VU University Amsterdam
Faculty of Sciences
Stochastics and Financial Mathematics

MASTER THESIS

Optimal emergency medical service system design

Author: Bc. Barbora Zuzáková

Supervisors: Dr. Sandjai Bhulai
Dr. Geert Jan Kommer

2012

Abstract

In this thesis we describe an ambulance location optimization model that minimizes the number of vehicles needed to ensure a specific service level. In addition to this, we do not consider any predetermined set of base stations and we optimize over all possible placements. The model measures service level as the fraction of calls processed within a given critical time and considers response time to be composed of a random pre-trip delay and random travel time. Moreover we incorporate a probabilistic approach to the ambulance availability when requested to serve demands. The stochastic response time and unavailability of vehicles are critical issues to designing valid models corresponding with real life emergency medical service system. Models that do not account for the uncertainty in all of these components may overestimate the possible coverage for a given number of ambulances or underestimate the number of ambulances needed to provide a specified coverage. We illustrate the application of the model using recent data from the Agglomeration of Amsterdam. Our main target is to design an optimal net of base stations and propose a reallocation of currently used vehicles to provide higher coverage.

Keywords

Emergency medical service, facility location, ambulance allocation, stochastic programming.

Acknowledgment

I would like to express my gratitude to all those who supported me to complete this thesis. Mainly, I want to thank my supervisor Dr. Sandjai Bhulai from VU University Amsterdam who was abundantly helpful and offered invaluable assistance, support and guidance. I am grateful to Dr. Geert Jan Kommer from RIVM who mediated the internship stay which provided me with an important source of data. Finally, I would like to thank my friend Bc. Daniel Krýcha for his comments that substantially improved this thesis.

Amsterdam, May 21, 2012 Barbora Zuzáková

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years there has been an important evolution in the development of ambulance location and relocation models. Since time is fateful in emergency situations, it is critical that vehicles are located so as to ensure an adequate coverage and a quick response time. There are also high costs involved in obtaining and maintaining emergency medical service (EMS) equipment and employing highly qualified staff, therefore it is important to make sure that the resources are used efficiently.

The design of EMS systems involves several interconnected strategic decisions, such as the number and locations of ambulance stations, the number and locations of the vehicles, and the dispatch system used. In this paper we focus on the optimal allocation of ambulance stations along with determining the number of vehicles settled in each one. The service area of an EMS system is often modeled by defining a network consisting of a set of geographical nodes (for instance postal codes), where each node represents a source of requests for response emergency vehicles.

The first models (deterministic models) were very basic and did not take into account the fact that some coverage is lost when an ambulance is dispatched to a call. Nevertheless, these early models served as a sound basis for the development of all subsequent models. Probabilistic models work with the busy fraction of vehicles or random demand. The latest models are dynamic. They can be used to periodically update ambulance positions throughout the day.

In this thesis we explore probabilistic models. The stochastic nature of emergency service requests and the unavailability of emergency vehicles when intervening are critical issues in constructing efficient models. In real life situations, the future emergency service requests from demand sites are not known with certainty. Furthermore, due to the various constraints, such as budget or capacity, pre-allocated emergency vehicles may not be sufficient to cover all demand in the service area within an acceptable time. Since unmet demands in emergency situations may result in loss of life, it is critical to design systems that guarantee target levels of coverage, which are quantified using risk measures on random unmet demand. These target levels are specified in advance based on the previous experience. We consider an EMS system design problem of locating the response facilities (ambulance stations) and determining the number of vehicles to allocate to each facility.

Coverage is determined by the response time to calls, which is generally defined as the time from when a call for ambulance service arrives until paramedics reach the patient. The most obvious and significant component of the response time is the travel time between the ambulance station and the demand location. The majority of the existing operations research literature on ambulance location focuses on travel times, but this is not the only component of the response time. It also includes any delays prior to the trip. Such delays can consist of time spent on the phone obtaining the address and establishing the seriousness of the call, time spent deciding which ambulance to dispatch, time to contact the paramedic crew, time needed for reaching and preparing the vehicle and start the drive. These delays are significant. Furthermore, the response time for a given emergency call can be affected by the availability of the ambulances because sometimes the closest ambulance is busy and another ambulance must respond. Queueing delays (when no ambulances are available) can also occur, but in practice with real systems, they occur infrequently. In the rare situations when all ambulances are busy, incoming calls are typically responded using some type of backup system, such as supervisor vehicles or fire engines.

The model we develop in this thesis is supposed to be applied on ambulance location in the urban area. In rural areas, ambulance availability will be less of a concern and geographical coverage will be of greater concern. In our research we provide the extension of the main model to overcome this difficulty. Our work is motivated by real-world ambulance location projects for the Agglomeration of Amsterdam.

The thesis is organized as follows. In Section 2 we provide an overview of literature relevant to our work, mainly we focus on the papers dealing with coverage models. In Section 3 we analyze the problem of designing an EMS system and we propose a stochastic formulation which includes probabilistic constraints. We present two stochastic models together with their exact solution method. Section 4 is devoted to analysis of available data and verification of model assumptions, in particular we prove that all statements about stochastic distributions are reasonable. In Section 5 we present the computational results and discuss extensive numerical results collected on a large set of test problems. Conclusions are stated in Section 6. In Appendix A we recall important definitions from probability theory that are relevant for our work. Appendix B contains tables and figures that either relate to geographical issues of our formulations or expand the part of computational results. The list of references can be found at the end of this thesis.

Chapter 2

Literature review

Historical overview

The first models were unsophisticated integer linear programming formulations, which dealt with the static and deterministic location problem and they ignored stochastic considerations. The weakness of these formulations has been recognized in the early 1980s, when the issue of server congestion began to be handled in a deterministic setting through backup coverage. Probabilistic models were later introduced in order to cope with server congestion in a stochastic setting. Most of the probabilistic EMS literature extends classical deterministic location problems by additionally focusing on server congestion. The aim is to guarantee an adequate service level with respect to server availability when demands arise. A variety of models has been introduced so far. They differ in the area of focus such as vehicle availability maximization, coverage maximization, cost minimization, temporal coverage, disaster preparedness. Typical recommendations provided by the different models are the numbers and types of vehicles, the locations of vehicle base stations, types and capacities of vehicle base stations.

There exists a rich literature on emergency vehicles setting models. The survey by Marianov and ReVelle (1995) provides an overview of the most important models published until that date. The Brotcorne, Laporte and Semet (2003) review is less general since it focuses on ambulance services, but it inevitably covers some of the same material, albeit with a different emphasis. This article also provides a brief introduction to dynamic relocation models. For other extensive reviews we refer to Goldberg (2004). Stochastic programming formulations using probabilistic constraints, also called chance constraints, are widely applied in stochastic EMS design models.

Comprehensive reviews on facility location under uncertainty can be found in Owen and Daskin (1998), and Snyder (2006).

Maximum expected covering models

The maximum expected covering location model (MEXCLM) suggested by Daskin (1983) can be considered as the basis of subsequent research in the area of coverage models, since the majority of the others are extensions. The MEXCLM model stems from the seminal work of Chapman and White (1974) and Adel and White (1978). They embedded probabilistic constraints in a classical set covering location model (SCLM), first formulated by Toregas, Swain, ReVelle and Bergman (1971), in order to take into account the randomness of service availability. Nevertheless, their model was never implemented due to mathematical difficulties. MEXCLM developed by Daskin (1983) simplifies the model of Chapman and White (1974) assuming independence and common busy probabilities among servers. The model maximizes the expected value of population coverage given that a fixed number of facilities have to be placed in a network. Daskin (1983) developed an exchange-based heuristic that approximates the solution of the problem for all values of the probability of a vehicle being busy. As a generalization of the maximum covering model, ReVelle and Hogan (1989) propose chance constrained stochastic models which maximize the demand covered with a given probability value. Earlier studies performed by Daskin (1983) and Revelle and Hogan (1989) assumed server independence and a system-wide server business probability. Although calls for service may arrive independently, the assumption of independence among service providers may not be justified. Batta, Dolan and Krishnamurthy (1989) proposed some corrective terms to handle unrealistic situations of independence between servers.

An important model that provides more realism (i.e., uses fewer simplifying assumptions) is a hypercube model suggested by Larson (1974). This model allows busy fractions to vary between ambulances and can accommodate ambulances responding to calls outside their assigned districts. The most relevant source to our research is paper by Ingolfsson, Budge and Erkut (2008) where they used an extension of the approximate hypercube model that allows multiple servers at a station. Although this article forms the basis of our research, there are some significant differences between the models. Mainly Ingolfsson, Budge and Erkut (2008) consider response times to be composed of a random delay and deterministic travel times in their computational experiments, our model considers both components of response time to be random. For more information about modeling travel time variability we refer to Ingolfsson, Budge and Erkut (2008), Marianov and ReVelle (1996), Daskin (1987) or Goldberg and Paz (1991).

Non-covering models

Although our model is from the family of coverage models, there are several non-covering models that influenced our work. We point out the paper Noyan (2010) which inspired us to use a scenario approach in the implementation. Noyan described two types of stochastic programming formulations that determine the optimal location and allocation decisions minimizing the total cost while meeting the target service level. The level of service is measured by keeping the unmet demand values below some prescribed target values. The model is based on stochastic versions of the classical capacitated fixed charge facility location problem (CFLP) introduced by Beraldi, Bruni and Conforti (2004) and Beraldi and Bruni (2009). They assume that the main uncertainty is due to the stochastic call arrival process, and they propose stochastic programming formulations under probabilistic constraints to ensure that all requests are served with a prescribed high probability. Uncertainty has been computationally incorporated in different ways. In Noyan (2010) and Beraldi and Bruni (2009) as well as in our work the scenario planning techniques are applied. Handling a larger set of scenarios is significant in modeling uncertainties of real life. Beraldi and Bruni (2009) also introduce a two-stage stochastic programming problem, where the second stage decision variables are associated with scenarios to represent the assignment of vehicles to demand nodes under each scenario.

Chapter 3

Problem formulation, properties and exact solution approach

3.1 General dispatch procedure

The chain of events leading to the intervention of an ambulance vehicle to the scene of an accident includes the following four steps: accident detection and reporting, call screening, vehicle dispatching, actual intervention by paramedics. The entire time to tackle the accident is often called operating time.

When a call arrives to the call center, the operator (dispatcher) has to enter important information, estimate the sincerity of the emergency call, and labels it with a priority. The call center operator also gives instructions to tide over until an EMS vehicle arrives to take over the first aid. There are also cases when the situation does not require EMS care. Making a decision on the type and number of ambulances to dispatch is of highest importance. If EMS intervention is necessary, the dispatcher contacts an available EMS vehicle according to certain guidelines, and sends it to the location where the call originates from. The time a call is handled together with the time needed for mobilization of an ambulance crew is called the pre-trip delay. Obviously this time should be modeled as a random variable since we cannot know with certainty its exact value.

An other important part of the operating time is the travel time to a patient. In all EMS models it is important to know the travel time between each two locations. It would be impossible to make good operational decisions if the travel times for EMS vehicles were unknown. Many previous researches showed that treating driving times as deterministic values can underestimate expected coverage (for instance see Ingolfsson, Budge and Erkut (2008)). Therefore in our research we consider the travel time as a random variable. Intuitively, the distribution of the travel time also depends on the period of the day (at least the parameters of the distribution differ). One possible approach is to provide required coverage for the busiest period (when the average travel time prolongs and more vehicles are needed) and for the rest of the periods just reduce the number of operating vehicles.

The pre-trip delay together with the travel time to the patient is called the response time. This time is always random, nevertheless national governments set specific standards. For example in the US in urban areas, 95% of requests should be served within 10 minutes; in rural areas, they should be served within 30 minutes.

Service time includes also time spent by the paramedics crew treating the patient at his location. The crew also makes a decision whether to transport the patient to a hospital or not. The travel time to a hospital matches the same travel time model as the travel time to a patient. The time necessary for conveying the patient to medics in the hospital cannot be known in advance and is therefore treated as a random variable. As soon as the paramedics crew is not needed in the hospital, it returns to a base station. The whole operating time is illustrated in Figure 3.1.

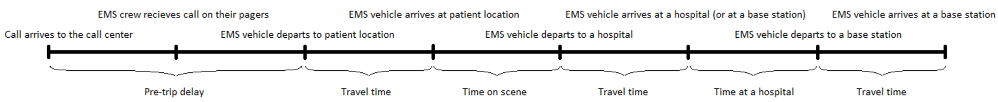

Figure 3.1: Operating time of an EMS system.

3.2 Problem formulation

Our aim is to design a system that minimizes the number of ambulances needed to provide a specified service level. We extend the ambulance location modeling paradigm by incorporating uncertainty in ambulance availability and in response times. The majority of covering models that we listed in the literature review use deterministic (average) travel times. In practice, the driving time varies depending on the time of the day or the day of the week. In this study we have several driving time models available for different time periods of the day (rush hours, off-peak hours and night hours). Our models can be first used to solve the problem of determining the facility locations and the number of vehicles allocated to each facility according to a selected driving time model (we will choose the busiest period of the day). Then, given the optimal number of facilities and the allocation of the emergency stations, modified versions of the first model, we can optimize the number of operating vehicles at every time period. The optimal allocation decisions for each time period would provide us with the number of vehicles required at each facility for each time period according to the specified risk preferences.

The assumption made by early covering models is that if an ambulance is available within a specified maximum distance of a demand point, then the demand point is covered. EMS systems typically measure performance based on the fraction of calls responded to within a specified time standard. However, for a given ambulance location and a demand point, it is not possible to know with certainty whether the call will be responded within the time standard. It depends on the pre-trip delay and the travel time as well as the availability of the ambulance, which cannot be predicted with certainty. Moreover, delays are usually not explicitly treated in papers dealing with coverage models, although it is easy to incorporate a constant (average) delay into all coverage models by simply subtracting the delay from the specified maximum response time. In our research we will estimate a distribution function of the delay and incorporate it as a random variable. Our model does not rely solely on average response times, and hence it is not limited by the resulting strict classification of demand points as covered or not covered. It allows incorporation of randomness in pre-trip delays, and computes an expected coverage for each demand point, given the ambulance locations and estimates of the ambulance availabilities. Traditional models assign 1 or 0 to each demand point, i.e., if the point can be reached within the specified time standard then 1 is assigned or else 0 is assigned. As suggested in Ingolfsson, Budge and Erkut (2008) we increase the model realism by replacing the 0-1 sequences by real numbers, which are better estimates of the fraction of calls related to different demand points that can be reached within the specified time standard.

We also incorporate some extensions suggested in Ingolfsson, Budge and Erkut (2008). We add limits for the number of ambulances settled at each station and introduce new variables to decide which stations to open. Exclusion of these variables causes that the optimal solution is found with respect to predetermined station location.

3.3 Model formulation

In this subsection we present our assumptions and notation, and describe the model properties. First we start with a simpler model where the number and the exact locations of ambulance stations are given. The objective of our interest is to set the optimal number of vehicles settled in those stations as to ensure an adequate coverage. Afterwards we expand the model by including the second type of decision variables to involve the decision on the optimal number and locations of ambulance stations. Since both of the models are from the family of coverage models, it is crucial to understand well what this term means, therefore we first give the exact definition. We say that a demand node is covered by a facility location if the time needed to reach the node is less than or equal to an acceptable value, which is known as a response time standard.

In the following part we state the model's inputs and decision variables.

Inputs

- M: finite set of m candidate facility locations indexed by i, i.e., places where stations can be located.
- N: finite set of n demand nodes indexed by j.
- b_i: the maximum number of ambulances at station i.
- d_j: total demand generated at node j.
- α, δ: parameters which specify the coverage objective; the calls should be responded to in at most δ time units with probability of at least α.
- w_{ij}: the probability that the response time is less than or equal to δ time units, provided that the call arrives from node j and the i-th station responds, i.e., $w_{ij} = \mathbb{P}(R_{ij} \leq \delta)$, where R_{ij} is a random response time composed of random pre-trip delay and random travel time.
- ρ_i: the busy fraction for an ambulance at station i, i.e., the probability that an ambulance at station i is not available to respond to calls.

Decision variables

- x_i: number of vehicles located at station i.

We stress that we do not monitor how available vehicles are assigned to particular emergency calls, i.e., we do not consider any dispatching procedure. The probability w_{ij} can be interpreted as the fraction of vehicles located at station i that responds to the demand generated at node j. However, these vehicles are not determined to respond to calls at node j in real life dispatching problems. As soon as the number of vehicles located at each facility is determined, different models may be utilized to find the solution for the dispatching and reallocating procedure.

We consider the whole problem formulation related to a certain amount of time. The length of this time period is taken as a reasonable time relevant to the service time, i.e., the time required for an ambulance crew to intervene at the patient location and to return to the original location before responding to another service request. Similar to other papers, as Beraldi, Bruni (2009) or Noyan (2010), we consider the time unit to be equal to one hour.

Let us denote x as an m-dimensional vector whose elements express the number of vehicles located at each station. Then we arrive the following optimization problem:

$$
\begin{aligned}
(P1) \qquad & \text{maximize} \qquad && c(x) = \sum_{j \in N} d_j c_j(x) && (3.1)\\
& \text{subject to} && \sum_{j \in N} d_j c_j(x) \geq \alpha \sum_{j \in N} d_j, && (3.2)\\
& && x_i \leq b_i, \qquad \forall i \in M, && (3.3)\\
& && z(x) \equiv \sum_{i \in M} x_i = b, && (3.4)\\
& && x_i \in \mathbb{Z}_+, \qquad \forall i \in M, && (3.5)
\end{aligned}
$$

where

$$c_j(x) = \sum_{i \in M} (1 - \rho_i) w_{ij} x_i, \qquad \forall j \in N. \tag{3.6}$$

This model maximizes the expected coverage $c(x)$, subject to a constraint on the total number of ambulances $z(x)$ given by Equation (3.4). For the time being, we assume b to be a given fixed number, but in the next subsection we will introduce an algorithm for how to overcome this simplification and change b into a decision variable. The objective function maximizes the system wide coverage $c(x)$ which is computed as the sum of a coverage that can be provided to the individual node by the whole system, $c_j(x)$, multiplied by the total demand generated at this node, d_j. Constraint (3.2) ensures that the specific fraction of demand is covered. Constraint (3.3) is a capacity constraint, it guarantees that the number of vehicles located at each station does not overflow its capacity. Constraint (3.5) is the integrality and nonnegativity constraint.

