want to test the optimal value of x_b (the number of ambulances reserved for **A0** at a base), as there are no **A0** ambulances present currently. Therefore, we take $x_b = 0$. Additionally, as recommended in the 2023 Dutch Ambulance Reference Framework, 30 ambulances are enough to service the entire region of Groningen. Based on a database of the ambulances in The Netherlands [31], we can find the current allocations of ambulances to each city (see Table 4).

Table 4: Ambulance allocations by city.

City	Current Ambulance Allocations
Appingedam	3
Groningen	9
Leens	1
Niebert	2
Sappemeer	2
Stadskanaal	4
Ter Apel	1
Uithuizen	1
Veendam	1
Vlagtwedde	1
Winschoten	4
Winsum	1

In Figure 10 and Figure 11, the results of running the baseline experiment can be observed. The current loss probabilities are portrayed in three periods: from 00:00 until 08:00, from 08:00 until 16:00, and from 16:00 until 24:00. By dividing the day into sections, we are able to more accurately simulate the ambulance assignments throughout the day.

Figure 10: A1 Loss probabilities for each city based on current allocations in E1.

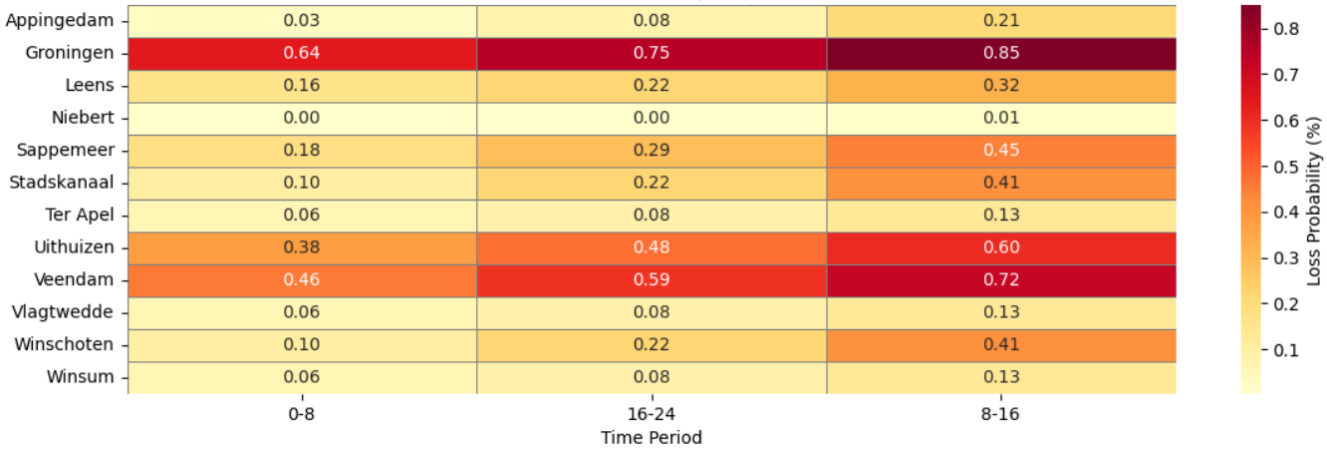
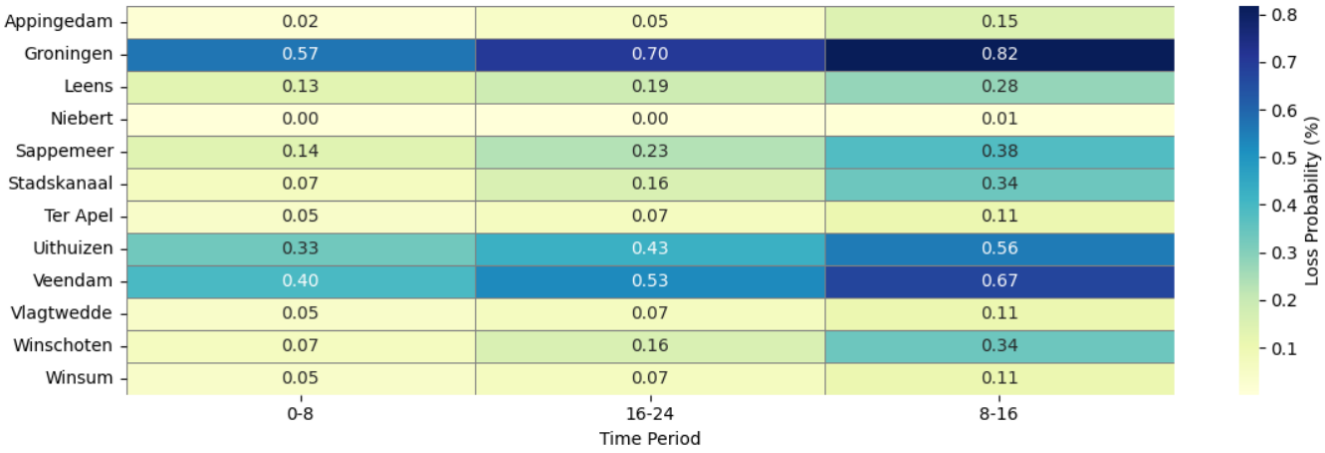


Figure 11: A2 Loss probabilities for each city based on current allocations in E1.



The results from the baseline experiment are clear: the current system is underperforming, and in most cities, calls are not attended to within the specified time threshold. Particularly, we see that even though Groningen has the highest assignment of ambulances, it nevertheless arrives late to calls in 50-80% of the time. Additionally, Uithuizen and Veendam struggle to attend to calls with their current assignment of ambulances.

Lastly, the highest demands of EMS takes place in the midday (08:00-16:00). This is because at this time of the day, most of the people are awake, and it is more common for accidents to take place. That is not to say that accidents don't take place at night or early in the morning, but when accidents take place at home, and no one is around to watch, ambulances are less often called. Therefore, that concludes why the loss probabilities are lower from 00:00 to 08:00.

5.2.3 E2: Addition of A0

In this second experiment, we introduce the new emergency call category **A0**. Based on RIVM's understanding, 6% of A1 calls can be reclassified as **A0** to represent urgent cases needing faster response. This information is then used to generate the corresponding arrival rates for each period of the day. With the arrival rates, the number of ambulances that must be placed in each city is calculated. The objective of this experiment is to assess the system's performance to the addition of **A0** calls, keeping all other constants from the Baseline experiment constant.

Figure 12: A0 Loss probabilities for each city based on current allocations in E2.

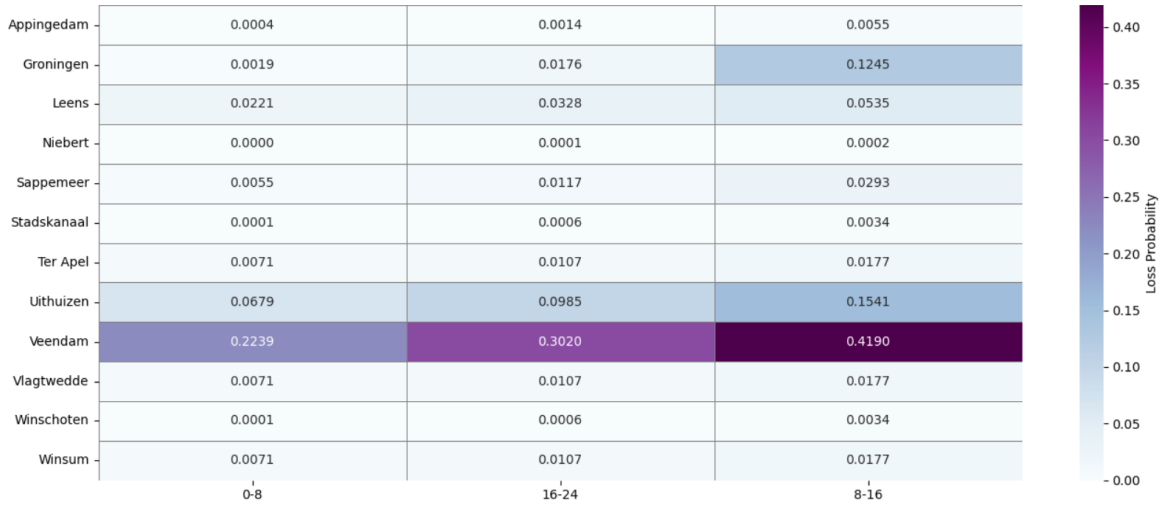


Figure 13: A1 Loss probabilities for each city based on current allocations in E2.