Now assume that the number of vehicles at the i-th station is equal to x_i and define a random variable X_i as the number of ambulance vehicles at the i-th station that are available to respond to a call. Then the probability that X_i equals k, $k = 0, \ldots, x_i$ can be expressed as:

$$\mathbb{P}(X_i = k) = \binom{x_i}{k}(1-\rho_i)^k \rho_i^{x_i - k}.$$

Thus X_i has Binomial distribution $Bi(x_i, (1-\rho_i))$. Taking the expected value of X_i, i.e., $\mathbb{E}X_i = (1-\rho_i)x_i$, we arrive to Formula (3.6).

The total demand generated at each demand node j, d_j, is not known in advance, therefore the model formulated as (P1) cannot be used straight away. In the spirit of Noyan (2010), we introduce the scenario approach. We assume that we are given a discrete set of scenarios, a set of realizations of service requests generated in the whole system. Generating of scenarios is described in Section 4.3. Let S be the finite set of scenarios and denote as π^s the probability that the exact scenario $s \in S$ is generated. Then d_j^s becomes the realization of demand from node j under scenario s. Thus we can reformulate the previous model as follows:

$$
\begin{aligned}
(P2) \qquad & \text{maximize} \qquad && c(x) = \sum_{s \in S} \pi^s \sum_{j \in N} d_j^s c_j^s(x) && (3.7)\\
& \text{subject to} && \sum_{s \in S} \pi^s \sum_{j \in N} d_j^s c_j^s(x) \geq \alpha \sum_{s \in S} \pi^s \sum_{j \in N} d_j^s, && (3.8)\\
& && x_i \leq b_i, \qquad \forall i \in M, &&\\
& && z(x) \equiv \sum_{i \in M} x_i = b, &&\\
& && x_i \in \mathbb{Z}_+, \qquad \forall i \in M, &&
\end{aligned}
$$

where

$$c_j^s(x) = \sum_{i \in M}(1-\rho_i^s) w_{ij} x_i, \qquad \forall j \in N. \qquad (3.9)$$

The reason why the busy fractions are dependent on scenarios will be clear as soon as we introduce the initial estimation of busy fraction in Subsection 4.2.6.

Now let us introduce the extended version of the previous model by taking into account the decision on facility locations. We will use the following inputs and decision variables:

Inputs

- M: finite set of m candidate facility locations indexed by i.
- N: finite set of n demand nodes indexed by j.
- b_i: the maximum number of ambulances at station i.
- β: the maximum number of ambulance stations in the whole system.
- d_j: total demand generated at node j.
- α, δ: parameters which specify the coverage requirements.
- w_{ij}: the probability that the response time to a call generated at node j and responded from the station at node i is less than or equal to δ time units.
- ρ_i: the busy fraction for an ambulance at station i.

Decision variables

- x_i: number of vehicles located at station i.
- y_i: binary indicator, i.e., y_i equals 1 if a facility is located at node i, equals 0 otherwise.

Then the optimization problem is:

$$
\begin{aligned}
(P3) \qquad & \text{maximize} \qquad && c(x) = \sum_{j \in N} d_j c_j(x) \\
& \text{subject to} && \sum_{j \in N} d_j c_j(x) \geq \alpha \sum_{j \in N} d_j, \\
& && \sum_{i \in M} y_i \leq \beta && (3.10) \\
& && x_i \leq b_i y_i, \qquad \forall i \in M, && (3.11) \\
& && z(x) \equiv \sum_{i \in M} x_i = b,
\end{aligned}
$$

$$x_i \in \mathbb{Z}_+, \qquad \forall i \in M,$$
$$y_i \in \{0,1\}, \qquad \forall i \in M,$$

where

$$c_j(x) = \sum_{i \in M} (1 - \rho_i) w_{ij} x_i, \qquad \forall j \in N.$$

Let us argue in detail the significance of the labeled constraints (3.10). Assume the problem of designing the optimal net of station locations in the whole system. We allow set M to be equal to set N, i.e., any node is considered as a possible location for a station. Then we can get the optimal solution as a densely settled net of stations, each being operated by a low number of vehicles. Intuitively, this solution is not a suitable one, since establishing and maintaining new stations involves high costs. Therefore constraint (3.10) replaces the budget limitation that is not included in our model and ensures adequate expenses for creating a new system. Constraint (3.11) is the version of a capacity constraint introduced in the first model. It guarantees that the number of vehicles located at the particular station cannot exceed its capacity in case that the facility location is opened.

We hereby state the scenario approach of the extended model:

$$
\begin{aligned}
(P4) \qquad & \text{maximize} \qquad & & c(x) = \sum_{s \in S} \pi^s \sum_{j \in N} d_j^s c_j^s(x) \\
& \text{subject to} \qquad & & \sum_{s \in S} \pi^s \sum_{j \in N} d_j^s c_j^s(x) \geq \alpha \sum_{s \in S} \pi^s \sum_{j \in N} d_j^s, \\
& & & \sum_{i \in M} y_i \leq \beta \\
& & & x_i \leq b_i y_i, \qquad \forall i \in M, \\
& & & z(x) \equiv \sum_{i \in M} x_i = b, \\
& & & x_i \in \mathbb{Z}_+, \qquad \forall i \in M, \\
& & & y_i \in \{0,1\}, \qquad \forall i \in M,
\end{aligned}
$$

where

$$c_j^s(x) = \sum_{i \in M} (1 - \rho_i^s) w_{ij} x_i, \qquad \forall j \in N, s \in S.$$

In the next section we focus on the exact solution method that can be applied to solve the above stated models.

3.4 Exact solution approach

The assumption that the busy fractions ρ_i are exogenous parameters is not realistic, since they depend on the number of ambulance vehicles in the stations. As the objective of the optimization, they cannot be known in advance. To overcome this limitation, we propose to use an algorithm that iterates between solving the location optimization problem and estimating the busy fractions. The exact value of the busy fractions at each step is influenced by the actual number of vehicles in the system which follows the bisection algorithm. Assume that the set of scenarios is given, then the algorithm of finding the optimal solution for the first model proceeds as follows:

Algorithm 1

STEP 1. Set n equal to 1. Choose an initial value for the total number of ambulances, b^1. We propose to set $b^1 = \sum_{i \in M} b_i$, i.e., the maximum acceptable number of ambulances in the system. Set the initial values of the lower and upper bounds of the number of ambulances in the system as $b^1_{\text{lower}} = 0$, $b^1_{\text{upper}} = \sum_{i \in M} b_i$ and the initial solution of the optimization problem with the number of ambulances equal to b^1_{upper} as $x^1_{\text{upper}} = (b_1, \ldots, b_m)$.

STEP 2. Attempt to maximize coverage with b^n ambulances as follows:

STEP 2a. Set the busy fraction ρ_i^s to an initial estimate of the busy fraction (using Formula (4.2)) corresponding to b^n ambulances in the system.

STEP 2b. Using the busy fractions ρ_i^s, find the solution x^n of the modified version of optimization problem (P2):

$$
\begin{aligned}
(P5) \qquad & \text{maximize} \qquad && c(x) = \sum_{s \in S} \pi^s \sum_{j \in N} d_j^s c_j^s(x) \\
& \text{subject to} && x_i \le b_i, \qquad \forall i \in M, \\
& && z(x) \equiv \sum_{i \in M} x_i = b^n, \\
& && x_i \in \mathbb{Z}_+, \qquad \forall i \in M,
\end{aligned}
$$

where $c_j^s(x)$ is defined as

$$c_j^s(x) = \sum_{i \in M} (1 - \rho_i^s) w_{ij} x_i, \qquad \forall j \in N, s \in S.$$

STEP 3. If the condition

$$c(x) \geq \alpha \sum_{s \in S} \pi^s \sum_{j \in N} d_j^s$$

is satisfied then set the new values of parameters b_{upper}, b_{lower} and x_{upper} as follows:

$$\begin{aligned} b_{\text{upper}}^{n+1} &\longleftarrow b^n, \\ b_{\text{lower}}^{n+1} &\longleftarrow b_{\text{lower}}^n, \\ x_{\text{upper}}^{n+1} &\longleftarrow x^n, \end{aligned}$$

otherwise

$$\begin{aligned} b_{\text{upper}}^{n+1} &\longleftarrow b_{\text{upper}}^n, \\ b_{\text{lower}}^{n+1} &\longleftarrow b^n, \\ x_{\text{upper}}^{n+1} &\longleftarrow x_{\text{upper}}^n. \end{aligned}$$

Set the new value of b equal to

$$b^{n+1} = \left\lfloor \frac{1}{2} \left(b_{\text{upper}}^{n+1} + b_{\text{lower}}^{n+1} \right) \right\rfloor.$$

If $b^{n+1} = b_{\text{lower}}^{n+1}$ and $n = 1$ then the problem has no solution. If $b^{n+1} = b_{\text{lower}}^{n+1}$ and $n > 1$ then stop the algorithm. The optimal solution is given by the vector x_{upper}^{n+1} and the corresponding optimal number of ambulances in the system is equal to b_{upper}^{n+1}. If the condition $b^{n+1} = b_{\text{lower}}^{n+1}$ is not satisfied then increase the value of n by one and go to STEP 2.

Remark. The algorithm of finding the optimal solution of the extended model, i.e., the one which includes decision variables y_i, $i \in M$, proceeds analogously except for replacing the problem (P5) by the following one:

$$
\begin{aligned}
(P5) \qquad & \text{maximize} \qquad & c(x) = \sum_{s \in S} \pi^s \sum_{j \in N} d_j^s c_j^s(x) \\
& \text{subject to} \qquad & \sum_{i \in M} y_i \leq \beta, \\
& & x_i \leq b_i y_i, \qquad \forall i \in M, \\
& & z(x) \equiv \sum_{i \in M} x_i = b^n, \\
& & x_i \in \mathbb{Z}_+, \qquad \forall i \in M, \\
& & y_i \in \{0, 1\} \qquad \forall i \in M,
\end{aligned}
$$

In this case the initial values for the total number of ambulances b^1 is derived as follows: assume the permutation χ of set M such that

$$b_{\chi_1} \geq b_{\chi_2} \geq \cdots \geq b_{\chi_m}.$$

That is, the stations are arranged in descending order of their capacities. Then we define b_1 as:

$$b^1 = \sum_{i=1}^{\beta} b_{\chi_i}.$$

We emphasize that the algorithm does not guarantee to converge to an unique solution. Sometimes two or more similar systems can be designed. In such cases, the planner can compare them in terms of another criteria, for instance, the number of active stations which highly influences the costs involved in maintaining the system.

Chapter 4

Data description and analysis

4.1 Data description

Our data set consists of 49,426 EMS calls and their properties received in 2010 in the Amsterdam region. The first information are enrolled at the dispatching center. The date, day of the week, and exact time when the call was generated are registered immediately. Then the dispatching operator records the type of the call. Not all of the entries are emergency calls, sometimes ambulance vehicles transfer patients between different hospitals, or transport them home from the hospital respectively. These planned transfers make part of our database, but their number is negligible in comparison with emergency calls, the planned transport creates only 5.86% of all records. Since we want to design a system for providing the first aid, all information about planned calls is irrelevant and therefore removed from the database. The operator has to estimate the sincerity of the accident and label it on the given scale. In Amsterdam the calls are divided into two groups, A1 for more urgent accidents requiring fast intervention and A2 for the less serious ones. The rest of the information is registered by the paramedics crew. Every vehicle is equipped with a special mobile device which is determined to log exact events like the time when the crew is informed about a new accident, time when it leaves the station or any base location, time it arrives at the patient location to provide the first aid. After having treated the patient paramedics have to inform the call center if the transport to a hospital is necessary or not. If it is so, they log the exact time when they leave the patient location and time when they arrive at the hospital. The last information collected by the crew is the time when the vehicle is free again. Scarce mistakes that occur during the recording process were removed from the data set when apparent. There are given standards which specify the time limit of every action depending on the urgency level, for the boundaries see Table B.1 in Appendix B. Further we have the information about the postal code where the intervening vehicle was located when announcing the EMS call and also the postal code of the hospital where the patient was eventually transported.

First we considered each of 103 postal codes in the Amsterdam region as a demand node. However the diversity throughout these areas, originating from large differences between interarrival rates of accidents, caused instability in obtained results. Probabilities of particular scenarios were highly influenced by the regions that had extremely low presence of accidents and the optimal solution of base stations locations varied depending on the generated scenarios. Therefore we decided to use the splitting of the Amsterdam region into 19 larger districts and treat each district as a demand node. The list of the districts with graphical illustration is displayed in Table B.2 and Figure B.1 in Appendix B.

Beside the database of EMS calls the travel time model makes the indispensable input for our computations. The travel time model is a square matrix whose elements express the average travel time between every two pairs of postal codes in the Amsterdam region (the calculation of the matrix is out of the scope of this paper). However in the real life situation the travel times between nodes are not deterministic values, therefore we endowed the travel time modeling with a parameter that compensates the deviation between the theoretical and real travel time. Naturally we also had to adjust the matrix with respect to the partition of the region into 19 districts. We were provided with the mapping from the set of the districts to the set of the postal codes, each district was assigned a central postal code. This mapping enabled us to reduce the original travel time matrix in the spirit of the districts taking the time distances between the central postal codes.

According to our computational experience it is necessary to divide the day into several homogeneous periods, in which the number of accidents per hour is approximately similar, and make every analysis for each period separately. Assuming the independent arrival process with the same parameter during all day will lead to disapproving a fitted distribution for the interarrival time by statistical tests, although the fitted distribution abstracting the parameter would be correct. The histogram in Figure 4.1 reflects the difference between the number of accidents during day time. While at 4 a.m. and 5 a.m. there are around 1,200 EMS calls received, in the afternoon at 1 p.m. and 2 p.m. there are nearly 4,000 calls. Therefore we suggest to divide the day into three periods as indicated in Figure 4.1 and Table 4.1. The adequacy of such a division

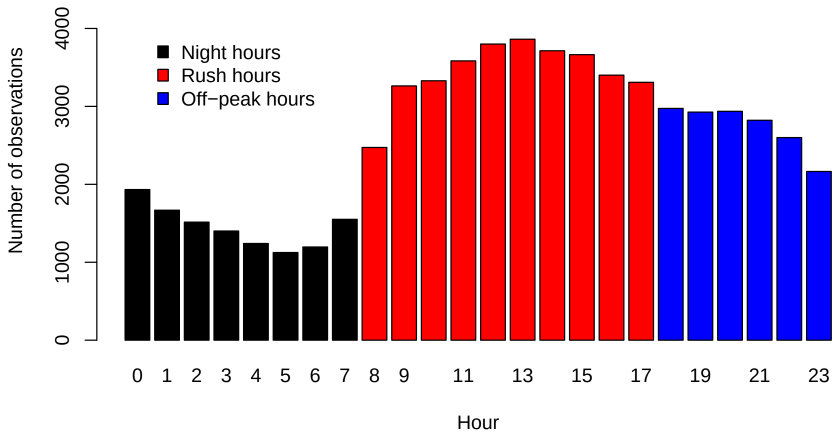

Figure 4.1: Histogram of received EMS calls.

will be proved in the following section.

period	time
night hours	0:00 - 8:00
rush hours	8:00 - 18:00
off-peak hours	18:00 - 0:00

Table 4.1: Day periods.

4.2 Problem data

Before running the optimization solver itself we need to estimate or set the following parameters:

- M set of m station locations (indexed by i), N a set of n demand nodes (indexed by j). We have already indicated the choice of the set of demand nodes and the choice of the set M is discussed in the next chapter where we summarize the obtained results for the concrete model.

- b_i, β, the maximum number of ambulances at station i, the maximum number of stations in the whole system respectively. These values will be specified in Chapter 5.
- Parameters δ and α. In our computational experiment we use the common values introduced in Chapter 5.
- λ_j the positive arrival rate of calls generated at demand node j, whose estimation is discussed in Subsection 4.2.1. Knowledge of interarrival rates is crucial when generating scenarios.
- Travel time distribution between the i-th station and the j-th node, or the j-th station and the assigned hospital respectively, based on the travel time model. Modeling of the travel time distribution is presented in Subsection 4.2.3.
- The probabilities w_{ij}, the probabilities that the response time is less than or equal to δ time units, provided that the call arrives from node j and a vehicle from the i-th station responds. Subsection 4.2.4 is devoted to the calculation of these probabilities.
- An assignment of a particular hospital to each demand node which is discussed in Subsection 4.2.5.
- The initial estimation of busy fraction ρ_i for the ambulances at station i, described in Subsection 4.2.6.

Remark. All calculations presented in the next subsections were performed in Microsoft Office Access (2007), Microsoft Excel (2010) and R (version 2.12.1).

4.2.1 Estimation of the arrival rates

We want to estimate a positive arrival rate λ_j for each demand node j. We assume independent Poisson arrival processes at the nodes (for a definition of the Poisson process see Appendix A), i.e., interarrival times for each demand node are exponentially distributed with parameter λ_j. We denote the system wide arrival rate with λ. Then the following relation has to hold:

$$\lambda = \sum_{j \in N} \lambda_j .$$

To illustrate the accuracy of the assumption on arriving process of accidents we present in detail results obtained by analyzing data of events reported during rush hours. The overview of other results (off-peak and night hours) can be found in Table B.3 in Appendix B. First we will demonstrate that the interarrival time between two accidents happened at different demand nodes has the exponential distribution. Then we will show that the interarrival time between accidents for each demand node is exponentially distributed as well. In this case the average of individual rates should approach the overall rate.

We analyzed data of all rush hours EMS calls received in 2010, which is a set of more than 26,000 observations. Our aim was to find a distribution that would best fit the available data. There are several methods for fitting a distribution, in our case the maximum likelihood was applied. Figure 4.2 displays the empirical distribution of the interarrival time, which is well approximated by the exponential distribution.

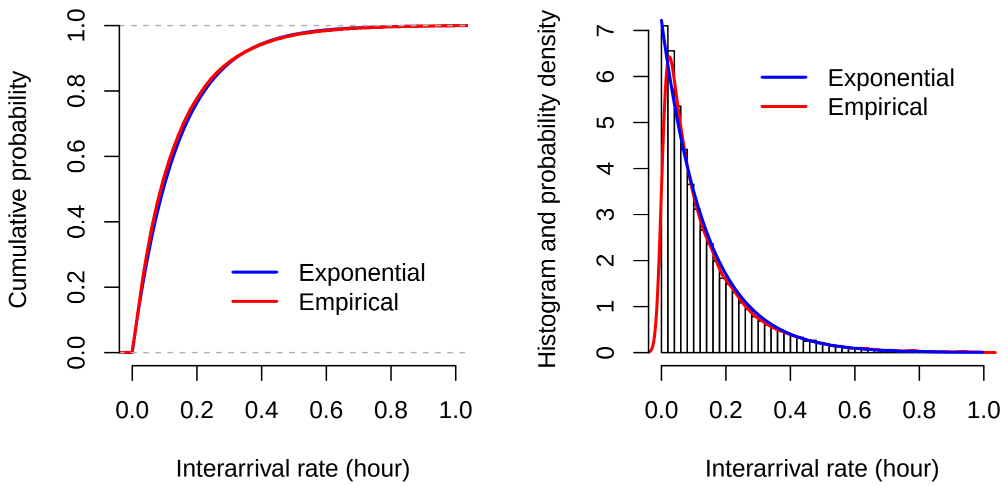

Figure 4.2: Empirical distribution of the interarrival time in rush hours and the fitted exponential distribution.

The accuracy of the exponential distribution was verified by a two-sided Kolmogorov -Smirnov test and Anderson-Darling test. Note that both of them test suitability of concrete distribution, i.e., the rate of exponential distribution has to be specified (we set this value equal to the estimated interarrival rate 7.22). We tested the null hypothesis that interarrival times are exponentially distributed on the confidence level 0.05. Both tests ran ten times on different random samples of size 20. Table 4.2 summarizes the obtained results. Kolmogorov-Smirnov accepted the null hypothesis in all tests and Anderson-Darling in 9 tests out of 10.

No.	p-value	No.	p-value
1	0.5108	6	0.9606
2	0.9784	7	0.4705
3	0.5635	8	0.7762
4	0.7872	9	0.9598
5	0.6599	10	0.5870

No.	p-value	No.	p-value
1	0.0011	6	0.1567
2	0.7494	7	0.6703
3	0.1755	8	0.6661
4	0.7878	9	0.6445
5	0.8178	10	0.6773

Table 4.2: P-values of Kolmogorov-Smirnov tests (left) and Anderson-Darling tests (right) of the exponential distribution of the interarrival time in rush hours.

The interarrival times ranged from 0 seconds to 30.09 hours with a sample mean of 0.1385, i.e., accidents appear every 8 minutes and 30 seconds on average. The rate of the fitted distribution is equal to 7.22, which is exactly the reciprocal value of the sample mean. This number can be interpreted as the average number of accidents recorded every hour. Table B.3 in Appendix B summarizes obtained results when analyzing night hours and off-peak hours. At this point let us stress an important result. The interarrival rate of rush hours is the highest one, i.e., rush hours are the busiest, therefore in further modeling our aim is to cover firstly the demand generated during this period and then reduce the number of ambulances needed to operate in night hours and off-peak hours.

Nevertheless, in our model we assume independent Poisson arrival processes at each node, therefore the analysis of demand nodes is crucial for us. However the previous results can be used as feedback, since the overall rate should equal the sum of the individual rates. Indeed, the sum of the individual rates equals 7.15 which is close to the value of the overall rate (7.22). For each demand node we first estimated the interarrival rate between the succeeding accidents based on fitting an exponential distribution based on maximum likelihood method, then the distribution was verified by the Kolmogorov-Smirnov and the Anderson-Darling test. We state the complete list of interarrival rates for all districts and periods in Table B.4 accompanied by the graphical illustration in Figures B.2, B.3 and B.4 all in Appendix B. Here we want to demonstrate the suitability of the exponential distribution. From the set of 19 demand nodes the Kolmogorov-Smirnov test approved the exponential distribution as a suitable distribution in all 19 cases and the Anderson-Darling test in 17 cases (all testing was carried out on the confidence level 0.95).

To illustrate the procedure of the individual approach we display only results for a randomly chosen node as all of them were treated similarly (the district of Amsterdam Nieuw-West was chosen). The number of observations registered at this node is equal to 3,852. The rate of the fitted exponential distribution equals to 1.0356, which negligibly differs from the reciprocal value of the sample mean (1.0358). Figure 4.3 displays the empirical distribution of the interarrival time, which is approximated by the exponecial distribution.

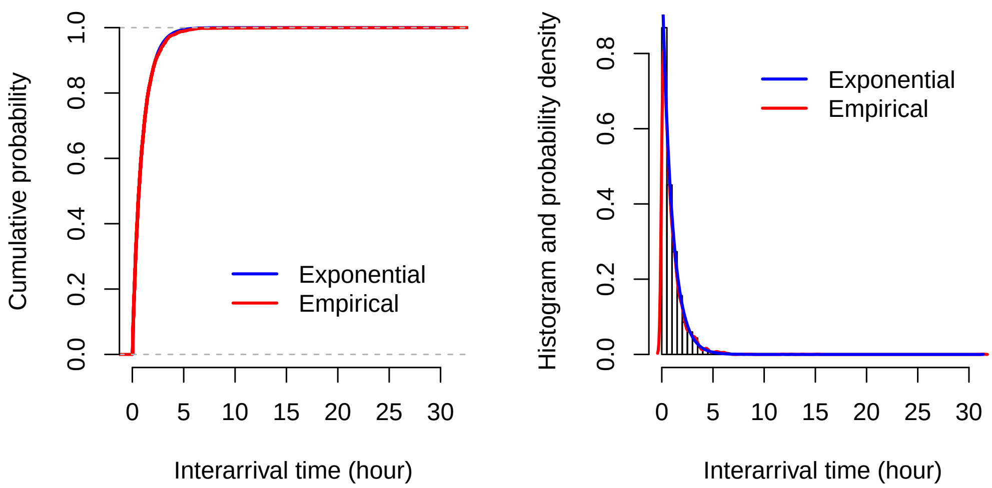

Figure 4.3: Empirical distribution of the interarrival time in rush hours at a demand node and the fitted exponential distribution.

The accuracy of the exponential distribution was verified again by the two-sided Kolmogorov-Smirnov test and Anderson-Darling test. As in the previous case, the tests were repeated ten times on different random samples of size 20. From Table 4.3 we can draw the conclusion that the assumption of independent Poisson arrival processes at the node is satisfied.

No.	p-value	No.	p-value
1	0.2066	6	0.6496
2	0.0110	7	0.1081
3	0.2813	8	0.8419
4	0.2075	9	0.4087
5	0.0824	10	0.9997

No.	p-value	No.	p-value
1	0.1047	6	0.5760
2	0.0102	7	0.1076
3	0.0611	8	0.6974
4	0.1039	9	0.1157
5	0.0708	10	0.9962

Table 4.3: P-values of Kolmogorov-Smirnov tests (left) and Anderson-Darling tests (right) of the exponential distribution of the interarrival time in rush hours at a demand node.

4.2.2 Estimation of the distribution function of the pre-trip delay

In this subsection we discuss estimation of the distribution function for the pre-trip delay. The estimation of the distribution function requires some simplifying assumptions. In particular, we assume that pre-trip delays do not depend on nodes where the calls are generated or on a particular ambulance station. We also exclude dependence on the day time periods. These arguments sound reasonable since the first part of the pre-delay is caused by operators in the call center and the second by a dispatched crew which needs several minutes to be able to set off from the station. Both of them try to reduce the time required to fulfill their tasks so that they can provide first aid as soon as possible.

Computational experience

We analyzed data from the Amsterdam region to illustrate the significance of pre-trip delays. We used the a data set consisting of more than 45,000 (45,668) of EMS calls received in one year regardless of distinguishing between nodes where the calls were generated, times when they were announced and the station which responded.

Figure 4.4 displays the empirical distribution of the pre-trip delays, which is well approximated by the lognormal distribution. We again applied the maximum likelihood method.

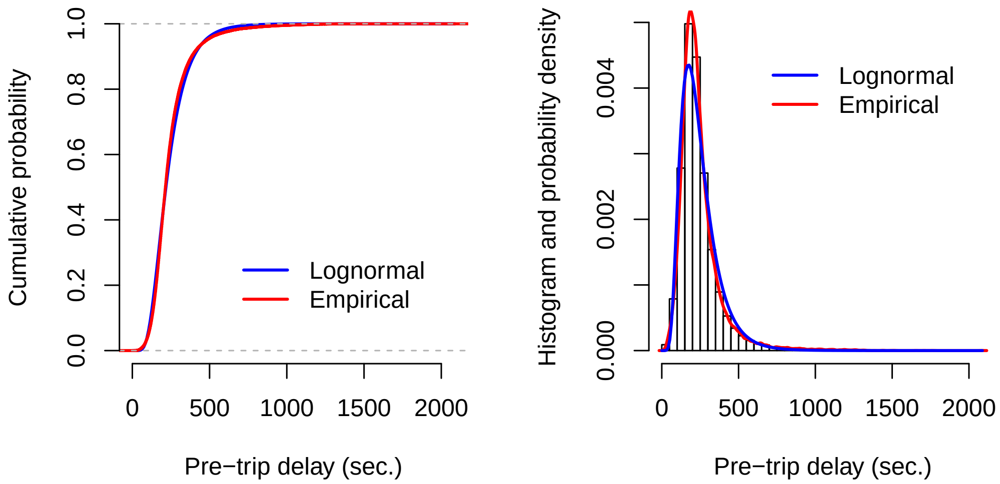

Figure 4.4: Empirical distribution of the pre-trip delay and the fitted lognormal distribution.

The accuracy of the lognormal distribution was verified by the two-sided Kolmogorov -Smirnov test and the Shapiro-Wilk test for normality applied on logarithmic values of the data. We tested the null hypothesis that pre-trip delays are lognormally distributed on the confidence level 0.05. The tests ran ten times on different random samples of size 20 of the original data, or their logarithmic values respectively. Table 4.4 summarizes the obtained results. In the first case, we applied the Kolmogorov-Smirnov test, the lognormal distribution was confirmed as a relevant distribution by each test, in the second case, the Shapiro-Wilk test, the lognormal distribution was confirmed by 8 tests out of 10, therefore we conclude that the pre-trip delays can be modeled as a random variable with lognormal distribution.

No.	p-value	No.	p-value
1	0.7428	6	0.3715
2	0.8407	7	0.6609
3	0.2757	8	0.8983
4	0.7676	9	0.2957
5	0.6121	10	0.3986

No.	p-value	No.	p-value
1	0.0265	6	0.9271
2	0.0216	7	0.1531
3	0.3395	8	0.3858
4	0.1030	9	0.7928
5	0.0674	10	0.1308

Table 4.4: P-values of Kolmogorov-Smirnov tests (left) and Shapiro-Wilk tests (right) of the lognormal distribution of the pre-trip delay.

The delays ranged from 10 second to 2088 seconds (which is approximately 35 minutes), with an average of 243.99 seconds and a standard deviation 135.92 seconds. The variation in the delay is too large (the standard deviation is more than 55% of the sample mean) which endores modeling the pre-trip delays as a random variable instead of considering it as a constant. The fitted distribution has mean equal to 242.81 and a standard deviation of 164.54. Note that these values are very close to the sample mean and the sample standard deviation.

4.2.3 Estimation of the distribution function of the travel time

We will model the travel time similarly as Marionov ReVelle (1996), i.e., they assumed the travel time from station i to node j as normally distributed with known mean and variance. As the mean we set the value t_{ij} given by the travel time model. We will assume that the variance parameter σ^2 does not depend on the distance between the station and the node. One could argue that treating this parameter as a constant independent of the distance is a limiting assumption, since a longer distance or permanent traffic barriers (crossing river, channels) involve a higher delay. However this fact is already incorporated when constructing the travel time model. Therefore we can assume σ^2 to be a constant for each station-node pair, which describes only random fluctuations around the mean. The distribution of the travel time from the i-th station to the j-th node can be written as:

$$T_{ij} = t_{ij} + K,$$

where K is a random variable with normal distribution $N(0, \sigma^2)$.

We will present in detail the results obtained by analyzing rush hours. The results for night hours and off-peak hours are stated in Table B.5 in Appendix B. We have at our disposal travel times realizations between nodes and stations, nodes and hospitals respectively, which include more than 23,000 items. Analyzing the distribution of the travel time is equivalent to analyzing the distribution of fluctuations K. Figure 4.5 displays the empirical distribution and the empirical density of fluctuations in rush hours. The fluctuation is well fitted by a normal distribution with standard deviation $\hat{\sigma}$ which equals to 438.46 (and a mean equal to 0). Note that the sample mean μ is not equal to 0 (237.6), this fact causes that $\hat{\sigma}$ is always bigger than the sample standard deviation σ (368.51).

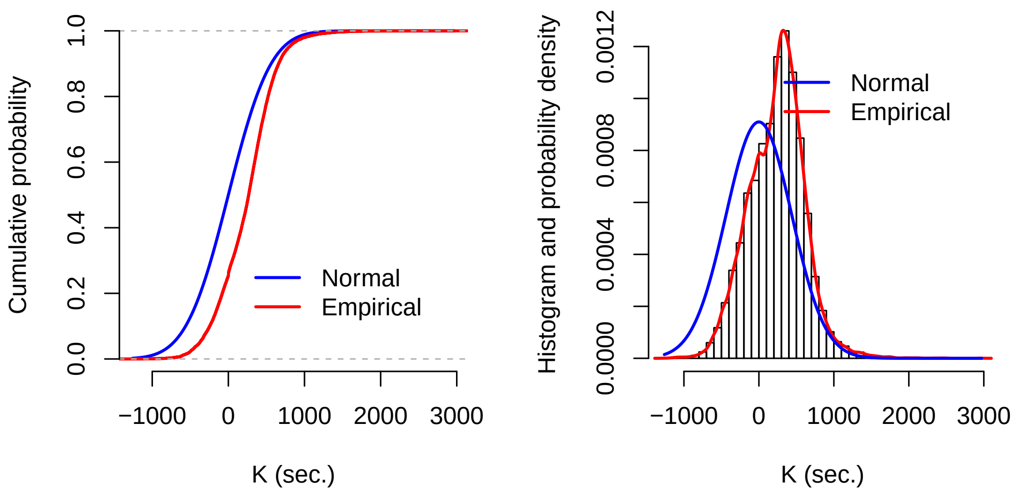

Figure 4.5: Empirical distribution of the fluctuation in rush hours and the fitted normal distribution.

The accuracy of the fitted normal distribution was verified by the Kolmogorov-Smirnov and the Shapiro-Wilk tests. Note that the Kolmogorov-Smirnov test examines suitability of the normal distribution with specific parameters, on the other hand the

Shapiro-Wilk test serves to testing normality without any specified parameters. In the case of the fluctuation in rush hours, Shapiro-Wilk approved the normal distribution on the level of 0.05 in 9 tests out of 10, the Kolmogorov-Smirnov test in 7 cases out of 10, see Table 4.5.

No.	p-value	No.	p-value
1	0.0259	6	0.1662
2	0.1611	7	0.2163
3	0.0124	8	0.0834
4	0.0157	9	0.1526
5	0.0734	10	0.0541

No.	p-value	No.	p-value
1	0.9161	6	0.2596
2	0.4608	7	0.6155
3	0.0843	8	0.1592
4	0.4131	9	0.2351
5	0.0074	10	0.6627

Table 4.5: P-values of Kolmogorov-Smirnov (left) and Shapiro-Wilk tests (right) of a normal distribution of the fluctuation.

4.2.4 Calculation of probabilities w_{ij}

In our computation of values w_{ij} we use one of the methods suggested by Ingolfsson, Budge and Erkut (2008). The probability w_{ij} is defined as the probability that the response time is less than or equal to a given constant δ. The response time is composed of two random variables (the travel time and the delay). Suppose that the distribution function $H_{ij}(t)$ of the travel time T_{ij} from the i-th station to node j as well as the distribution function $F(t)$ for the delay are available. Moreover assume that the travel time and the delay are independent. Now we can use the convolution theorem (see Appendix A) to calculate the probabilities w_{ij}, i.e.,

$$w_{ij} = \mathbb{P}(R_{ij} \leq \delta) = \int_0^\delta T_{ij}(\delta - x)\ \mathrm{d}F(x). \tag{4.1}$$

Note that we use an adaption of the convolution theorem since in our case we integrate over the interval $(0, \delta)$ instead of $(-\infty, \infty)$ (pre-trip delays and travel times cannot take negative values).

4.2.5 Determining the assignment of a hospital to each demand node

Let us denote as $H(j)$ a function that assigns to the j-th node a particular hospital and as t_{jh} the average travel time from the j-th node to the h-th hospital. Note that

t_{jh} are the values provided by the travel time model. We will assume the following order:

$$H(j) = \arg\min_{h \in H} \; t_{jh},$$

where H is a set of regions where the hospitals are located. In other words, the j-th node is always assigned to the nearest hospital. At this point we stress that we made a simplifying assumption that each hospital has unlimited capacity.