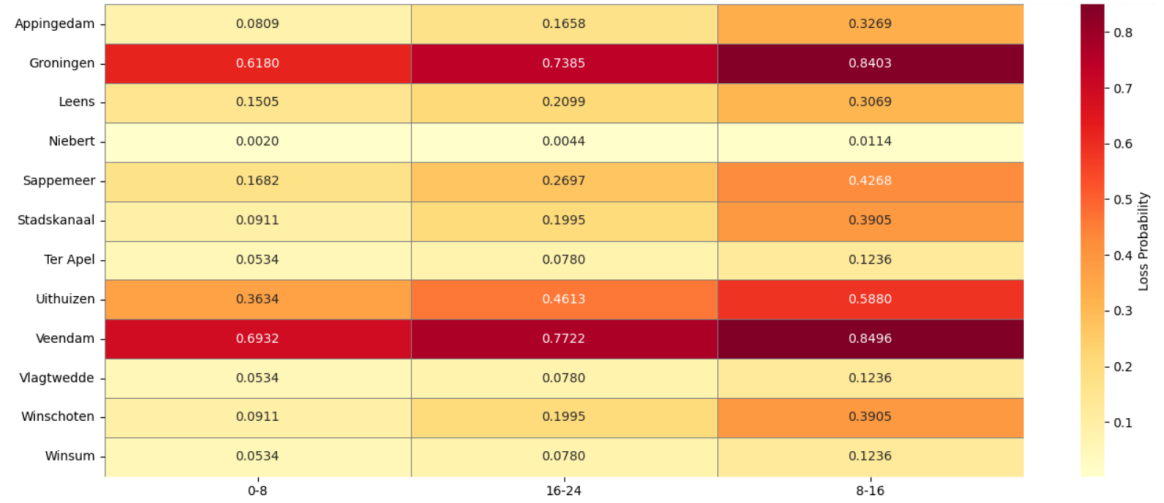
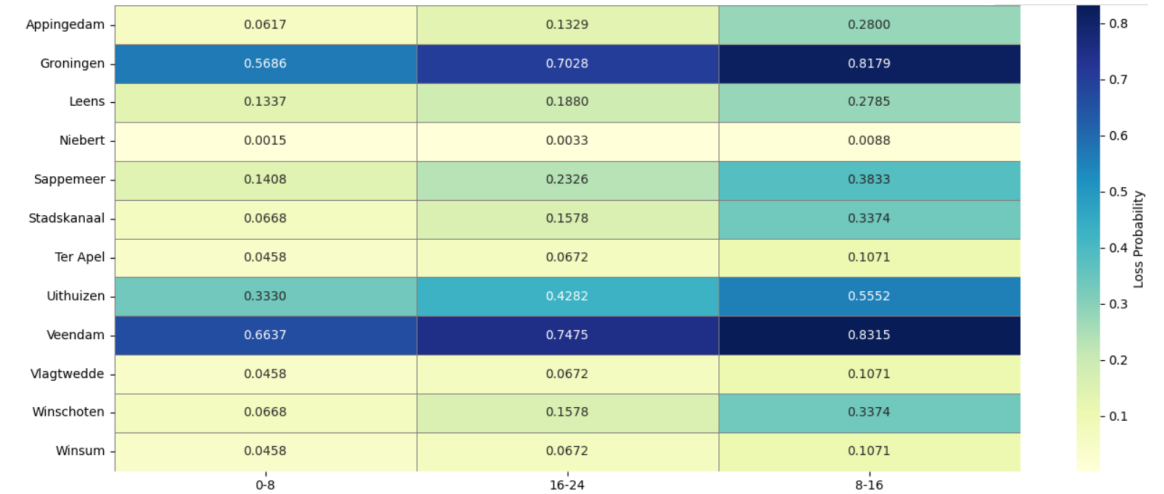


Figure 14: A2 Loss probabilities for each city based on current allocations in E2.



As Figure 12 shows, in the experiment **A0** calls are successfully added to the system. The loss probability for the **A0** calls is below 5% in most cities except for Veendam and Uithuizen (as well as Groningen during the midday). This again indicates that for both of these cities, the current allocation of ambulances is not optimal, and likely more ambulances are needed. In addition, the increase in loss probabilities during the midday in Groningen, indicates once again that this is a busy period of the day, and hence more ambulances should be available during this time of day. These results can also be seen from Figures 13 and 14, in which both Veendam and Uithuizen cannot properly attend to A1 and A2 calls with the current assignment of ambulances. Therefore, it is not an isolated event taking place only in **A0** calls.

The addition of **A0** calls leads to an increased loss probability for A1 and A2 calls as a result of the urgent nature of **A0** calls, which requires the model to prioritize them. Additionally, since there are a fixed number of ambulances at each base, when we reclassify 6% of A1 calls as **A0** calls, we reduce the availability of ambulances for A1 and A2.

5.2.4 E3: Sensitivity Analysis

In this last experiment, the objective is to analyze the sensitivity of the Dutch EMS to different operational constraints and scenarios with regards to the addition of **A0** emergency calls. We explore the following three scenarios: first, we assess how the system performs with varying demands. Then, in our second scenario we evaluate the impact of changing ambulance availability, in order to simulate breakdowns in ambulance vehicles. Lastly, we assess how different values of x_b affect the loss probabilities at each of the cities in the Groningen region.

Scenario 1: Increased demand

In this scenario, we assess how the system performs under increased demand. By applying a multiplier to the daily arrival rates for **A0**, A1, and A2 calls, we simulate a scenario in which the Dutch EMS system experiences a significant surge in call volume. While it is clear from E2 that most of the cities in the region of Groningen require more ambulances, in this scenario we maintain the current allocations across the cities and assess how increasing the demand across all call types will affect the EMS system's performance. Therefore, this serves as a way for policymakers at the RIVM to better understand the way in which future variations in demand may affect the overall EMS system.

We test different variations in the demand, namely: 0.5, 1.2, 1.5, and 2.0. These values represent how much the demand (arrival rates) for each call category is increased or reduced. After running these variations, we generate visualizations to assess the performance of the different cities at different times of the day.

Figure 15: Loss probability in each city for different changes in demand (00:00 – 08:00am).