4.2.6 Initial estimation of the busy fraction

We assume that the average fraction of time an ambulance is busy (not available to respond to calls) is given by the average server utilization for the z-server queueing system, assuming that the number of lost calls due to queueing is negligible, i.e., $\rho_i = \lambda\tau_i/z$. Here we consider an ambulance vehicle as a server and the z-server queueing system as a system consisting of z ambulance vehicles. The estimation of parameter λ was already described in Subsection 4.2.1. Provided that the patient, after being treated at the primary location, is taken to a hospital, the average service time τ (during which an ambulance is linked to a call) can be broken down to the average travel time to the call from the i-th station $\mathbb{E}T_i^{\text{to call}}$, the average on-scene time $\mathbb{E}T^{\text{on scene}}$, and the average time spent traveling to and remaining at a hospital $\mathbb{E}T^{\text{ to hospital}}$, $\mathbb{E}T^{\text{at hospital}}$, respectively. In case the patient is just treated at his/her location and not taken to a hospital, the average service time τ reduces to the sum of the average travel time to the call from the i-th station $\mathbb{E}T_i^{\text{to call}}$, and the on-scene time $\mathbb{E}T^{\text{on scene}}$ (for illustration see Figure 4.6 and Figure 4.7). Note that the average on-scene time and the time spent traveling to and remaining at a hospital do not depend on the initial location of ambulance, i.e., on index i.

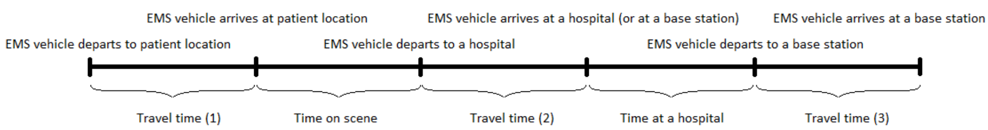

Figure 4.6: Service time provided that the patient is taken to a hospital.

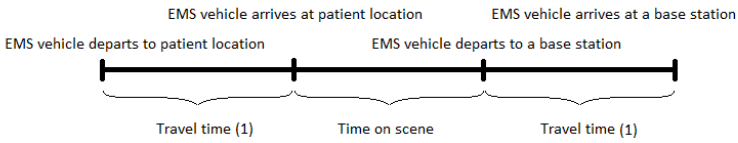

Figure 4.7: Service time provided that the patient is treated at his/her location.

Assume that the probability that a patient is taken to a hospital equals $p \in (0,1)$, then the average fraction of time an ambulance is busy is given as

$$\rho_i = \frac{\lambda}{z}\left(\mathbb{E}T_i^{\text{to call}} + \mathbb{E}T^{\text{on scene}} + p\left(\mathbb{E}T^{\text{ to hospital}} + \mathbb{E}T^{\text{at hospital}}\right)\right).$$

The average travel time to a call and the average travel time to a hospital can be expressed as:

$$\mathbb{E}T_i^{\text{to call}} = \sum_{j \in N} h_j \mathbb{E}T_{ij}$$

$$\mathbb{E}T^{\text{to hospital}} = \sum_{j \in N} h_j \mathbb{E}T_{jh}$$

This leads to the following formula for approximating ρ_i as a function of x:

$$\rho_i(x) = \frac{\lambda}{z(x)}\left(\sum_{j \in N} h_j\left(\mathbb{E}T_{ij} + p\ \mathbb{E}T_j^h\right) + \mathbb{E}T^{\text{on scene}} + p\ \mathbb{E}T^{\text{at hospital}}\right), \tag{4.2}$$

where h_j defined as

$$h_j = \frac{d_j}{\sum_{j \in N} d_j}, \qquad j \in N,$$

is the fraction of the total demand generated at the demand node j.

First we want to emphasize that the number of servers $z(x)$ is a variable. The validity of this formula requires some approximations. In particular, we exclude the time spent traveling back to a station from a hospital or from a patient location from the average service time since the ambulance is available to respond to incoming calls during this time. On the other hand, we assume that an ambulance always sets off from a station and turns back to the same station.

Remark. When applying the exact solution approach, h_j becomes scenario dependent and according to Formula (4.2) ρ_i are scenario dependent as well. Therefore we used notation ρ_j^s in Chapter 3.

Computational experience

Estimations of the random travel times T_{ij} between the i-th station and the j-demand node, and T_j^h between the j-th demand node and the assigned hospital were discussed in Subsection 4.2.3. Estimations of the average time spent on scene, at hospital respectively, are based on the sample mean of realizations of the random variable $T^{\text{on scene}}$, $T^{\text{at hospital}}$ respectively. The parameter p, the probability that a patient is taken to the hospital, is simply computed as the number of calls when the patient was transported to the hospital divided by the overall number of calls. Based on the available data we obtained the following results:

$$\begin{aligned} \mathbb{E}T^{\text{on scene}} &= 1,273.21, \\ \mathbb{E}T^{\text{at hospital}} &= 1,140.19, \\ p &= 0.69. \end{aligned}$$

Remark. Note that the previous results are stated in seconds. All the travel times needed for computing busy fractions have to be converted into hours.

4.3 Scenario generating

Assume that λ_j, the interarrival rate of calls for each demand node, is a known parameter (their estimations were described in Subsection 4.2.1). We know that accidents at each demand node appear according to an independent Poisson process, i.e., the interarrival time between two successive accidents is exponentially distributed and the number of accidents in an interval of length of one unit of time (one hour) obeys a Poisson distribution. When modeling the hourly situation at each demand node we use the following algorithm:

Algorithm 2

STEP 1. Set the initial value for the total number of accidents appeared at demand node j, d_j equal to 0.

STEP 2. Generate the first value x of a random variable with exponential distribution with parameter λ_j and set an auxiliary variable t equal to x mod 1.

STEP 3. Repeat the procedure described in STEP GENERATE while $t < 1$.

STEP 4. If $t \geq 2$ stop the algorithm and return the current value d_j. Otherwise increase the value of d_j by one and apply STEP GENERATE. Repeat this procedure until $t \in [1, 2)$.

STEP GENERATE. Generate a new value of x of a random variable with exponential distribution with parameter λ_j. Set the new value of t as $t + x$.

Since we assumed the independent Poisson arrival process at each node, the probability that the number of accidents at demand node j is equal to the simulated value d_j is given by (see Appendix A):

$$\mathbb{P}(N_j(1) = d_j) = \frac{\lambda_j^{d_j}}{d_j!} \exp(-\lambda_j).$$

We observed that λ_j belongs to interval (0,1) for each $j \in N$, in other words there is no node with more than one accident per hour on average. Therefore the most common value of d_j will be 0 and 1. Moreover we will assume that accidents at a particular node appear independently of accidents happened at the others. The previous assumption allows us to generate scenarios piecewise, i.e., first we model the situation for each demand node separately and then we group the obtained results into the overall scenario. Finally we arrive to the formula expressing the probability of a concrete scenario π^s which has the form:

$$\pi^s = \prod_{j \in N} \mathbb{P}(N_j(1) = d_j^s) = \prod_{j \in N} \frac{\lambda_j^{d_j^s}}{d_j^s!} \exp(-\lambda_j).$$

The assumption of independent behavior for each demand was already confirmed in Subsection 4.2.1 where we showed that the sum of the individual interarrival rates is approximately equal to the overall interarrival rate.

In the majority of our computational experiment presented in the next chapter we handled the set of scenarios of cardinality 5,000. We presume that this set is large enough to model the real-life situation.

Chapter 5

Computational results

This section is devoted to the presentation and discussion of the computational experiments carried out to assess and validate our models. In the first part we discuss the solution of the model formulated as (P2) with a predetermined set of facility locations, the second part deals with the model formulated as (P4). In order to demonstrate the computational performance of our methods, we display results obtained for rush hours. Here we would like to point out that solving the same issue for off-peak hours and night hours is not our main concern as we use another approach for these periods. In the Section 5.3 we present illustrative results to give some insight about the variance of solutions of problems (P2) and (P3). Section 5.4 provides the analysis of overall solution of designing EMS system taking into account the changes due to different time periods.

Remark. All calculations were carried out using R (version 2.12.1), in particular package `lpsolve` which is dedicated to solve linear, integer and mixed integer problems.

5.1 Optimizing with the predetermined set of base station locations

In order to assess the contemporary situation, we have first considered the problem with predetermined locations of the base stations, i.e., we apply the model formulated as (P2). In fact, the set of location points is assumed to be equal to the existing base station locations which are displayed in Table B.7 in Appendix B. As far as capacity of a station is concerned, we considered the same value for the capacity for each candidate facility, i.e., $b_i = \gamma$ for all $i \in M$. Beside specifying the base station locations we need to have the information about the positions of hospitals as to be able to assign the nearest hospital to each demand node. The list of hospitals and their locations can be also found in Appendix B in Table B.6.

5.1.1 Numerical results

The first set of experiments was carried out to demonstrate the results for the usual choice of the critical time δ and the system-wide confidence level α and design the real-life EMS system. The other computations were aimed to confirm the presumption about the system behavior when these parameters exceed their common values. For instance, it seems reasonable that with decreasing value of the critical time δ, the diversification of ambulance vehicles between the base stations is more significant. Equally we could state that with the higher system-wide confidence level α the diversification should not change whereas the number of vehicles in each station should raise.

As indicated in Section 4.2, we assume that each demand node has a circular coverage area and we consider two common values to define the critical time δ of the coverage radius: 10 and 15 minutes. In order to show the influence of the system-wide confidence level on the problem's solution, we have set the value of parameter α equal to 0.95 and 0.99. Finally, each case has been solved for different values of the station capacity γ: 5, 20. In all the tests we have carried out a large number of scenarios (5,000) to ensure the stability of the solution. Table 5.1 summarizes the obtained results for each combination of the above mentioned parameters. Particularly we were interested in the total number of vehicles that satisfy the minimum coverage condition and their exact distribution between the base station locations, further we recorded the number of iterations of the bisection algorithm (denoted as ν) and the fraction of the expected coverage (denoted as ϕ). Note that ϕ has to be always equal to or larger than α. The table shows that locating the ambulance vehicles to the base station in region 8 (Amsterdam Zuid) is the most profitable choice in order to increase the system coverage. Indeed, region 8 has the highest interarrival rate of accidents among all regions considered as the base station locations, moreover it is situated in the middle of the agglomeration, therefore the average reaching distance (expressed in time units) for any EMS call is the shortest. Maintenance of the base stations in regions 11 (Amsterdam Zuidoost), 3 (Amstelveen) and 1 (Kudelstraat, Kalslagen and Aalsmeer) seems to be inconvenient. This fact can be explained in the case of regions 11 and 3 by the close presence of the base station in the region 8, and in the case of the region 1 by a low interarrival rate and its outskirt position. The results show that the optimal solution for both values of the parameter α coincide. Nevertheless we observed that this behavior is symptomatic for the most common choice of the parameters δ and γ, for extremely low values of these parameters the influence of α on the optimal solution is visible (this statement will be proved in the Subsection 5.1.2). The table also shows that the higher is the critical time δ, the lower is the number of required vehicles to ensure the specified fraction of the expected coverage. As expected, the upper bound on the number of vehicles that can be allocated to base stations has a significant effect on the number of facilities to be opened. When the capacity variable γ is set to be a low value the model results in more facilities, concretely the station in region 4 (Amsterdam Centrum) is opened. Since region 4 has lower interarrival rate of accidents than region 8 and is situated closer to the borderline of the agglomeration, the provided coverage decreases with disaggregation of the vehicles.

α	δ	γ	Number of vehicles					ϕ	ν
			Reg. 4	Reg. 8	Reg. 11	Reg. 3	Reg. 1		
0.95	10	5	3	5	0	0	0	1.09	6
0.95	10	20	0	8	0	0	0	1.12	8
0.99	10	5	3	5	0	0	0	1.09	6
0.99	10	20	0	8	0	0	0	1.12	8
0.95	15	5	2	5	0	0	0	1.04	5
0.95	15	20	0	7	0	0	0	1.07	7
0.99	15	5	2	5	0	0	0	1.04	5
0.99	15	20	0	7	0	0	0	1.07	7

Table 5.1: Optimal number and location of ambulance vehicles for different values of parameters α, δ, γ when locations of the base stations are predetermined.

Let us now pass to prove the stability of the proposed solving method, i.e., we want to show that the optimal solution does not depend on the set of generated scenarios (we used again the set of cardinality 5,000). As can be seen from Table 5.2, the particular set of scenarios influences negligibly the provided coverage ν whereas the total number of vehicles and their distribution stay constant. We remark that the other parameters were kept fixed on values $\alpha = 0.95$, $\delta = 15$ min., $\gamma = 20$. In our initial computations where we assumed the set of demand nodes to be equal to all postal codes in the Agglomeration Amsterdam this was not the case.

	Number of vehicles					ϕ	ν
	Reg. 4	Reg. 8	Reg. 11	Reg. 3	Reg. 1		
1	0	7	0	0	0	1.05	7
2	0	7	0	0	0	1.07	7
3	0	7	0	0	0	1.06	7
4	0	7	0	0	0	1.06	7
5	0	7	0	0	0	1.05	7
6	0	7	0	0	0	1.05	7
7	0	7	0	0	0	1.06	7
8	0	7	0	0	0	1.06	7
9	0	7	0	0	0	1.07	7
10	0	7	0	0	0	1.06	7

Table 5.2: Stability of the optimal solution when locations of the base stations are predetermined.

5.1.2 Solution sensitivity to input parameters

In order to provide an insight on about the sensitivity of the model to changes in input parameters, we present in the following section several graphical illustrations. The values of only a certain type of parameters are changed while the remaining ones are kept fixed. Namely, we were interested in investigating dependence of the optimal solution on the number of generated scenarios, the system-wide confidence level α, the critical time δ and the capacity of the base stations γ.

Let us first discuss the sensitivity of the optimal solution on the number of scenarios. We altered the number of scenarios from 0 to 5,000 with a step 20 while keeping the other parameters fixed on the values: $\alpha = 0.95$, $\delta = 15$ min. and $\gamma = 20$. Figure 5.1 displays the total number of active stations and vehicles in the system, the ratio between the expected and provided coverage and the number of iterations of the bisection algorithm associated with the concrete number of scenarios. The total number of stations together with the total number of vehicles assigned to cover demand as well as the number of iterations are consistent, whereas the the ratio between the expected and provided coverage stabilizes around the value 1.058 as soon as the number of scenarios exceeds 1 500. For smaller values we observed significant changes in the ratio ϕ. Therefore we can conclude that the method converges to the unique optimal solution.

Other experiments were carried out to analyze the effect of the system-wide confidence level on the optimal results. Figure 5.2 reports the total number of stations and vehicles, ratio ϕ and the number of iterations ν versus the confidence level α. The calculation of the optimal solution was realized for each element of the uniform sequence in the interval $[0.1, 1]$ with step size 0.01 while the other parameters were kept fixed on values: $\delta = 2$ min., $\gamma = 20$ and the number of scenarios equal to 5,000. As expected, the higher the required reliability value α is, the higher is the number of vehicles needed to ensure the emergency medical case, moreover the dependence is almost linear. Since we restricted the capacity γ to the value of 20, the number of active stations raises with higher number of vehicles in the system. The fraction ϕ traces exactly the value of parameter α. This behavior can be explained by keeping parameter δ fixed on a very low value. In this case the overall number of vehicles in the system is in general much higher for any confidence level α (in comparison with previous computations where δ was equal to 10 or 15 minutes), therefore the number of vehicles which influences the ratio ϕ is more sensitive to the variation of the reliability value. The number of iterations fluctuates between 7 and 8, thus we can conclude that ν does not depend on the parameter α.

The next figure (5.3) displays the behavior of the optimal solution in relation to parameter δ. We compared the results for nearly 120 different values of δ, we started at the value of 1 minute and raised it with every new computation by 15 seconds up to the value of 30 minutes. For the other parameters we chose standard values, i.e., $\alpha = 0.99$, $\gamma = 20$ and we involved 5,000 scenarios. We can observe that for very low values of δ the optimization problem does not have any solution. As soon as δ exceeds the threshold of 2 minutes, a network that satisfies all conditions can be designed. However, the number of vehicles in such a system is enormously high. The higher δ is, the lower number of vehicles is required to respond to emergency calls and the total number of active stations increases as well. Now let us focus on the explanation of the shape of the curve of the ratio ϕ. As it was mentioned in the previous paragraph, in the case of low δ the optimal number of vehicles is extremely high and even small changes in the critical time leads to amending the number of ambulances. Jumps of the ratio ϕ exactly reflect the changes in the number of vehicles. Therefore for smaller values of δ they occur frequently whereas for higher sparsely. The total number of iterations again remained independent of the parameter δ.

Finally, we compared the results for different bounds of the capacity of stations γ. As we can observe from Figure 5.4, the optimal solution for the parameter γ lower than 3 does not exist (the rest of the system parameters were set as: $\alpha = 0.95$, $\delta = 5$ min.

and we again used 5,000 scenarios). If we solve the optimization problem with such specified parameters and unlimited capacity of stations we arrive at 11 vehicles needed to ensure the required medical care. Therefore when the capacity γ is set to be equal to or higher than 11 ambulances per station the optimal solution remains stable, i.e., all vehicles are settled in one particular station (in our case it is the station in region 8). For other values of γ $(3, 4, \ldots, 11)$ a distribution between stations is required, thus we can observe either higher number of vehicles in the system or a lower ratio ϕ. In contrast with previous computations, we can conclude that the number of iterations of the bisection algorithm depends on the capacity γ. This behavior can be explained by setting the initial value of vehicles in the system b^1 introduced in *Algorithm 1* as a sum of the station capacities, i.e., the higher γ is, more iterations are necessary to find the optimal solution.