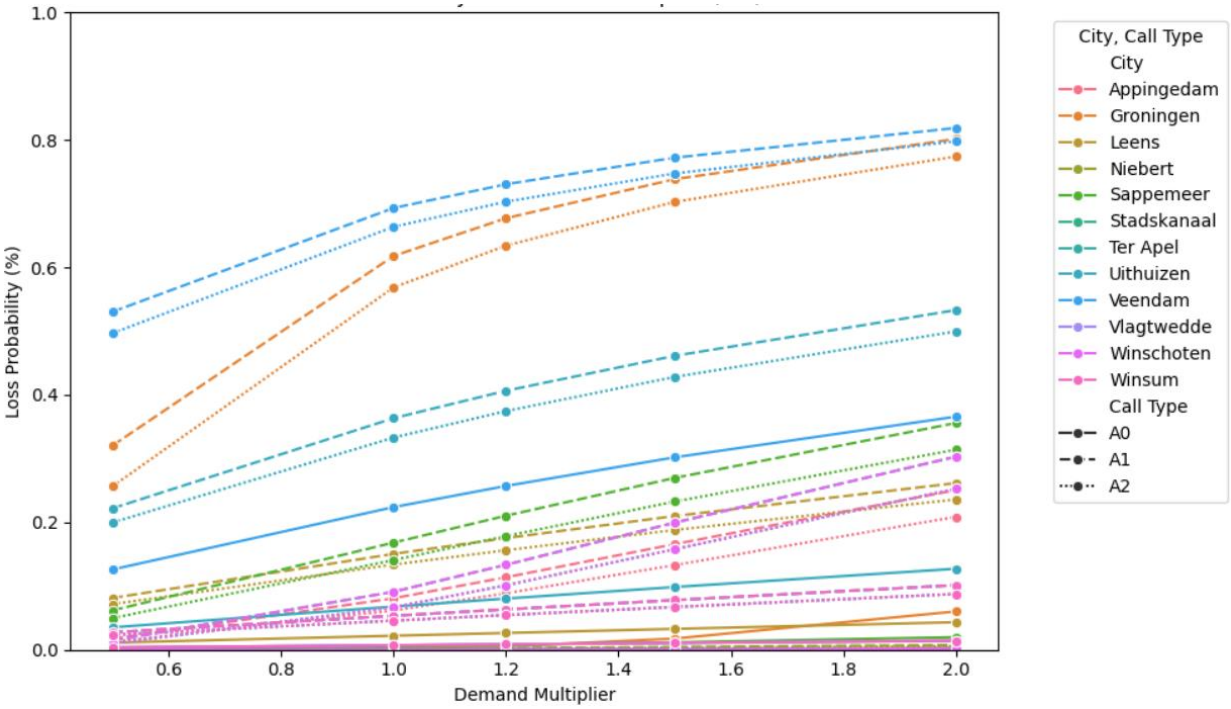


Figure 16: Loss probability in each city for different changes in demand (08:00 -16:00).

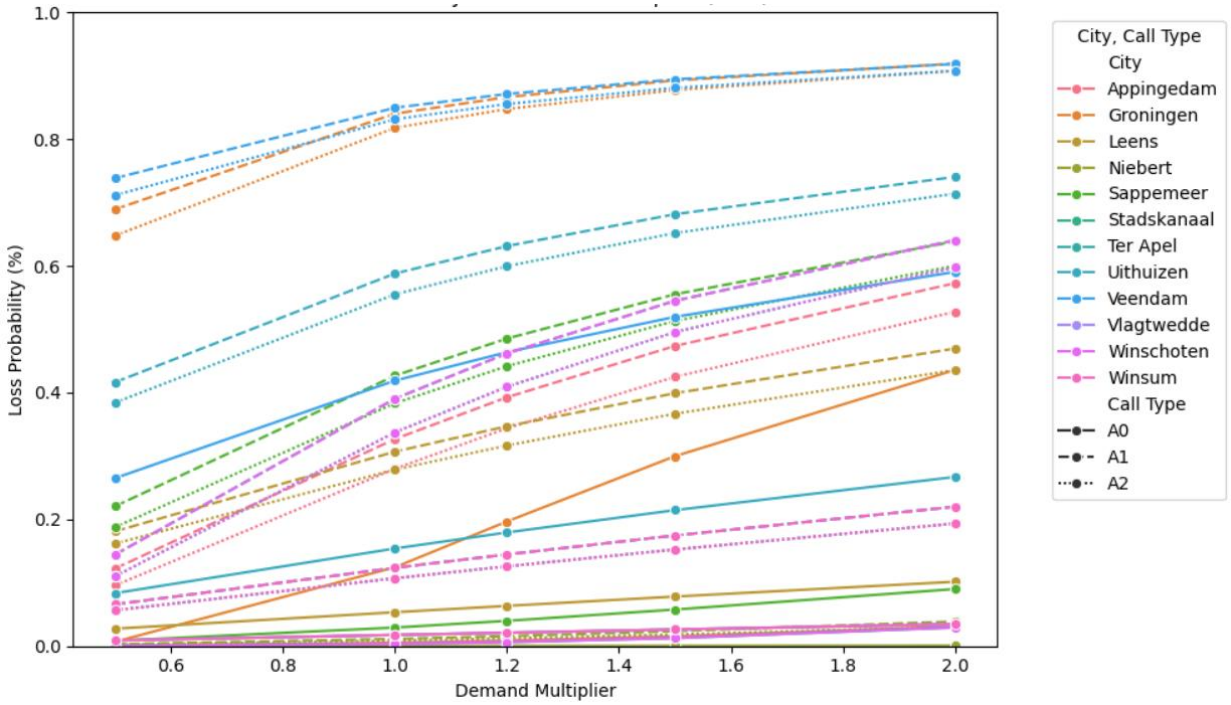
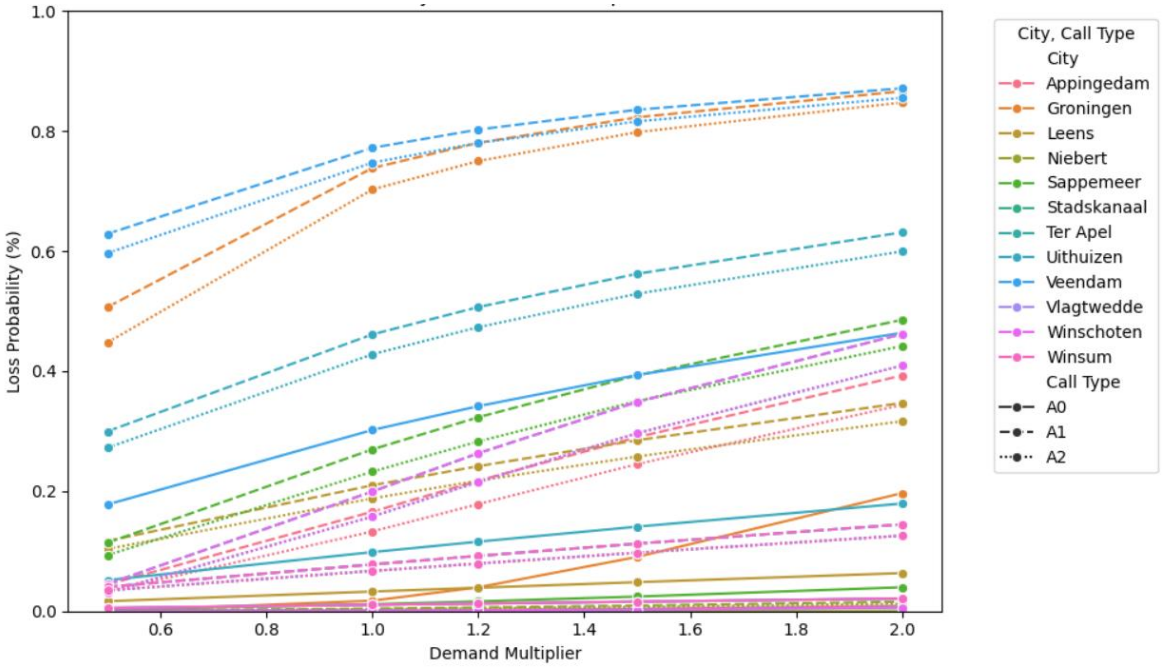


Figure 17: Loss probability in each city for different changes in demand (16:00 – 24:00).



In Figures 15-17, the loss probability for the different cities is mapped out with the varying demands. Across all periods (morning, afternoon, and evening), a consistent trend can be observed, namely that increasing demand leads to higher loss probabilities. This is to be expected, as all systems have their breaking points, and in the case of EMS systems, this is represented as a lack of resources. The effect produced by the increased demand on the loss probability is particularly significant in Groningen, Uithuizen, and Veendam. The reason for this can be attributed to the baseline demand already being high, resulting in elevated loss probabilities even at lower demand multipliers. Additionally, as it was noted in experiment E2, ensuring a rapid response for the most urgent cases (**A0** calls), leads to A1 and A2 calls experiencing noticeable increases in loss probability. In Figure 17, for instance, it is noticeable how all the solid lines (**A0** calls) have much lower loss probabilities than the dotted lines (A1 and A2).

In addition to the differences that arise as a result of the different call categories, the time of day has a noticeable effect on the loss probabilities as well. The morning period (0-8) generally shows lower loss probabilities across all cities and call types. Although the evening period (16-24) shows a slight reduction in loss probabilities compared to the midday, the system remains stressed especially for A1 and A2 calls. These results suggest that the current ambulance allocations are insufficient to handle A1 and A2 calls in high-traffic areas, particularly during midday.

Scenario 2: Ambulance availability

The second scenario explores the impact of reduced ambulance availability. We simulate how each city would react to shortages or surpluses in ambulance availability. By changing the ambulance allocation in each city, we can evaluate how the system's service levels for **A0**, A1, and A2 calls are affected, and ultimately, reach the best allocation of ambulances for each city.

Figure 18: Average daily loss probability for different ambulance allocations to A0 calls in each city

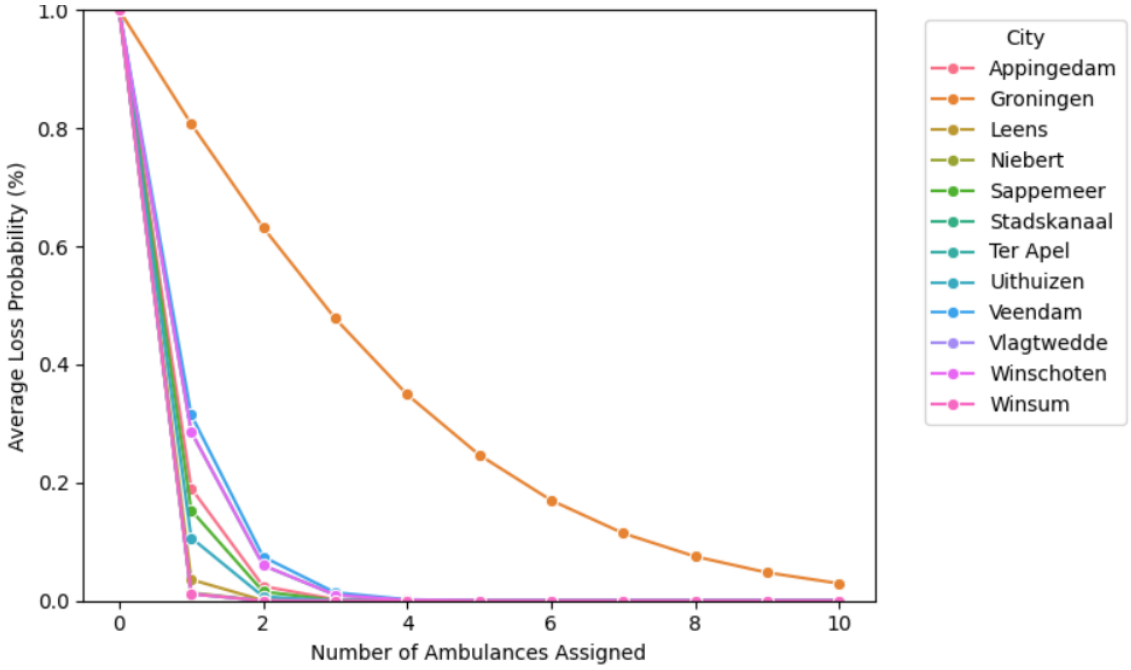


Figure 19: Average daily loss probability for different ambulance allocations to A1 calls in each city

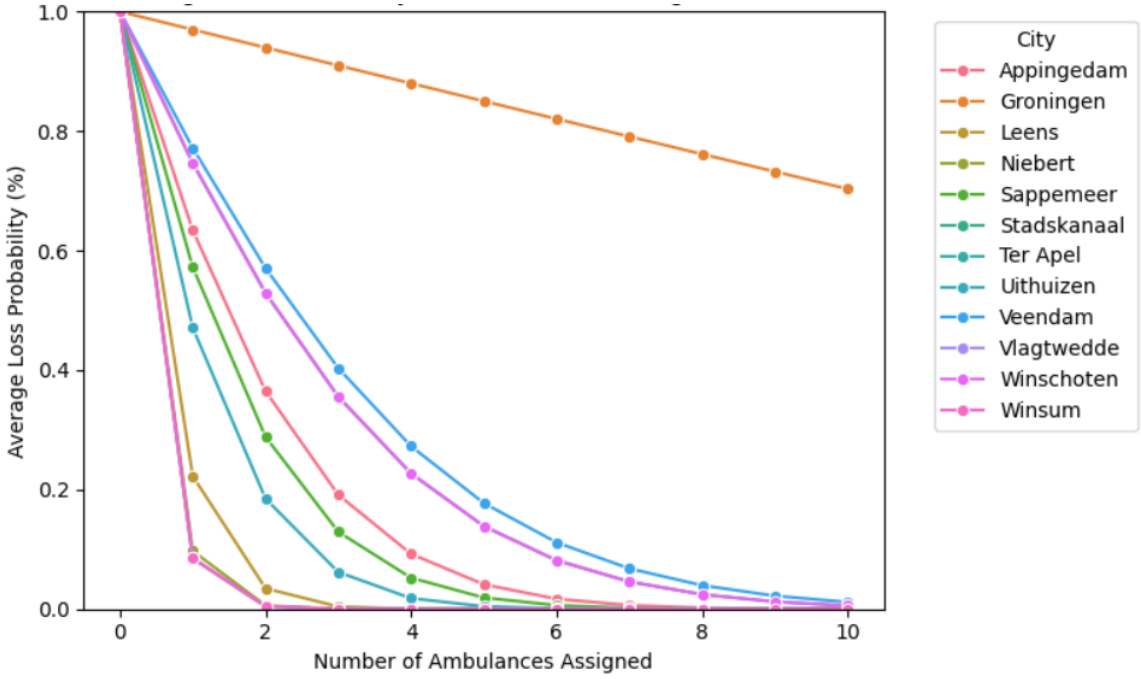
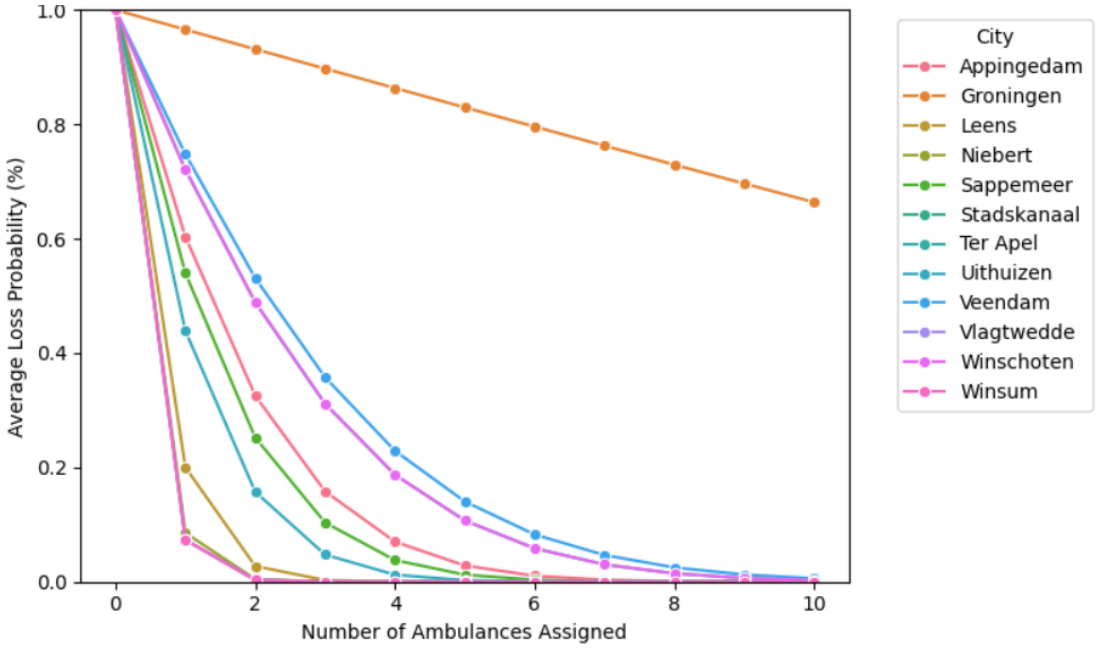