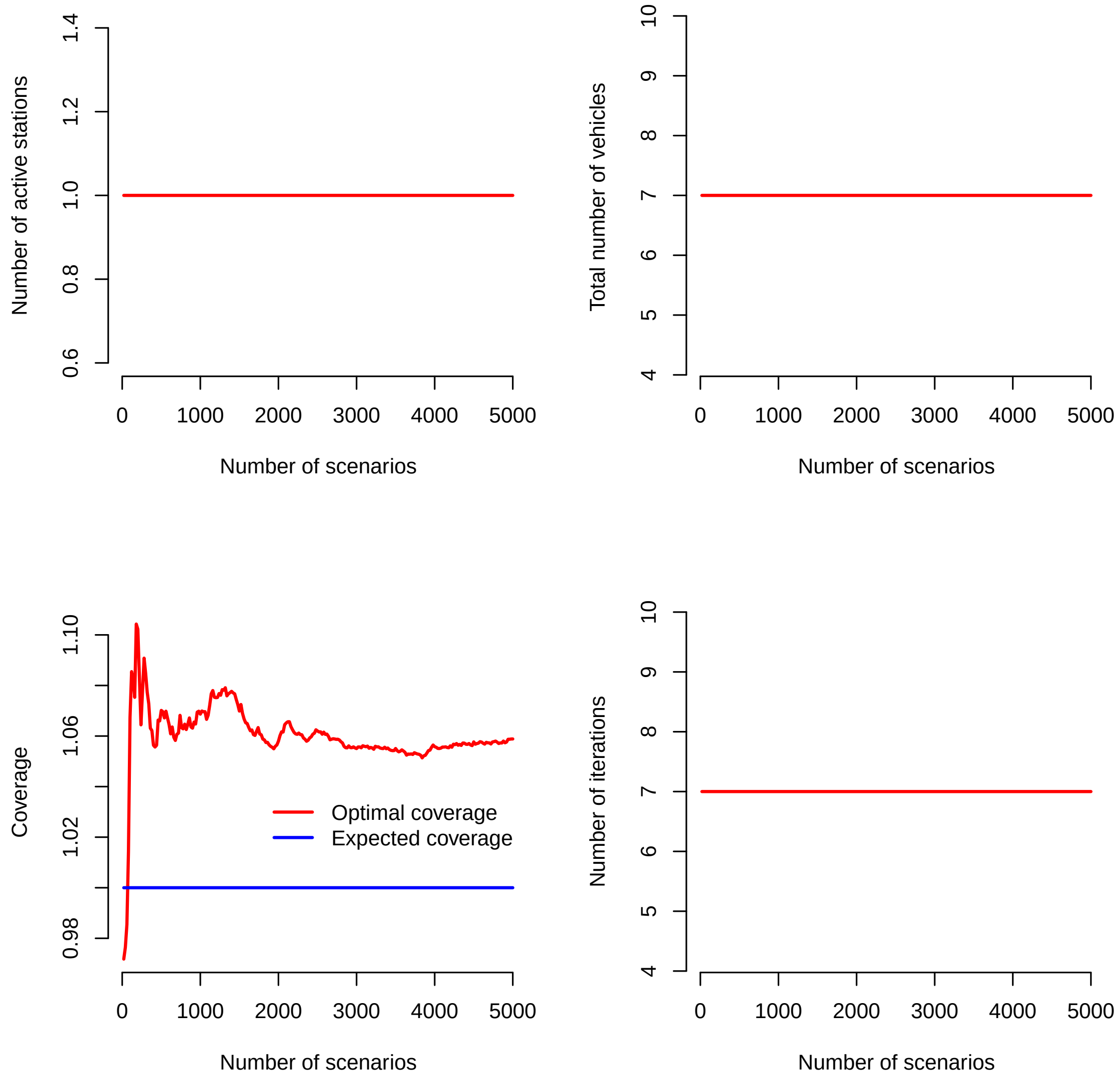

Figure 5.1: Sensitivity of the total number of active stations and vehicles, the coverage and the number of iterations to the number of scenarios when locations of the base stations are predetermined.

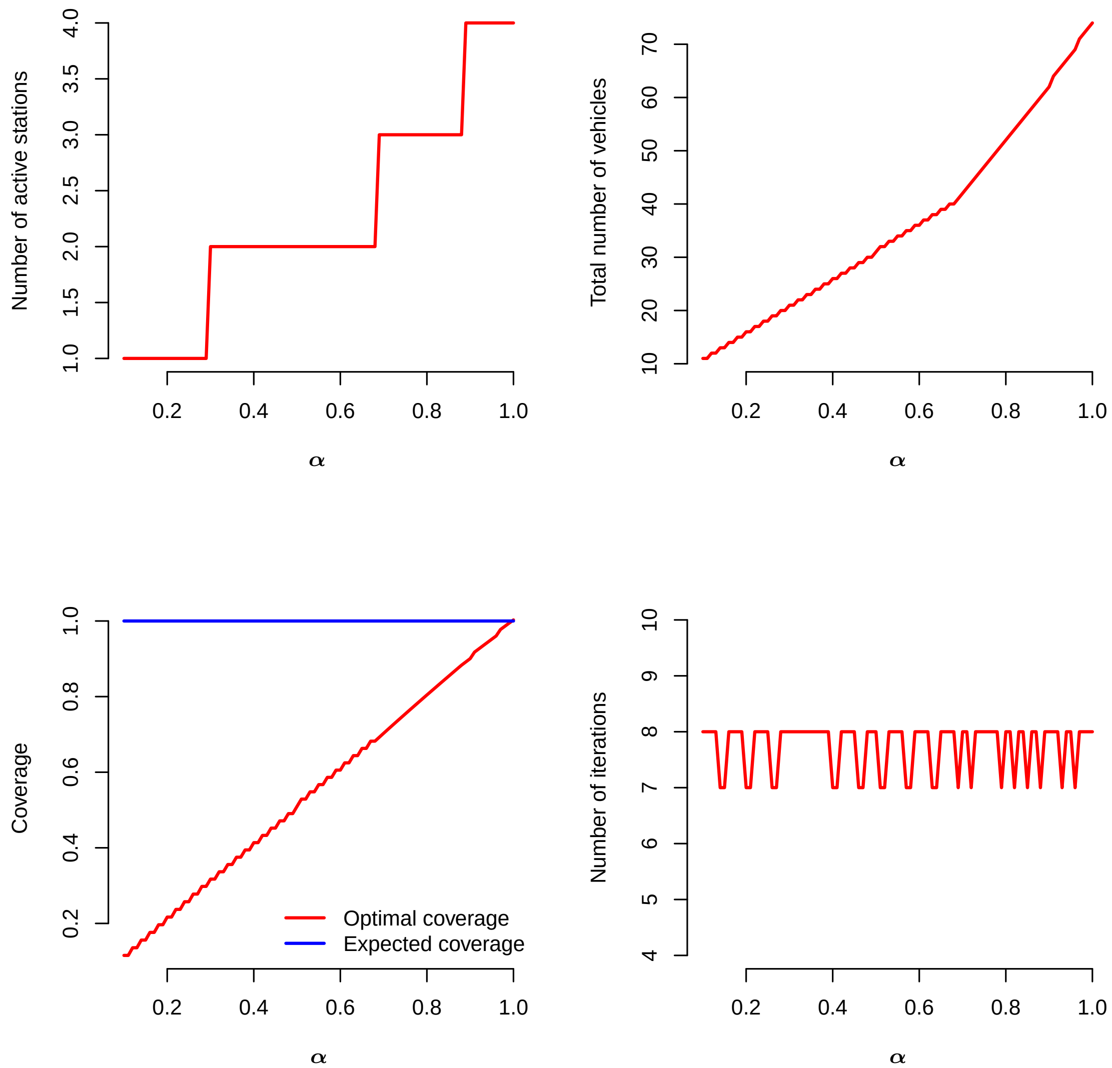

Figure 5.2: Sensitivity of the total number of active stations and vehicles, the coverage and the number of iterations to the parameter α when locations of the base stations are predetermined.

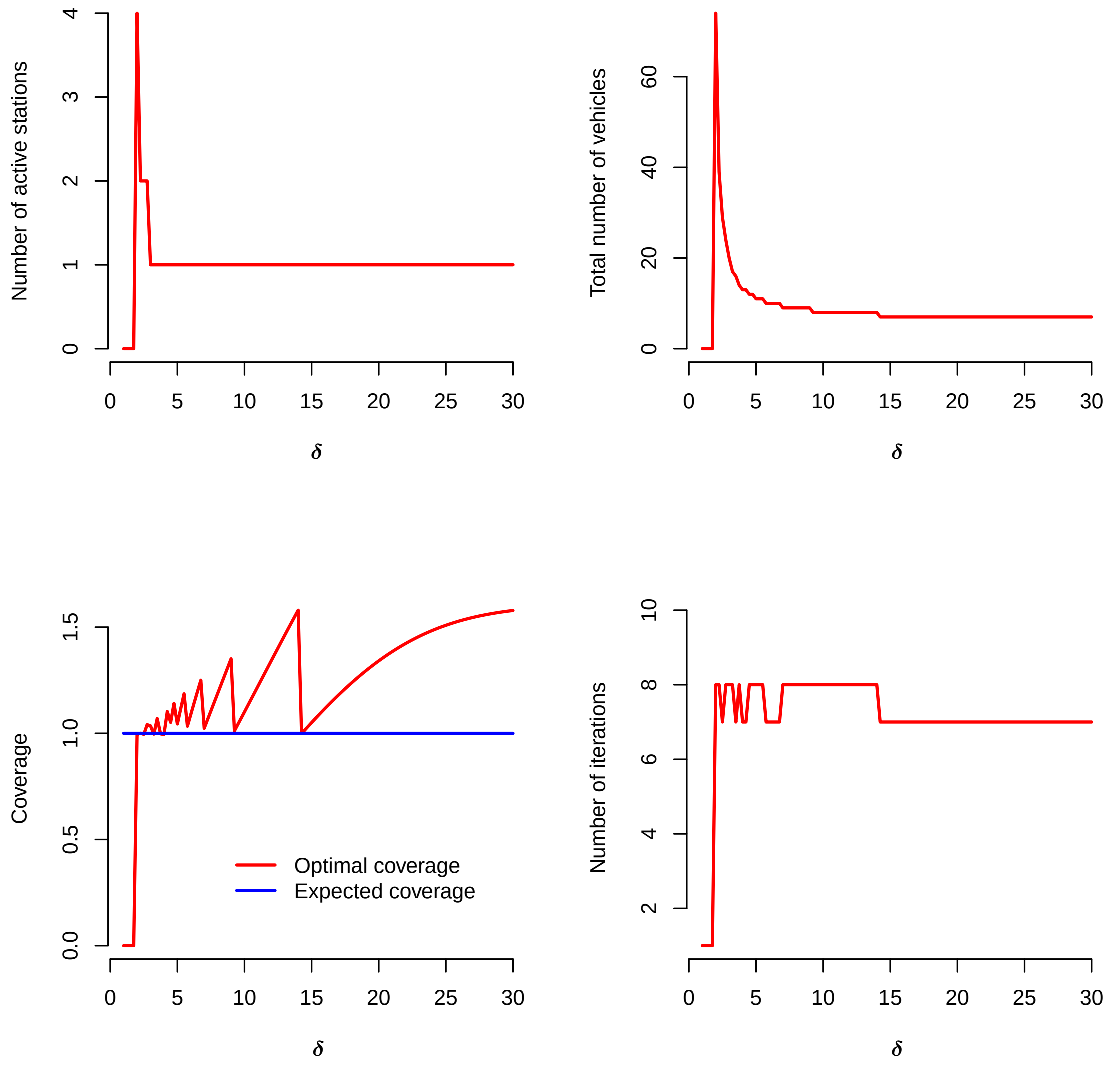

Figure 5.3: Sensitivity of the total number ofactive stations and vehicles, the coverage and the number of iterations to the parameter δ when locations of the base stations are predetermined.

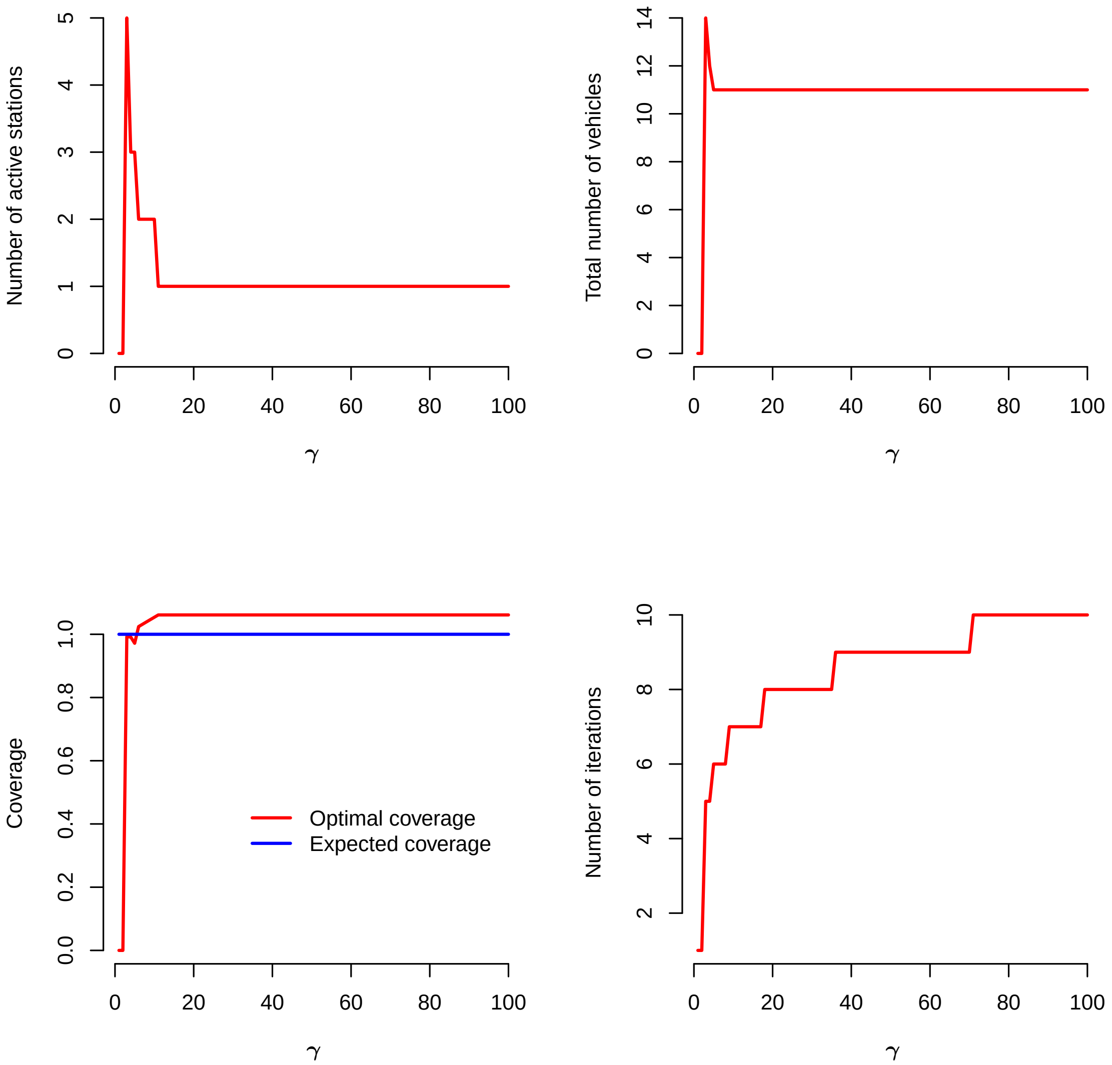

Figure 5.4: Sensitivity of the total number of active stations and vehicles, the coverage and the number of iterations to the capacity γ of the base station when locations of the base stations are predetermined.

In the previous part we discussed sensitivity of the optimal solution on the input parameters, in particular we observed changes in the optimal number of vehicles, in the ratio ϕ and in the number of iterations ν. However we did not mention how the distribution of vehicles between stations differs with varying these parameters. First let us focus on the case of the system-wide confidence level α. In order to draw a conclusion about behavior of the optimal solution we set the parameter δ equal to 2 minutes and combined different values of α and γ. It seems reasonable that the higher is the reliability level α, the higher is the number of ambulances in the system, but their locations should remain consistent. Indeed, Table 5.3 gives the evidence of the previous statement. The algorithm first fulfills the capacity of the facility location that maximizes the provided coverage and as soon as the boundary is met it targets on the other base stations. The other experiment was carried out to investigate changes in the vehicle location depending on the critical time δ. Unlike the previous case, one would assume that the lower δ is, the vehicles scatter throughout the net of the base stations. Nevertheless the results that are stated in Table 5.4 do not confirm our presumptions (we remark that for each instance we used the confidence level equal to 0.95). With increasing value of δ the number of vehicles descends though the distribution does not change dramatically. When we applied the same experiment on the model where we divided the region of Amsterdam according to its postal codes instead of using partition on 19 districts, we observed a variation in distribution of vehicles. The lower the critical time was, the more facility locations were dedicated to serve as the base stations even though the capacity of none of them was fully filled. We can conclude that distribution of ambulances highly depends on the partition of the examined region and on the diversity in the interarrival rates.

γ	α	Total number of vehicles	Number of vehicles				
			Reg. 4	Reg. 8	Reg. 11	Reg. 3	Reg. 1
100	0.80	45	0	45	0	0	0
	0.85	47	0	47	0	0	0
	0.90	50	0	50	0	0	0
	0.95	52	0	52	0	0	0
	1.00	54	0	54	0	0	0
50	0.80	45	0	45	0	0	0
	0.85	47	0	47	0	0	0
	0.90	50	0	50	0	0	0
	0.95	53	3	50	0	0	0
	1.00	55	5	50	0	0	0

	0.80	52	20	20	0	12	0
	0.85	57	20	20	0	17	0
20	0.90	62	20	20	2	20	0
	0.95	68	20	20	8	20	0
	1.00	74	20	20	14	20	0

Table 5.3: Distribution of ambulance vehicles for different values of α when locations of the base stations are predetermined.