Figure 20: Average daily loss probability for different ambulance allocations to A2 calls in each city



To generate the Figures 18-20, we use the average arrival rate for the day (from morning, midday, and evening). While using this value generalizes to some extent the fluctuations in real-time demand of EMS, it gives a good understanding of the needs for each city. Naturally, for each call type the loss probability decreases significantly as the number of ambulances increases. However, the rate of this decrease varies by city. For instance, Groningen always seems to require more ambulances than the rest of the cities in the region. This is a sensible conclusion, as Groningen has the largest population. Nevertheless, a loss probability of less than 5% can be reached in Groningen for **A0** calls, with 9 ambulances being reserved for these types of calls. On the other hand, the loss probability of A1 and A2 calls in Groningen does not reach a good result even after assigning 10 ambulances to the city's base. In contrast, other cities such as Winsum, Leens, and Niebert have steep declining loss probabilities with only 1 or 2 ambulances being required.

It is also clear that the **A0** calls require little allocations to be able to reach acceptable loss probabilities. Based on Figure 18, most cities can have a single ambulance reserved for **A0** calls, and this would already be enough to reach a loss probability of less than 5%. However, the same cannot be said about reaching acceptable loss probabilities for A1 and A2 calls, where more ambulances are needed.

Scenario 3: Varying ambulance prioritizations for A0 calls

In this last scenario, we investigate how the model responds to variations in x_b, the number of ambulances reserved for **A0** calls at each base. We simulate with the current ambulance allocations to each city, how many of the available ambulances should be prioritized for **A0** calls, in order to assess the extent to which giving priority to these calls affects the loss probabilities for the other types of calls.

Figure 21: A0 Loss probabilities for every city, with varying reserved ambulances for A0.

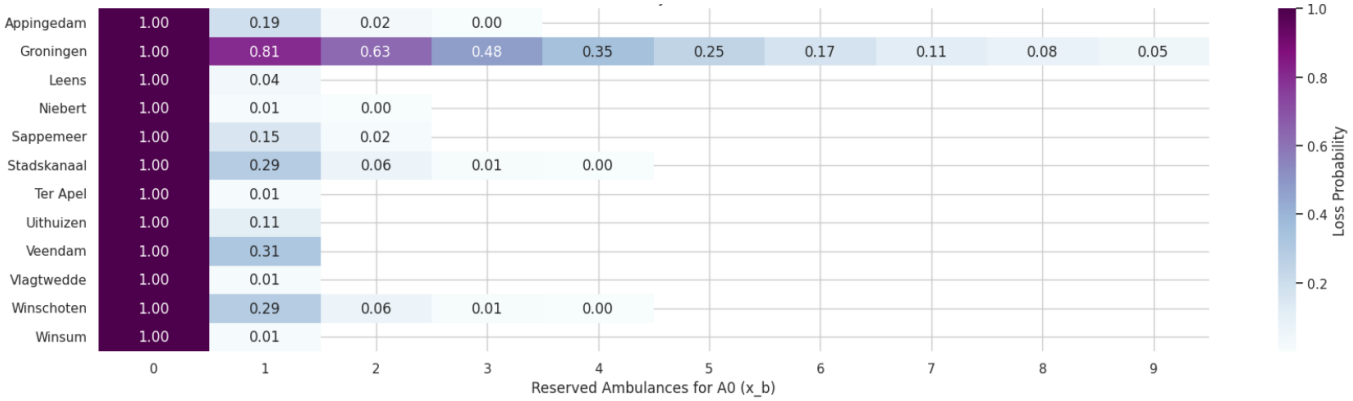


Figure 22: A1 Loss probabilities for every city, with varying reserved ambulances for A0.

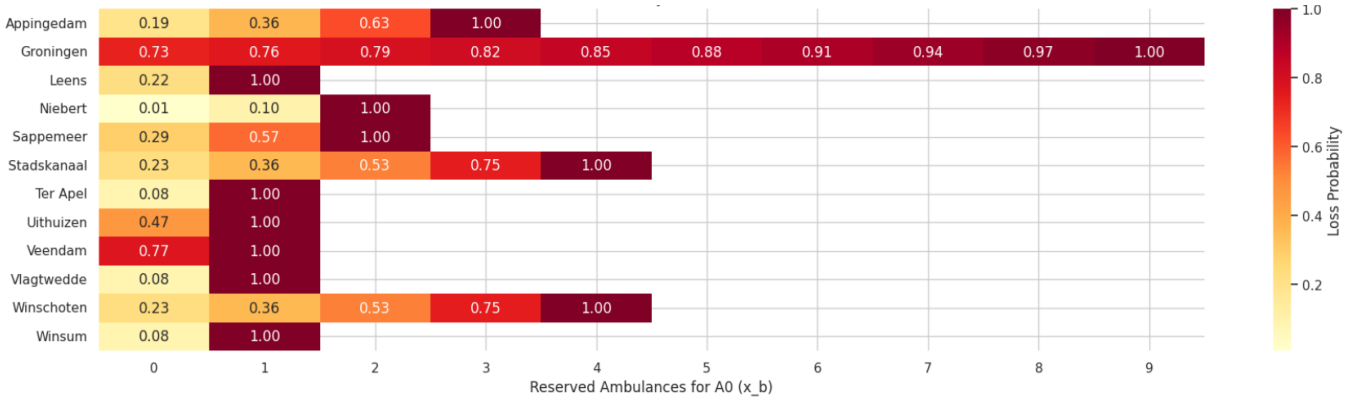
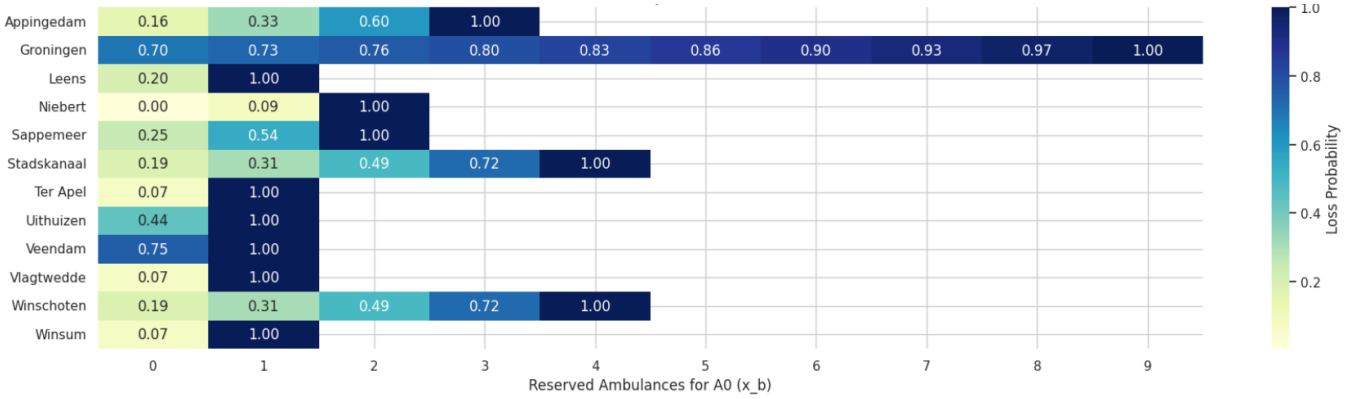


Figure 23: A2 Loss probabilities for every city, with varying reserved ambulances for A0.



In Figures 21-23 we can see the outputs of running scenario 3. Reserving ambulances specifically for **A0** calls has a notable impact in the performance of A1 and A2 calls. The results indicate that assigning even a single ambulances to **A0** calls in most cities, can significantly improve the service level for these high-priority emergencies.

However, in high-demand cities like Groningen, more ambulances need to be allocated to achieve similar improvements in **A0** service levels. Particularly, for the **A0** loss probability in Groningen to be below 5%, at least 9 ambulances must be reserved for these types of calls. With the current allocation of ambulances, this would mean that A1 and A2 calls in Groningen would not be able to be attended in time in 100% of the cases (assuming static ambulance deployments). Therefore, for the city of Groningen more ambulances have to be assigned, or more realistically, a dynamic approach has to be taken. The same is true for other smaller cities which only have one ambulance, as when that ambulance is reserved for **A0** calls, the A1 and A2 calls cannot be serviced in time.

5.3 Conclusion

Having completed the model design as well as data collection, our goal in this chapter has been to answer the question "*What is the performance of the selected solution approach for the optimization of service levels in The Netherlands?"*. After presenting the synthetic data that we generated for the 12 cities in the region of Groningen, three experiments were carried out to evaluate how the solution approach performs with respect to the current situation in the region of Groningen. We considered different allocations of ambulances to the dispatch call categories and found the optimal amount of ambulances that need to be reserved for **A0** calls in each of the region's cities.

In the first experiment we conducted, we assessed the current loss probability for each of the 12 selected cities. We simulated the arrival of A1 and A2 calls, without introducing **A0** calls. This gave us insight into the distribution of the demand in the cities, and which periods of a given day had the highest demand for EMS. Then, in the second experiment we introduced **A0** calls into the system, as recommended by the RIVM (reclassifying 6% of A1 calls as **A0**). The addition of the **A0** calls allowed us to understand better whether the current ambulance allocations are sufficient to handle the demand in the different cities for each of the call categories, and for each period of the day. In particular, we saw that during the midday, most cities have a harder time to deal with the A1 and A2 calls after the addition of **A0** calls. This is because less ambulances are available. Additionally, we observe how some cities with small populations, such as Veendam and Uithuizen, struggle to reach loss probabilities lower than 5%. This indicates that the current allocation of ambulances at these cities is not appropriate. Lastly, the third experiment involved different extreme scenarios to assess how the current system performs. In particular, as the demand for EMS augments, less critical calls (A1 and A2) perform worse, especially for cities with larger populations during the midday. This again indicates that more ambulances are needed. However, as the third experiment showcases, the increase in ambulances does not need to be large. For instance in Leens, Niebert, or Winsum, an increase of 1 ambulance would already be enough to handle most **A0**, A1, and A2 calls in time. The minimum number of ambulances required per city to keep loss probabilities at acceptable levels under varying demand conditions were determined. Additionally, in E3 a target number of ambulances for each city, as well as strategies to maximize efficiency by reallocating resources to better meet service demands across **A0**, A1, and A2 calls were determined.

6. Conclusion, Limitations, and Recommendations

This chapter marks the conclusion of the thesis. Our goal approaching this work, was to answer the question *"What is the optimal allocation strategy of ambulances to urgent life-threatening A0 calls to maximize the overall service levels of EMS in the Netherlands?"* We have developed an algorithm that is able to answer this question, and through various experiments, we investigated different configurations of model parameters to simulate real-world scenarios. In this chapter we answer the final question: *What conclusions and recommendations can be drawn?* Particularly, in Section 6.1 we address the final conclusions of this thesis; in Section 6.2 we review the limitations of our research as well as the thesis paper; then in Section 6.3 we provide recommendations for policymakers in The Netherlands; finally, in Section 6.4 we give ideas for future research that can be conducted based on our thesis.

6.1 Conclusion

This thesis has investigated the optimal allocation of ambulance vehicles in The Netherlands. The specific focus has been on high-urgency **A0** emergency calls, which have been recently implemented in the Dutch EMS. Through a series of experiments centered on the Groningen region, we tested how different demands and prioritizations affected the system, with insights not only on the number of ambulances required to meet a target response time for 95% of emergencies but also on how many ambulances should be specifically reserved for the **A0** call category.

The approach taken in this thesis focuses on the static allocation of ambulances, done in the tactical side of EMS planning. This means that the demands for the different cities are not known with exactitude when assigning ambulances to bases, and that dynamic relocations are not considered. Therefore, after an ambulance is finished with an emergency, it must return to the same base from which it was deployed. In contrast, most EMS allow ambulances which are closer to an emergency to be dispatched, especially if the emergency is highly urgent, and regardless of whether or not the ambulance was already deployed for another call or returning to base. Based on this, it may seem that the solution methodology chosen for this thesis is not a realistic way of assessing EMS responses, however, the decision to make prioritization before dispatching was done to assess the effectiveness of this method to the real world, as the introduction of the new call category could provide an opportunity for policymakers to reevaluate the way in which they carry out their operations.

To be able to generate significant and realistic results, the initial phase of this thesis focused on understanding the baseline performance of the EMS system in Groningen, using a sample of real data provided by the RIVM. This data, however, was limited, and initial findings revealed that the model tended to overfit, hindering the reliability of any results. To address this, a data augmentation process using random sampling was implemented to generate synthetic data that maintained the distributional characteristics of the original dataset but was broader in scale. While the initial goal of the study was to develop a solution applicable across an entire region, focusing on the cities with dispatch locations within them lead to more realistic results, as **A0** calls would not be reached within 6 minutes otherwise. If data were generated for the entire region, and not just these 12 cities, the results for loss probabilities would have been worse overall for **A0** calls. Highly urgent calls taking place outside of built-up areas will require a different solution

methodology, perhaps involving helicopters, CPR trainings for the local population, improved allocation of defibrillators, or relocations of ambulance bases.