γ	δ (min)	Total number of vehicles	Number of vehicles				
			Reg. 4	Reg. 8	Reg. 11	Reg. 3	Reg. 1
	1.65	110	10	100	0	0	0
	1.7	96	0	96	0	0	0
100	1.8	77	0	77	0	0	0
	2.5	28	0	28	0	0	0
	10	8	0	8	0	0	0
	1.65	127	50	50	0	27	0
	1.7	100	50	50	0	0	0
50	1.8	79	29	50	0	0	0
	2.5	28	0	28	0	0	0
	10	8	0	8	0	0	0
	1.65	no solution					
	1.7	no solution					
20	1.8	no solution					
	2.5	28	8	20	0	0	0
	10	8	0	8	0	0	0

Table 5.4: Distribution of ambulance vehicles for different values of δ when locations of the base stations are predetermined.

5.2 Optimizing throughout unlimited base station locations

In the following section we present and discuss the computational results carried out to assess and validate the model formulated as (P4), i.e., we want to design a system with optimal base station locations and assign them an adequate number of vehicles. We assume that a station can be established in any region, in other words we defined the set of the facility locations to be the same as the set of demand nodes. The choice of capacities for potential base stations is up to the system planner who might also impose different restrictions for the different location points according to the geographical and construction conditions. For simplicity in our calculations we considered the same value of the capacity for each facility location (we again used the notation γ). We remark that the set of hospitals remained the same and thus the assignment of the nearest hospital to each demand node did not change.

5.2.1 Numerical results

In the first subsection we present a design of the system for common choice of input parameters. We again observed the results for all instances where $\alpha = 0.95, 0.99$, $\delta = 10$ or 15 min. and $\gamma = 5, 20$. In each of the instances we utilized 5,000 of scenarios and restricted the maximal number of stations in the system β to 10. We refer to Table 5.5 to view the obtained results. In the column Optimal number and locations of stations we first display the number of operating stations, in the second row we state the serial number corresponding to the region where the stations are supposed to be established. The next column summarizes the overall number of ambulance vehicles and their exact distribution between the stations underneath. In addition we were interested in the number of iterations of the bisection algorithm ν and the ratio between provided and expected coverage ϕ. We can conclude that establishing a base station in region 8 (Amsterdam Zuid) is again the most profitable in the sense of increasing the system coverage. When the capacity of this station is fully filled and the target level of the emergency medical service is not met, a station on the second prefered position (a station in region 4) comes to use. It is worthwhile to emphasize that the current system takes into account stations in regions 4 and 8, and assigns them the majority of operating vehicles. The accuracy of this approach was confirmed by our computations. Moreover we demonstrated that stations in regions 1, 3 and 11 are dispensable for an efficient system since they did not appear in the results. As we discussed in the previous part of computational results, the higher the critical response time δ is, less ambulances are involved to ensure the specified fraction of the expected coverage. The number of base stations that should be opened is highly influenced by the upper bound on the number of vehicles that can be settled at each facility location, i.e., when γ is relatively low, the model results in more operating stations.

α	δ	γ	Optimal number and locations of stations	Optimal number of vehicles	ϕ	ν
0.95	10	5	2	8	1.08	7
			(4, 8)	(3, 5)		
0.95	10	20	1	8	1.11	9
			(8)	(8)		
0.99	10	5	2	8	1.08	7
			(4, 8)	(3, 5)		
0.99	10	20	1	8	1.11	9
			(8)	(8)		
0.95	15	5	2	7	1.04	6
			(4, 8)	(2, 5)		
0.95	15	20	1	7	1.06	8
			(8)	(7)		
0.99	15	5	2	7	1.04	6
			(4, 8)	(2, 5)		
0.99	15	20	1	7	1.06	8
			(8)	(7)		

Table 5.5: Optimal number and location of ambulance vehicles for different values of parameters α, δ, γ.

As in the previous part we give an evidence that the solution does not depend on the set of generated scenarios. We kept the input parameters fixed on values $\alpha = 0.95$, $\delta = 15$ min., $\gamma = 20$ and $\beta = 10$, we simulated 5,000 scenarios and recorded the results. Afterwards we repeated this procedure several times. Table 5.6 shows that different set of scenarios only causes negligible deviation in the ratio ϕ whereas the optimal solution remains the same.

	Optimal number and locations of stations	Optimal number of vehicles	ϕ	ν
1	1 (8)	7 (7)	1.06	8
2	1 (8)	7 (7)	1.07	8
3	1 (8)	7 (7)	1.06	8
4	1 (8)	7 (7)	1.06	8
5	1 (8)	7 (7)	1.06	8
6	1 (8)	7 (7)	1.05	8
7	1 (8)	7 (7)	1.06	8
8	1 (8)	7 (7)	1.07	8
9	1 (8)	7 (7)	1.06	8
10	1 (8)	7 (7)	1.05	8

Table 5.6: Stabiltiy of the optimal solution.

5.2.2 Solution sensitivity to input parameters

In order to give an idea about the sensitivity of the model to changes in input parameters, we again present several graphical illustrations. As in Subsection 5.1.2, we investigated dependence of the optimal solution on the number of scenarios, the system-wide confidence level α, the critical respond time δ, the capacity of the base station γ and additionally we add sensitivity analysis to the maximal acceptable number of stations in the system β. Beside the optimal number of vehicles, the fraction ϕ and the number of iterations ν we also recorded for each instance the total number of base stations in the system. Since the large share of the experiments arrived to the same conclusion as in Subsection 5.1.2, we discuss here mainly differences and supplementary observations.

First of all we focus on the sensitivity of the optimal solution on the number of scenarios. We generated sets of scenarios of different cardinality ranged from 0 to 5 000 with a step 20 whereas the other parameters remained on values: $\alpha = 0.95$, $\delta = 15$ min., $\gamma = 20$ and $\beta = 10$. As we can observe from Figure 5.5 in all cases one base station with consistent number of vehicles is required to cover the specified fraction of demand. Also in this case the number of scenarios influences the ratio ϕ which converges to its optimal value as the number of scenarios goes to infinity. We can again confirm that the number of iterations does not depend on the number of scenarios as it did not change during the whole experiment.

In the second experiment we were interested in the effect caused by changes in the system-wide confidence level. Figure 5.6 displays the obtained results. We solved the optimization model for all α starting at the value of 0.1, raising it by 0.01 in each step till the value of 1, while the other parameters were equal to $\delta = 2$ min., $\gamma = 20$, $\beta = 10$ and the number of scenarios was equal to 5,000. As before, the higher the required reliability value α is, the higher is the number of vehicles needed to ensure the emergency medical care. Moreover, as we set the maximal capacity of the station equal to 20, the number of operating stations raises with increasing values of α. The ratio ϕ again follows the values of the system-wide confidence level, this behavior was already explained in Subsection 5.1.2. In conclusion we note that the number of iterations did not depend on the value of α.

Further we observed the behavior of the optimal solution in relation to the critical time. We ranged δ from 1 minute to 30 minutes with a step of 15 seconds. For the other parameters we chose standard values ($\alpha = 0.99$, $\gamma = 20$, $\beta = 10$ and the number of scenarios equal to 5 000). The Figure 5.7 shows that for very low values of δ there is no optimal solution. As soon as δ is higher than 2, a network that would provide the requested level of service can be designed. However the number of vehicles in such a system is extremely high and the boundary for the maximal number of stations in the network is met. The higher this parameter is, the lower number of vehicles is involved and the number of base stations reduces as well. The changes of the ratio ϕ were already explained in detail before and therefore we refer reader to Subsection 5.1.2.

The next figure (Figure 5.8) illustrates dependence of the optimal solution on the maximal capacity of base stations. The parameter γ took values $1, 2, \ldots, 100$ whereas the other ones were kept fixed, $\alpha = 0.95$, $\delta = 5$ min., $\beta = 10$ and number of scenarios equal to 5 000. The conclusions are nearly identical as in the Subsection 5.1.2. However in this case the optimal solution exists even for $\gamma = 2$ which is reasonable result with respect to the fact that the total number of active stations in the system can be as high as 10. The number of stations and vehicles gradually increases till the capacity γ reaches the value of 11, then the solution stabilizes on one optimal station with 11 vehicles.

In addition we attach the sensitivity analysis which deals with the dependence of the optimal solution on the maximal number of stations in the network β. We examined all possible values of these parameter limited by the number of considered districts, i.e., up to the 19 facility locations. The other parameters were set as follows: $\alpha = 0.95$, $\delta = 2$ min., $\gamma = 10$, the number of scenarios 5,000. Figure 5.9 shows that the optimal solution for such specified values of parameters required 65 vehicles and thus we need to have at least 7 stations in the network to able to base them. Therefore when $\beta \leq 6$ there is no optimal solution. It is worthwhile to remark that in the case when δ is low (2 min.) we can observe that the vehicles are optimally based in 13 stations whose capacities are not filled. In other words the ambulances are spread over the network to ensure the adequate medical care. The total number of vehicles in the system together with the ratio ϕ stabilizes on the optimal value as soon as the problem has any solution. The number of iterations in order to reach the optimal solution increase as β raises which is a reasonable behavior with respect to the choice of the initial estimation of number of vehicles b^1 defined in Section 3.4.

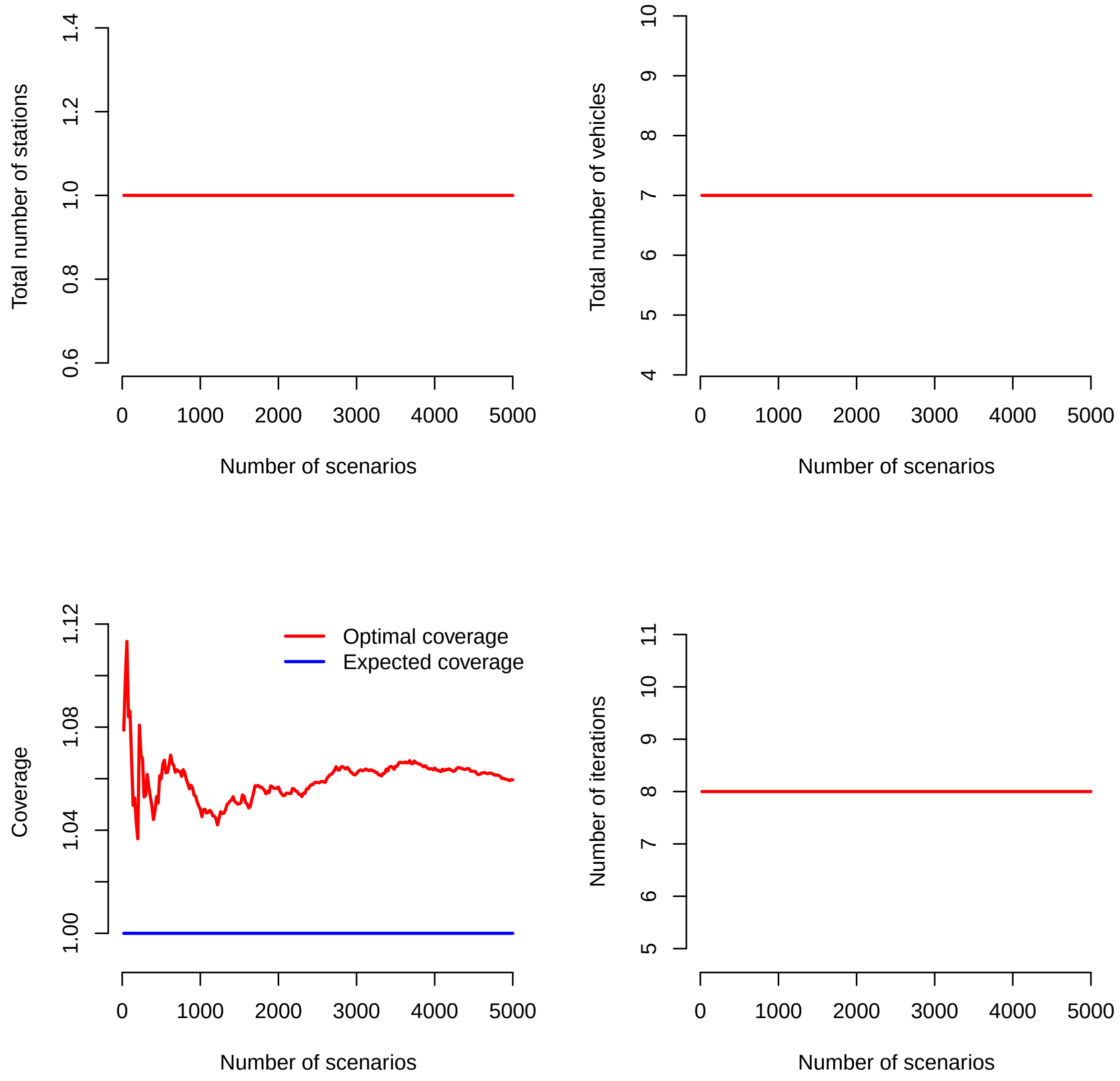

Figure 5.5: Sensitivity of the total number of stations and vehicles, the coverage and the number of iterations to the number of scenarios.

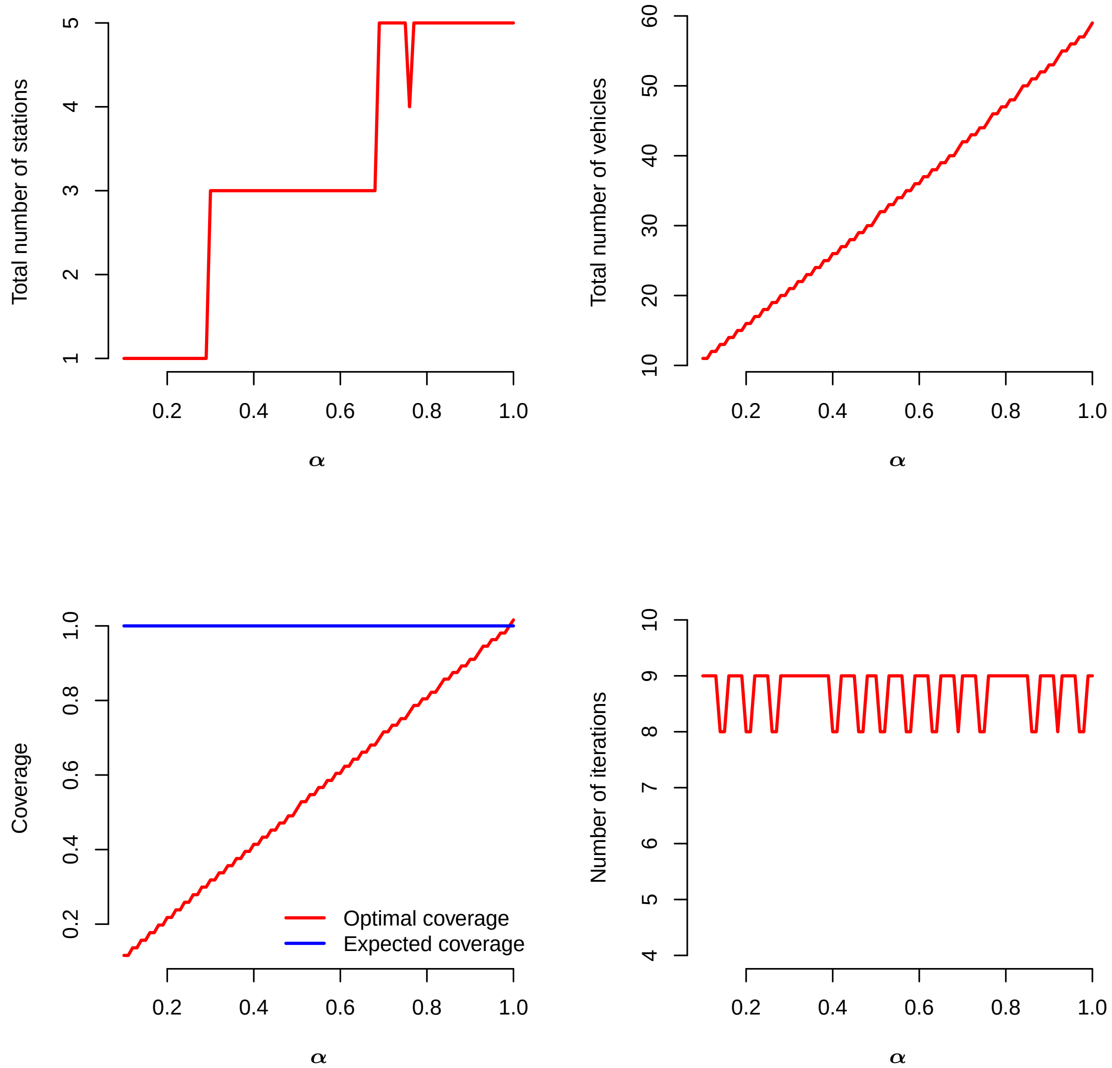

Figure 5.6: Sensitivity of the total number of stations and vehicles, the coverage and the number of iterations to the parameter α.