In experiment E1, we learned that the current allocation of ambulances was able to meet the response goals in most cities, based on the generated demand distributions. However, some cities (particularly Groningen, Veendam, and Uithuizen) struggled to maintain acceptable service levels due to elevated demand. Daily variability was also identified, meaning that different parts of the day lead to significant different needs for ambulance allocations.

Experiment E2 introduced the **A0** calls into the system, representing a new tier of high-priority emergencies. Results showed a notable decrease in service levels for A1 and A2 calls as more ambulances were diverted to handle **A0** calls. This happened as a result of bases having a strict number of ambulances, since we did not consider the addition of ambulance vehicles for this experiment. It is worth mentioning that to solve this problem in our solution, we simply have to add more ambulances to the relevant bases. In reality, a solution could be to use dynamic dispatching to redirect the ambulances that are already on the road towards more pressing emergencies.

For the last experiment, E3, three scenarios were designed to test the system's robustness under different conditions. In the first scenario, variations in demand were simulated to observe how fluctuations in call volumes would impact the service levels for **A0**, A1, and A2 calls. The results showed that while high-demand areas could absorb moderate increases without severe impacts on service levels, significant demand surges led to rapid declines in response times, particularly for A1 and A2 calls. The second scenario in E3 examined the effects of reduced ambulance availability, simulating potential vehicle breakdowns or other operational disruptions. This scenario identified the cities and call types most vulnerable to shortages, which confirmed that high-demand areas require a buffer in the number of ambulance vehicles to maintain stability under stressed situations. The third scenario for E3 evaluated the best prioritization strategy for ambulances assigned to **A0** calls, for each of the bases. This analysis suggested that setting aside even a small number of ambulances exclusively for **A0** calls could improve response times for the majority of the cities without severely compromising A1 and A2 call response levels. Below is the resulting static allocation recommendations for the region of Groningen, which allow for all of the **A0**, A1, and A2 calls to have a loss probability below 5% (except for Groningen city, which due to its high demand, requires a dynamic ambulance allocation to be able to attend all patients with fewer resources).

Table 5: Recommended allocations by city

City	Total Ambulances Allocated	Reserved for A0
Appingedam	7	2
Groningen	?	9
Leens	3	1
Niebert	3	1
Sappemeer	6	2
Stadskanaal	9	3
Ter Apel	2	1
Uithuizen	6	2
Veendam	10	3
Vlagtwedde	3	1
Winschoten	10	3
Winsum	2	1

Overall, while the Dutch EMS system in the region of Groningen can be adjusted to prioritize **A0** calls effectively, doing so without expanding resources significantly impacts response levels for A1 and A2 calls. As Table 5 shows, the proactive assignment model we have tested provides a foundation for achieving higher service levels, but this will have to be coupled with effective dispatching strategies and alternate solutions to the difficult calls outside of built-up areas and especially in the city of Groningen. Lastly, the experiments we conducted provide a detailed understanding of how changes in demand, resource constraints, and prioritization policies interact, offering policymakers valuable insights for refining ambulance deployment strategies in densely populated areas such as The Netherlands.

6.2 Limitations

The main limitation of our chosen model arises from the assumption that once ambulances are stationed in a base, they stay in that location until a call arrives. This fundamental assumption of our model limits its ability to generate an effective solution in real-world circumstances. This is because the Erlang Loss model does not account for real-time fluctuations in demand, and instead assumes that the demand is known in advance, as well as the assumption that different dispatch locations or ambulance service providers do not interact with each other. In reality, part of the dispatching of ambulances occurs as a result of dynamic reallocations from demand fluctuations, which send ambulances from locations where the demand is low at the moment, to locations where there is a higher chance of a call taking place. Additionally, in high-urgency **A0** emergency calls where there is a need to arrive quickly to the emergency site, if a dispatch center is not able to provide assistance in time, other ambulance providers will intervene to help the patient in need.

While it is true that dynamic dispatching strategies may help to improve the response times of our allocation model, the core issue lies in the forecasting of the demand for high-urgency **A0** emergencies. If the demand for a given region is not well known, then the allocation of ambulances to that region will never be truly optimal. Our results are therefore limited by the quality of the data provided to us, as well as the data that we have generated. Particularly, our model assumes that 50% of calls take place during the midday, 20% in the morning, and 30% in the evening. If this assumption is wrong, the entirety of our results could change and become irrelevant to the real-life scenario. Optimally, to generate significant results, real data would be needed on the ambulance dispatches and their demand. Additionally, we were limited to calculating the allocation of ambulances to different dispatch call categories based on governmental reports from 2023, which may not be as relevant for the current year of 2024. Since the Erlang Loss model relies on knowing the demand in advance, a demand estimation based on insufficient or wrong data, leads to sub-optimal results.

Another limitation of our solution approach is that we do not include the number of allocations per ambulance base. The reason for this is partly due to the lack of data given to us, but also due to the inefficacy of the Erlang Loss model to do this, since it assumes independence between bases, and in reality, and effective solution for allocating ambulances to different bases should take into account these dependencies. Additionally, another limitation of our solution approach is that the service rate is assumed to be the same for all ambulances, regardless of the type of emergency. In practice, certain emergencies (such as **A0**) might require faster or more complex responses, which could affect the service rate.

6.3 Recommendations

In this section, we outline the final short-term and long-term recommendations to policymakers and/or stakeholders of the Dutch EMS and specifically the ambulance sector.

Short Term

Based on our findings, policymakers at the RIVM should explore how the implementation of **A0** calls affects the overall EMS performance in different regions of The Netherlands, specifically using the prioritization methodology we present. Additionally, a larger dataset of real ambulance demand should be used to assess whether our findings still hold. In particular, we would recommend starting to collect more ambulance deployment data to generate accurate demand forecasts, as our solution approach relies on demand being present beforehand. Having a more accurate demand will provide insight into the model's effectiveness to unpredicted scenarios and identify any necessary adjustments. It would be interesting for the RIVM to also explore the effectiveness of our model on weekly and monthly data, as we have only explored the effect that the introduction of **A0** calls has on daily loss probabilities.

Long Term

Over an extended period of time, we would recommend the RIVM to generate a dynamic allocation strategy which implements the **A0** calls, so that a simulation of strategic deployments can be assessed across the entire Dutch EMS. Our model could be used in the tactical level to form the baseline of optimal ambulance allocations before the dispatching takes place. Then, within the simulation, running some dispatching strategies and assessing whether the initial placement of ambulances helps the EMS to reach calls in time. Of course, the dynamic simulation would be also beneficial to test whether the assumption of the Erlang Loss model of bases not communicating between each other is indeed realistic.

Lastly, as it was highlighted in our results, many of the small cities were not able to handle properly the calls within rural or isolated cities. Therefore, we recommend assessing different solutions such as helicopter or drone dispatches, the use of cars or motorcycles to reach A0 calls in particular, or the redistribution of ambulance bases or defibrillator stations.