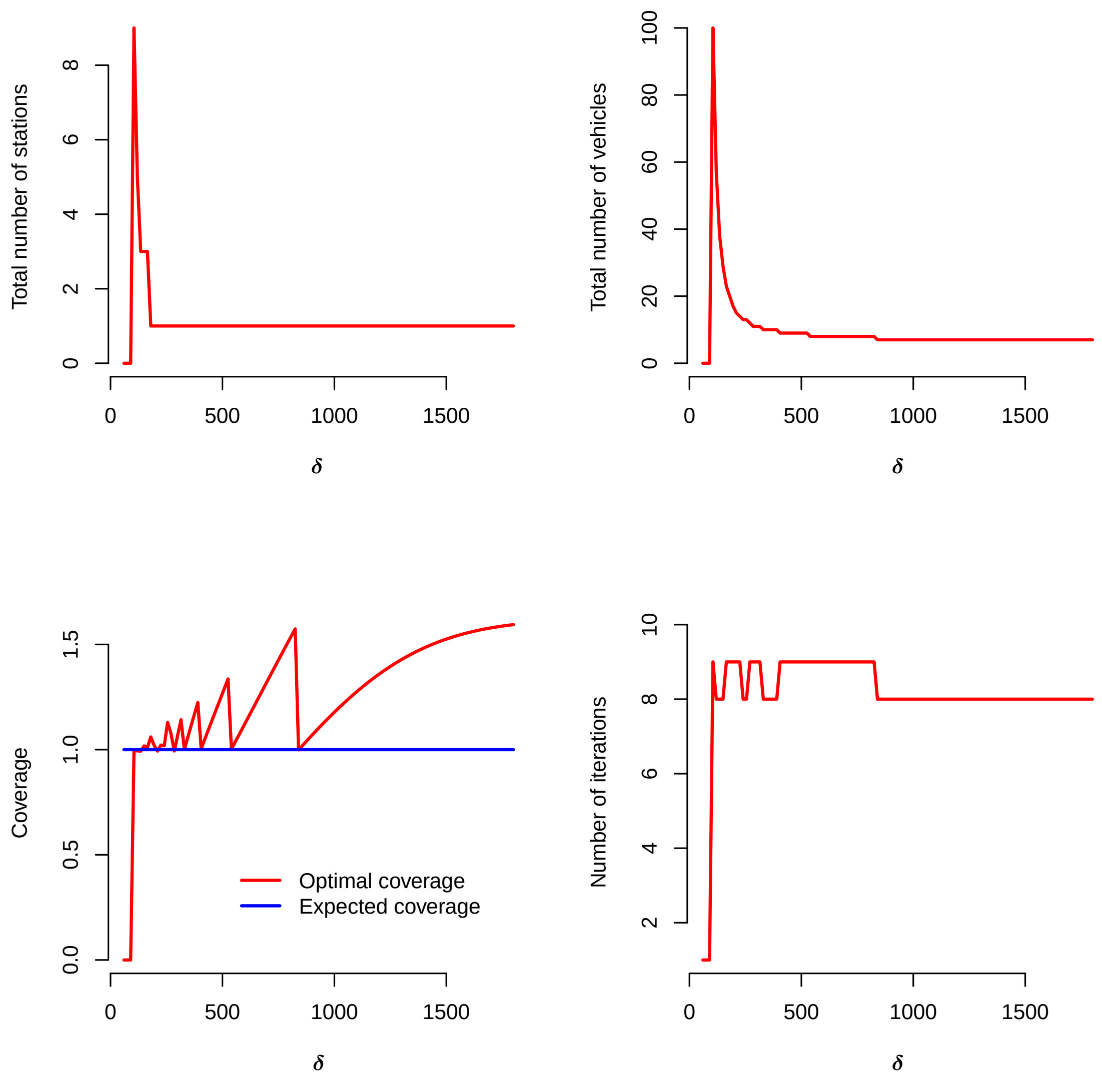

Figure 5.7: Sensitivity of the total number of stations and vehicles, the coverage and the number of iterations to the parameter δ.

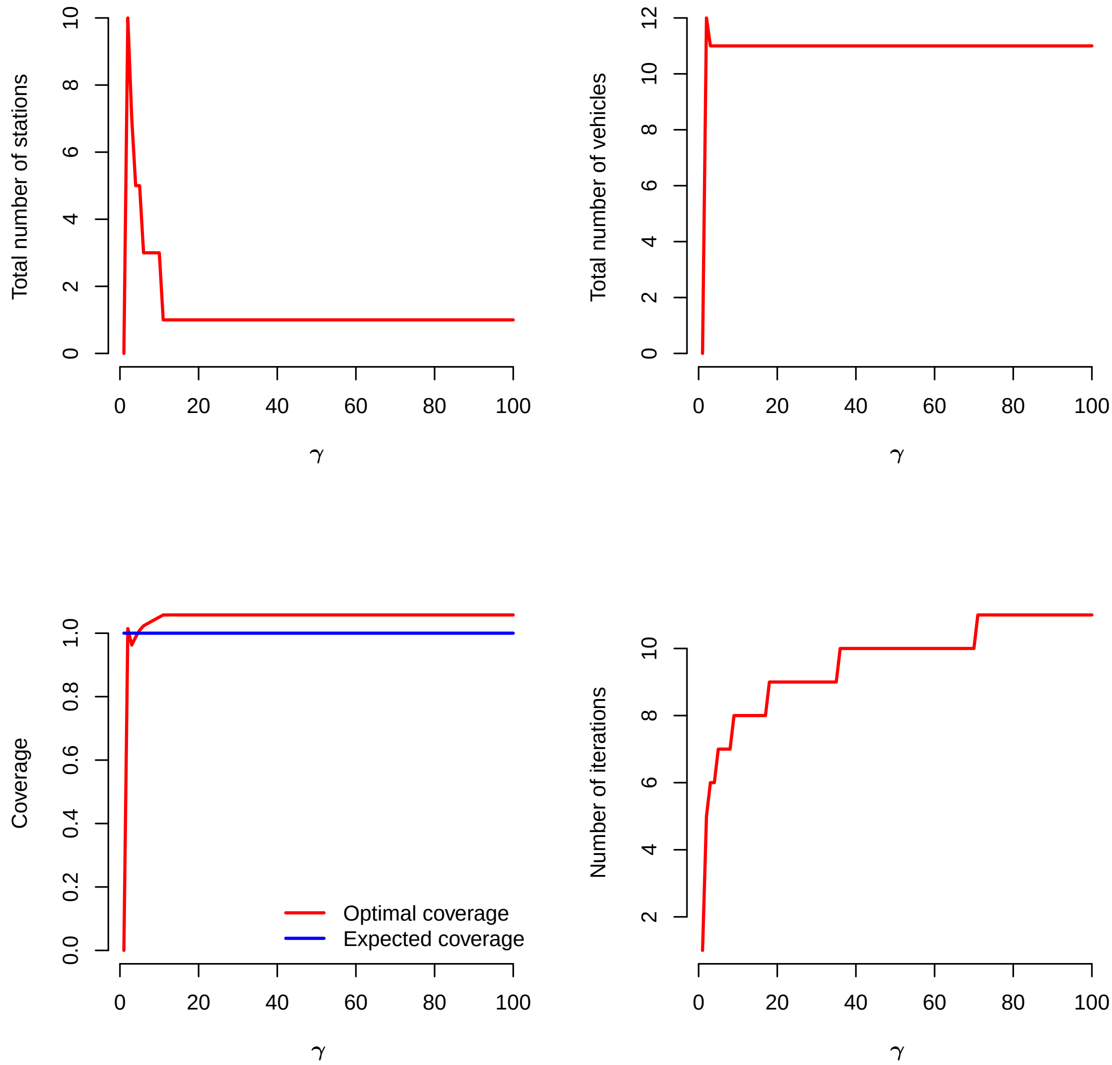

Figure 5.8: Sensitivity of the total number of stations and vehicles, the coverage and the number of iterations to the capacity of the base station.

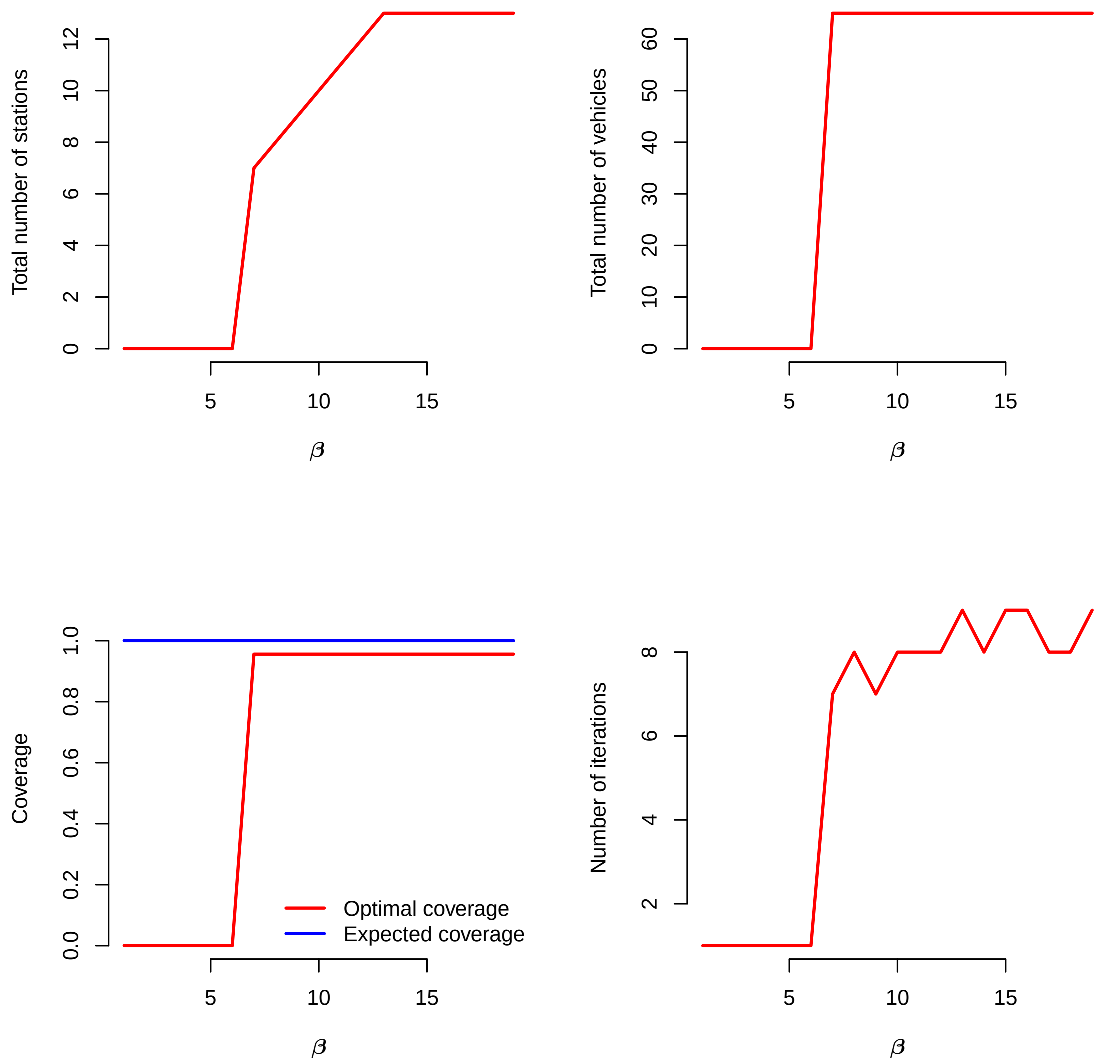

Figure 5.9: Sensitivity of the total number of stations and vehicles, the coverage and the number of iterations to the parameter β.

In the last part of this subsection we explore the distribution of vehicles throughout the network related to low value of the critical respond time. Results in Tables 5.7 and 5.8 are not surprising, we observed the same interpretation in the case where the base stations set was predetermined. In the first experiment we defined the value of δ equal to 2 minutes, β equal to 10 and we took different combinations of α and γ. As the first mentioned table shows, the higher α is, the higher is the number of ambulances but their distribution between stations does not change. The other table gives evidence that with lower δ, the vehicles scatter throughout the net of the base stations. In all instances the algorithm fills the capacity of the facility location that maximizes the ratio ϕ. As soon as the boundary has met it targets on the other base stations. We want to highlight that the third prefered facility location is station in region 7 (Amsterdam Nieuw-West) which is not in the list of current stations. Our computations proved that establishing a station in region 7 instead of maintaining any station in region 1, 3, or 11 would improve the provided medical care.

γ	α	Optimal number and locations of stations	Optimal number of vehicles
100	0.80	1 (8)	46 (46)
	0.85	1 (8)	48 (48)
	0.90	1 (8)	51 (51)
	0.95	1 (8)	53 (53)
	1.00	1 (8)	56 (56)
50	0.80	1 (8)	46 (46)
	0.85	1 (8)	48 (48)
	0.90	2 (4, 8)	51 (1, 50)
	0.95	2 (4, 8)	53 (3, 50)
	1.00	2 (4, 8)	56 (6, 50)

20	0.80	3 (4, 7, 8)	48 (20, 8, 20)
	0.85	3 (4, 7, 8)	50 (20, 10, 20)
	0.90	3 (4, 7, 8)	53 (20, 13, 20)
	0.95	3 (4, 7, 8)	56 (20, 16, 20)
	1.00	3 (4, 7, 8)	59 (20, 19, 20)

Table 5.7: Diversification of ambulance vehicles for different values of the parameter α.

γ	δ	Optimal number and locations of stations	Optimal number of vehicles
100	1.65	2 (4, 8)	110 (10, 100)
	1.7	1 (8)	97 (97)
	1.8	1 (8)	77 (77)
	2.5	1 (8)	28 (28)
	10	1 (8)	8 (8)
50	1.65	3 (4, 7, 8)	115 (50, 15, 50)
	1.7	2 (4, 8)	100 (50, 50)
	1.8	2 (4, 8)	79 (29, 50)
	2.5	1 (8)	28 (28)
	10	1 (8)	8 (8)

20	1.65	8 (3, 4, 6, 7, 8, 9, 10, 13)	142 (2, 20, 20, 20, 20, 20, 20, 20)
	1.7	6 (4, 6, 7, 8, 9, 13)	115 (20, 20, 20, 20, 20, 15)
	1.8	5 (4, 6, 7, 8, 9)	84 (20, 20, 20, 20, 4)
	2.5	2 (4, 8)	28 (8, 20)
	10	1 (8)	8 (8)

Table 5.8: Diversification of ambulance vehicles for different values of the parameter δ.

5.3 Discussion on the current EMS system

The following section addresses the assessment of the contemporary EMS system in the Agglomeration of Amsterdam. Here we focus only on the instance of rush hours. Although the same procedure could be applied to off-peak hours and night hours, as it was mentioned in the previous part, our approach to those periods is different and it will be discussed in detail in the last part of this chapter.

Nowadays there are 5 base stations (Table B.7 Appendix B) where 31 ambulance vehicles are settled during rush hours. However only a fraction of them is dedicated to respond the emergency calls, the rest serves to so-called planned transport, i.e., transport of patients between hospitals or their home, relocating of vehicles under the dispatching logistics. Due to our aim of designing and assessment of the EMS system we take into account just the part of vehicles operating as emergency, thus we arrive at a number 12. The exact distribution of those vehicles between the stations is stated in Table B.8 in Appendix B). At this point we remark that our partition of the day into 3 periods slightly differs from the one made by the current system planner, anyway in further computations we relax this fact.

We have carried out 3 experiments, reaching the optimal solution in all the cases in an exact way. With the first one we intended to show that a mere redistribution of current number of vehicles in the system (12) increases the provided coverage and therefore leads to designing a more reliable net of EMS system. In our test we applied the model formulated as (P2) with fixed number of vehicles in the system equal to 12, we solved this optimization problem without running the bisection algorithm. We refer to the first experiment as Problem 1. In the second experiment (Problem 2) we basically used the same solution method as in the first part of this chapter. We predetermined the set of optimal base stations to be the same as the current locations of stations and found the minimal number of ambulances providing specified fraction of the coverage. The objective of the test was to prove that by a mere reduction of the overall number of vehicles one can achieve the requested coverage. The third experiment (Problem 3) refers to Section 5.2 where we deal with the optimal solution in the spirit of optimal station locations.

Table 5.9 and Figure 5.10 summarize the results of the experiments. In all the computations the system parameters were specified as follows: the confidence level $\alpha = 0.95$, the critical time $\delta = 15$ min., the capacity of base stations $\gamma = 20$ and maximal allowed number of stations in the system in Problem 3 $\beta = 10$. In each case we used a set of scenarios of cardinality 5,000. From the first solution we can conclude that the number of 12 vehicles served to EMS calls is overestimated since the provided coverage multiple times exceeds (4.27 times) the expected coverage. Nevertheless, the figure shows that redistributing (or rather gathering of all vehicles into the station in region 8) results in a better EMS scheme. Problem 2 and Problem 3 provide us with the same solution as the most convenient location for a station (region 8) is included in the predetermined set of facility location in the second experiment. In this case

	Optimal number and locations of stations	Optimal number of vehicles	ϕ
Problem 1	5	12	4.27
	(4, 8, 11, 3, 1)	(0, 12, 0, 0, 0)	
Problem 2	5	7	1.05
	(4, 8, 11, 3, 1)	(0, 7, 0, 0, 0)	
Problem 3	1	7	1.05
	(8)	(7)	

Table 5.9: Comparison of different optimization models.

we can conclude that the stations in regions 1, 3, 4 and 11 seem to be redundant for emergency medical care.

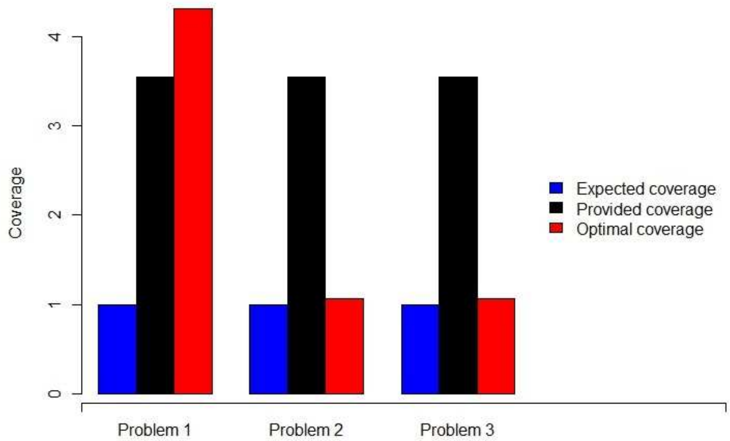

Figure 5.10: Fraction of expected coverage related to different optimization models.