6.4 Further Research

1. Look at the possibility to include dynamic elements in finding the optimal allocation of ambulances in bases. This would be done to test whether the selected allocations of ambulances are indeed optimal during the real-time dispatching of ambulances and how the dispatches interact with other ambulance bases, or even different dispatching prioritizations based on the type of emergency.
2. Explore the effectiveness of using different vehicle types to reach the **A0** emergencies, such as with helicopter, bicycle, motorcycle, or even car. As **A0** calls require a fast response time, perhaps the use of different, more agile, vehicles can lead to better results in rural areas.
3. Demand forecasting is of high importance, as the model developed rests upon de assumption that the demand is known in advance. While this topic could have been undertaken by us, we chose to dive into the allocation of ambulance vehicles to assess whether this simplistic approach could already make any changes in the service levels. Yet, the results we received are dependent on the quality of the data. As the saying goes: "Garbage in = garbage out".

4. With a better understanding of the demand for A0 calls, it would be also interesting to see where the real ‘hubs’ of high-urgency emergencies are located. Perhaps there is a need to centralize the ambulance locations, or alternatively, smaller locations dispersed across the region or city. Related to this would be the analysis of stationing ambulances on stand-by not in dispatch centers, but in specific spots of a city near high-demand streets or neighborhoods, to see whether this allows for faster response times.

Bibliography

[1] RIVM (2024). *About RIVM*. https://www.rivm.nl/en/about-rivm

[2] AZN (2024). Information About Ambulance Care. *Ambulancezorg Nederland*. https://www.ambulancezorg.nl/themas/kwaliteit-van-zorg/urgentie-indeling

[3] AZN (2024). Ambulance Care Sector Overview. *Ambulancezorg Nederland*. https://www.ambulancezorg.nl/en/ambulance-care-sector-overview/facts-figures-2020#:~:text=In%20the%20Netherlands%2C%20ambulances%20and,calls%20at%20the%20same%20time

[4] Heerkens, H., & Van Winden, A. (2017). Solving managerial problems systematically (1st ed.). *Noordhoff Uitgevers*. [PDF].

[5] LMS (2022). LMS Beleids- en Bestedingsplan 2023-2027 *[LMS Policy and Spending Plan 2023-2027]* (In Dutch). [PDF].

[6] Ingolfsson, A. (2013). EMS Planning and Management. *Operations Research and Health Care Policy*. International Series in Operations Research & Management Science, vol 190. Springer. https://doi.org/10.1007/978-1-4614-6507-2_6

[7] Volksgezondheid en Zorg (2024). Acutezorg | Regionaal | Ambulancezorg [Acute Care | Regional | Ambulance Care] (In Dutch). https://www.vzinfo.nl/acute-zorg/regionaal/ambulancezorg

[8] AZN (2022). Sectorkompass Ambulancezorg: tabellenboek 2021 [Sector Compass Ambulance Care] (In Dutch). *Ambulancezorg Nederland*. [PDF].

[9] AZN (2024). Regional Ambulance Service. *Ambulancezorg Nederland*. https://www.ambulancezorg.nl/en/themes/ambulance-care-in-the-netherlands/regional-ambulance-service/regional-ambulance-service

[10] Toregas, C., Swain, R., ReVelle, C., & Bergman, L. (1971). The Location of Emergency Service Facilities. *Operations Research, 19(6)*. https://dx.doi.org/10.1287/opre.19.6.1363

[11] Restrepo, M., Henderson, S.G., & Topaloglu, H. (2009). Erlang Loss Models for the Static Deployment of Ambulances. *Health Care Manag Sci*. https://doi.org/10.1007/s10729-008-9077-4

[12] Becker J., Kurland L., Höglund E., & Hugelius, K. (2023). Dynamic ambulance relocation: a scoping review. *BMJ Open*. https://doi.org/10.1136/ bmjopen-2023-073394

[13] Gendreau, M., Laporte, G., & Semet, F. (2001). A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*. https://doi.org/10.1016/S0167-8191(01)00103-X

[14] Restrepo, M. (2008). Computational methods for static allocation and real-time redeployment of ambulances [PhD Thesis]. *Cornell University*.

[15] Brotcorne, L., Laporte, G., & Semet, F. (2003). Ambulance location and relocation models. *European Journal of Operational Research*. https://doi.org/10.1016/S0377-2217(02)00364-8

[16] Belanger, V., Ruiz, A., & Soriano, P. (2019). Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. *European Journal or Operational Research*. https://doi.org/10.1016/j.ejor.2018.02.055

[17] Daskin, M. (1983). A maximum expected covering location model: formulation, properties and heuristic solution. *Transportation Science*. http://dx.doi.org/10.1287/trsc.17.1.48

[18] Larson, R. (1974). A hypercube queueing model for facility location and redistricting in urban emergency services. *Computers and Operations Research*. https://doi.org/10.1016/0305-0548(74)90076-8

[19] Iannoni, A., & Morabito, R. (2023). A review on hypercube queuing model's extensions for practical applications. *Socio-Economic Planning Sciences*. https://doi.org/10.1016/j.seps.2023.101677

[20] Church, R., & ReVelle, C. (1974). The maximal covering location problem. *Papers of the Regional Science Association*. https://doi.org/10.1007/BF01942293

[21] Takeda, R., Widmer, J., & Morabito, R. (2005). Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model. *Computers & Operations Research*. https://doi.org/10.1016/j.cor.2005.03.022

[22] Jagtenberg, C., Bhulai, S., & van der Mei, S. (2015). An efficient heuristic for real-time ambulance redeployment. *Operations Research for Healthcare*. http://dx.doi.org/10.1016/j.orhc.2015.01.001

[23] Marianov, V., & ReVelle, C. (1996). The queueing maximal availability location model: a model for the sitting of emergency vehicles. *European Journal for Operational Research*. https://doi.org/10.1016/0377-2217(95)00182-4

[24] Rastpour, A., Ingolfsson, A., Kolfal, B. (2020). Modeling red and yellow alert durations for ambulance systems. *Production and Operations Management*. https://doi.org/10.1111/poms.13190

[25] de Bruin, A.M., Bekker, R., van Zanten, L., & Koole, J.M. (2009). Dimensioning hospital wards using the Erlang Loss model. *Annals of Operations Research*. https://doi.org/10.1007/s10479-009-0647-8

[26] Erkut, E., Ingolfsson, A., & Erdogan G. (2007). Ambulance location for maximal survival. *Naval Research Logistics*. https://doi.org/10.1002/nav.20267

[27] Ingolfsson, A., Budge, S., & Erkut E. (2008). Optimal ambulance location with random delays and travel times. *Health Care Manage Sci*. https://doi.org/ 10.1007/s10729-007-9048-1

[28] Karsten, F., Slikker, M., & Houtum, G.J. van (2014). Doman extensions of the Erlang Loss function: their scalability and its applications to cooperative games. *Probability in the Engineering and Informational Sciences*. https://doi.org/10.1017/S0269964814000102

[29] Medhi, J. (2006). The evergreen erlang loss function. *OPSEARCH*. https://doi.org/10.1007/BF03398780

[30] AZG (2024). Onze ambulanceposten [Our ambulance dispatch centers] (In Dutch). *Ambulancezorg Groningen*. https://www.ambulancezorggroningen.nl/over-ons/onze-ambulanceposten

[31] HV Database (2024). https://hv-database.weebly.com/huidige-voertuigen.html

[32] Referentiekader (2023). Referentiekader spreiding en beschikbaarheid ambulancezorg 2023 [Reference framework for distribution and availability ambulance care 2023] (In Dutch). *RIVM*.

[33] AllChartsInfo (2024). Regional Statistics. https://allcharts.info/the-netherlands/

Appendix

Appendix A: Problem Cluster of Ambulance Allocations to A0 Calls