5.4 Optimization throughout the day

In the following section we focus on adaptation of the optimal locations of stations and vehicles to variable conditions throughout the day which are caused by different standard deviations of the fluctuation K in travel time modeling and mainly by difference in the interarrival rates. In the real-life system the planner determines the reference time (for instance rush hours) and designs the optimal system based on requirements related to this period. Afterwards the current system is adjusted to conditions specific for other periods, usually the number of vehicles operating in a less busy period is reduced or some of the stations can be temporarily out of service. Indeed, Table B.8 shows the number of operating vehicles in each period of the day in the Agglomeration of Amsterdam. Notice that their number is highest during rush hours. Moreover, a decreasing trend for the majority of base stations can be observed for the other periods. In our calculations we chose as the reference period the rush hours, since the overall interarrival rate of accidents is the highest and therefore the highest number of ambulance vehicles is involved to respond all emergency calls. As soon as we design the optimal system for rush hours, we adjust the number of vehicles in the stations operating in off-peak hours and night hours. In words of our mathematical models, we first compute the optimal solution for rush hours, the model formulated as (P4) where the facility locations are not specified. In the second step, we declare the set of the base stations to be given by the optimal solution in the previous step, and we apply the model formulated as (P2) to off-peak hours and night hours.

Table 5.10 summarizes the optimal solution for the parameters $\alpha = 0.95$, $\delta = 15$ min., $\gamma = 20$, $\beta = 10$ and the number of scenarios equal to 5 000. In rush hours 7 vehicles settled in the station in region 8 are required to provide the first aid, during off-peak hours one vehicle is redundant and in night hours we can drop another 2 ambulances. Figure 5.11 provides the comparison of obtained optimal solutions with the current situation. In all three cases we can conclude that the coverage provided by the actual net of stations and vehicles is excessively higher than requested fraction.

	Optimal number and locations of stations	Optimal number of vehicles	ϕ
rush hours	1	7	1.06
	(8)	(7)	
off-peak hours	1	6	1.09
	(8)	(6)	
night hours	1	4	1.14
	(8)	(4)	

Table 5.10: Optimal solution for each period of the day.

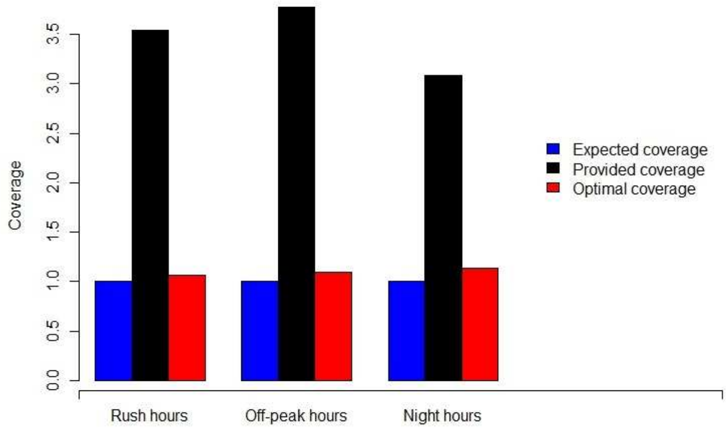

Figure 5.11: Comparing the optimal solution with the current situation for each period of the day.

Chapter 6

Conclusion and future work

In this thesis, it has been shown that the problem of designing and planning emergency medical service can be effectively posed as a stochastic programming problem. We have presented optimization models for allocating base stations and determining the number of vehicles so as to maximize system-wide expected coverage. Our models differ from previous related work mainly by considering pre-trip delays together with travel times as random variables. All models were linked to a specific risk level in order to consider the variability in the system and to reflect the decision makers' risk preference.

We carried out numerous experiments based on data collected in the Agglomeration of Amsterdam 2010. The presented numerical results illustrate how the location and allocation solutions change with respect to the different system-wide confidence level, the critical respond time and the constraints determined for facility locations. Finally, we paid special attention to the assessment of the current system design.

The future research will focus on developing similar a stochastic model which considers other feature of an EMS system, such as evaluating urban, industrial and rural areas with respect to population density by varying the risk parameters. In fact, according to the risk level specified for the given areas, the planner can choose different levels of reliability, achieving the balance between saving costs and guaranteeing a high quality of service.

Appendix A

Probability theory

Definition 1 *(Poisson process) The Poisson process can be defined in three different (but equivalent) ways:*

1. *Poisson process is a pure birth process: In an infinitesimal time interval* dt *there may occur only one arrival. This happens with the probability* λdt *independent of arrivals outside the interval.*

2. *The number of arrivals* $N(t)$ *in a finite interval of length* t *obeys the Poisson distribution with parameter* λt*,*

$$\mathbb{P}(N(t) = n) = \frac{(\lambda t)^n}{n!} exp(-\lambda t).$$

   Moreover, the number of arrivals $N(t_1, t_2)$ *and* $N(t_3, t_4)$ *in non-overlapping intervals* $(t_1 \leq t_2 \leq t_3 \leq t_4)$ *are independent.*

3. *The interarrival times are independent and obey the exponential distribution with parameter* λ*:*

$$\mathbb{P}(\textit{interarrival time} > t) = exp(-\lambda t).$$

Theorem 2 *(The convolution theorem) Let* X *and* Y *be independent random variables with distribution functions* F_X *and* F_Y*. Then the distribution function of a random variable* $Z = X + Y$ *is given as:*

$$F_Z(t) = \int_{-\infty}^{\infty} F_X(t - x)\ dF_Y(x) = \int_{-\infty}^{\infty} F_Y(t - x)\ dF_X(x), \qquad t \in \mathbb{R}. \qquad \text{(A.1)}$$

Appendix B

Attached tables and figures

The following attachment contains tables and figures that are related to our EMS system design. Their references and explanations can be found in the text therefore they are stated without any description.

fraction of service time	urgency	upper bound (min)
time handling a call	A1	10
	A2	20
time for mobilizing the vehicle crew	A1	10
	A2	20
travel time to patient	A1	30
	A2	60
time on scene	A1, A2	60
travel time to hospital	A1	30
	A2	60
time spent in hospital	A1, A2	60
travel time to base station	A1, A2	60

Table B.1: Service time restriction.

District no.	District name
1	Kudelstraat, Kalslagen and Aalsmeer
2	Oosteinde
3	Amstelveen
4	Amsterdam Centrum
5	Amsterdam Westpoort
6	Amsterdam West
7	Amsterdam Nieuw-West
8	Amsterdam Zuid
9	Amsterdam Oost
10	Amsterdam Noord
11	Amsterdam Zuidoost
12	Diemen Zuid
13	Diemen Centrum
14	Diemen Noord
15	Ouder-Amstel
16	Thamerdal
17	Zijdelwaard
18	Meerwijk
19	De Kwakel

Table B.2: List of districts in the Agglomeration of Amsterdam

	data size	range	$1/\mu$	λ
rush hours	26 853	0; 30.09	7.22	7.22
night hours	9 322	0; 24.45	3.133	3.133
off-peak hours	13 248	0; 18.45	5.936	5.936

Table B.3: List of interarrival rates for different day periods.

District no.	$\lambda_j^{\text{rush hours}}$	$\lambda_j^{\text{off-peak hours}}$	$\lambda_j^{\text{night hours}}$
1	0.1240	0.0819	0.0484
2	0.0482	0.0369	0.0150
3	0.5996	0.4236	0.2108
4	0.9434	0.9903	0.6565
5	0.2740	0.1807	0.0963
6	0.7352	0.6594	0.3364
7	1.0356	0.8007	0.3830
8	1.0072	0.7975	0.3849
9	0.7741	0.6353	0.3195
10	0.7184	0.5578	0.2681
11	0.5418	0.4139	0.2303
12	0.0575	0.0507	0.0249
13	0.0543	0.0573	0.0217
14	0.0181	0.0186	0.0154
15	0.0983	0.0680	0.0294
16	0.0384	0.0359	0.0172
17	0.0643	0.0358	0.0222
18	0.0211	0.0153	0.0093
19	0.0003	0.0004	0.0003

Table B.4: List of interarrival rates per district for different day periods.

	data size	range	$\hat{\sigma}$	μ	σ
rush hours	23 718	-1262; 2972	438.47	237.60	368.51
off-peak hours	10 417	-1189; 2589	370.33	169.78	329.12
night hours	7 471	-1141; 2166	340.95	141.31	310.28

Table B.5: List of fluctuations for different day periods.

Name	Address	District
Academisch Medisch Centrum	Meibergdreef 9 1105 AZ, Amsterdam Zuidoost	Amsterdam Zuidoost
BovenIJ Ziekenhuis	Statenjachtstraat 1 1034 CS, Amsterdam	Amsterdam Noord
Onze Lieve Vrouwe Gasthuis	Oosterpark 9 1091 AC, Amsterdam	Amsterdam Oost
Slotervaartziekenhuis	Louwesweg 6 1066 EC, Amsterdam	Amsterdam Nieuw-West
St. Lucas Andreas Ziekenhuis	Jan Tooropstraat 164 1061 AE, Amsterdam	Amsterdam West
VU Medisch Centrum	de Boelelaan 1117 1081 HV, Amsterdam	Amsterdam Zuid
Ziekenhuis Amstelland	Laan van de Helende Meesters 8, 1186 AM, Amstelveen	Amstelveen

Table B.6: List of hospitals located in the Agglomeration of Amsterdam

Station no.	Address	District
1	Karperweg 19-25 1075 LB, Amsterdam	Amsterdam Zuid
2	Meibergdreef 9 1105 AZ, Amsterdam Zuidoost	Amsterdam Zuidoost
3	Spinnerij 15 1185 ZN, Amstelveen	Amstelveen
4	Valckenierstraat 9-21 1018 XB, Amsterdam	Amsterdam Centrum
5	Zwarteweg 77A 1431 VJ, Aalsmeer	Kudelstraat, Kalslagen and Aalsmeer

Table B.7: List of stations located in the Agglomeration of Amsterdam

Station no.	District name	Number of vehicles		
		rush hours	off-peak hours	night hours
1	Amsterdam Zuid	5	4	3
2	Amsterdam Zuidoost	2	1	1
3	Amstelveen	1	1	1
4	Amsterdam Centrum	3	4	2
5	Kudelstraat, Kalslagen and Aalsmeer	1	1	1

Table B.8: Current number and location of ambulance vehicles in the Agglomeration of Amsterdam

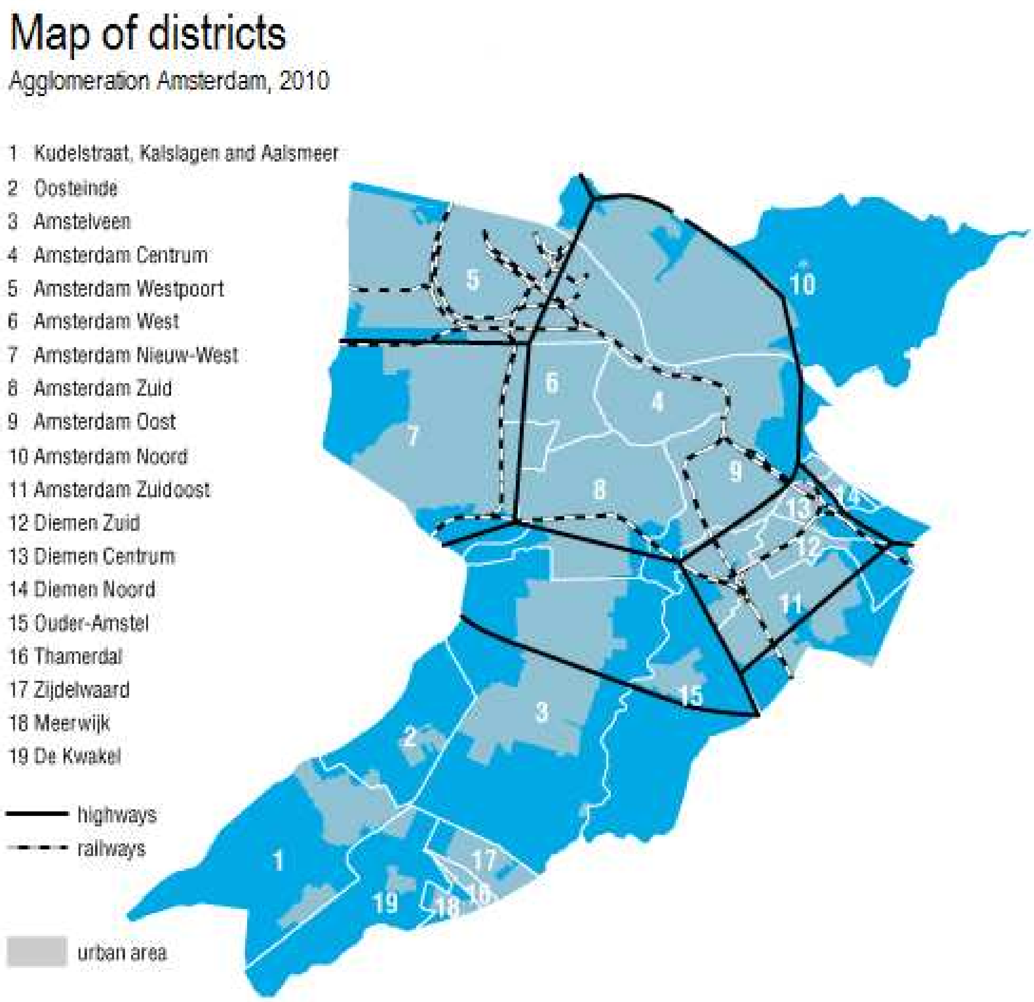

Figure B.1: Graphical illustration of districts in the Agglomeration of Amsterdam.

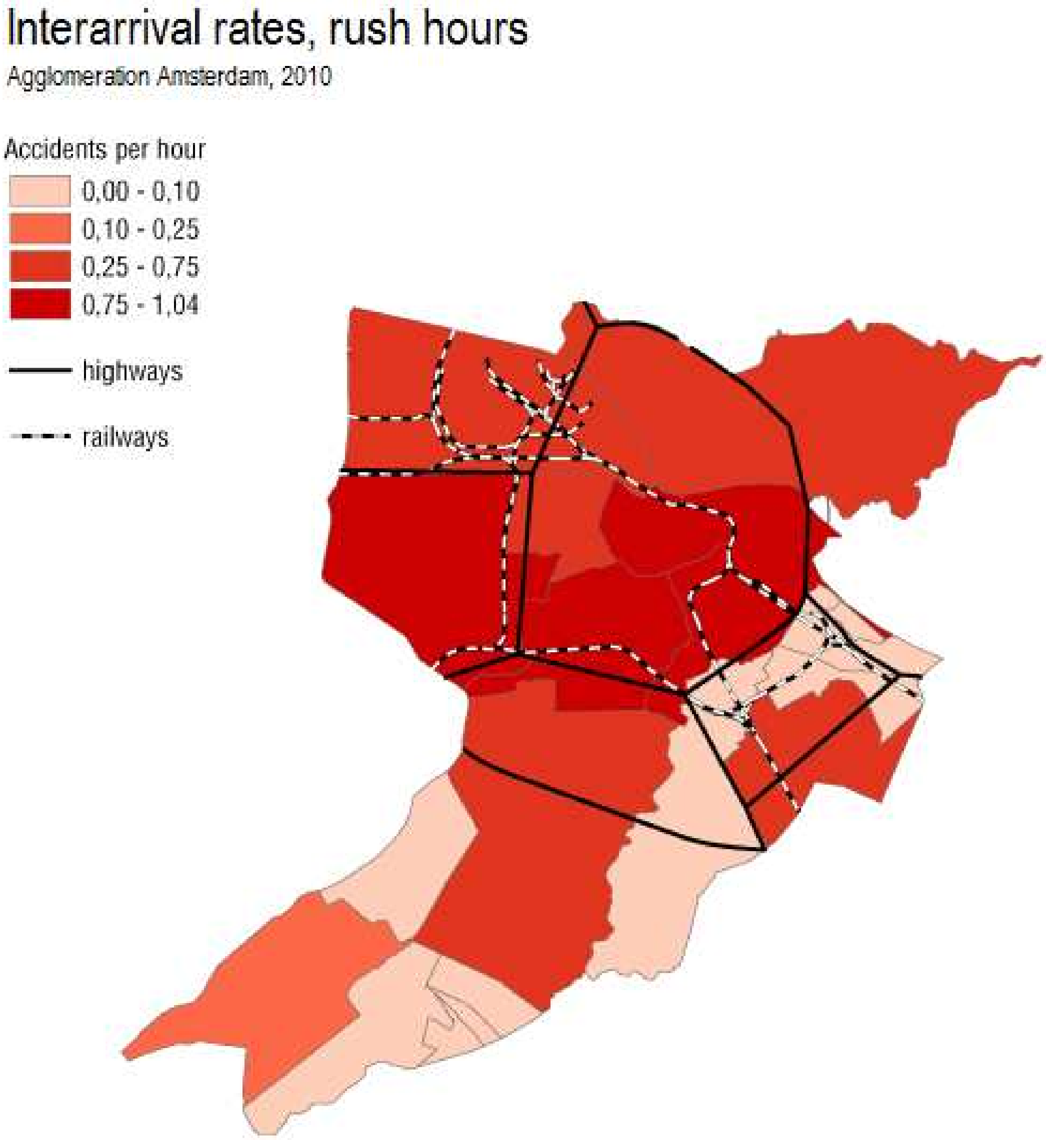

Figure B.2: Graphical illustration of interarrival rates in rush hours.

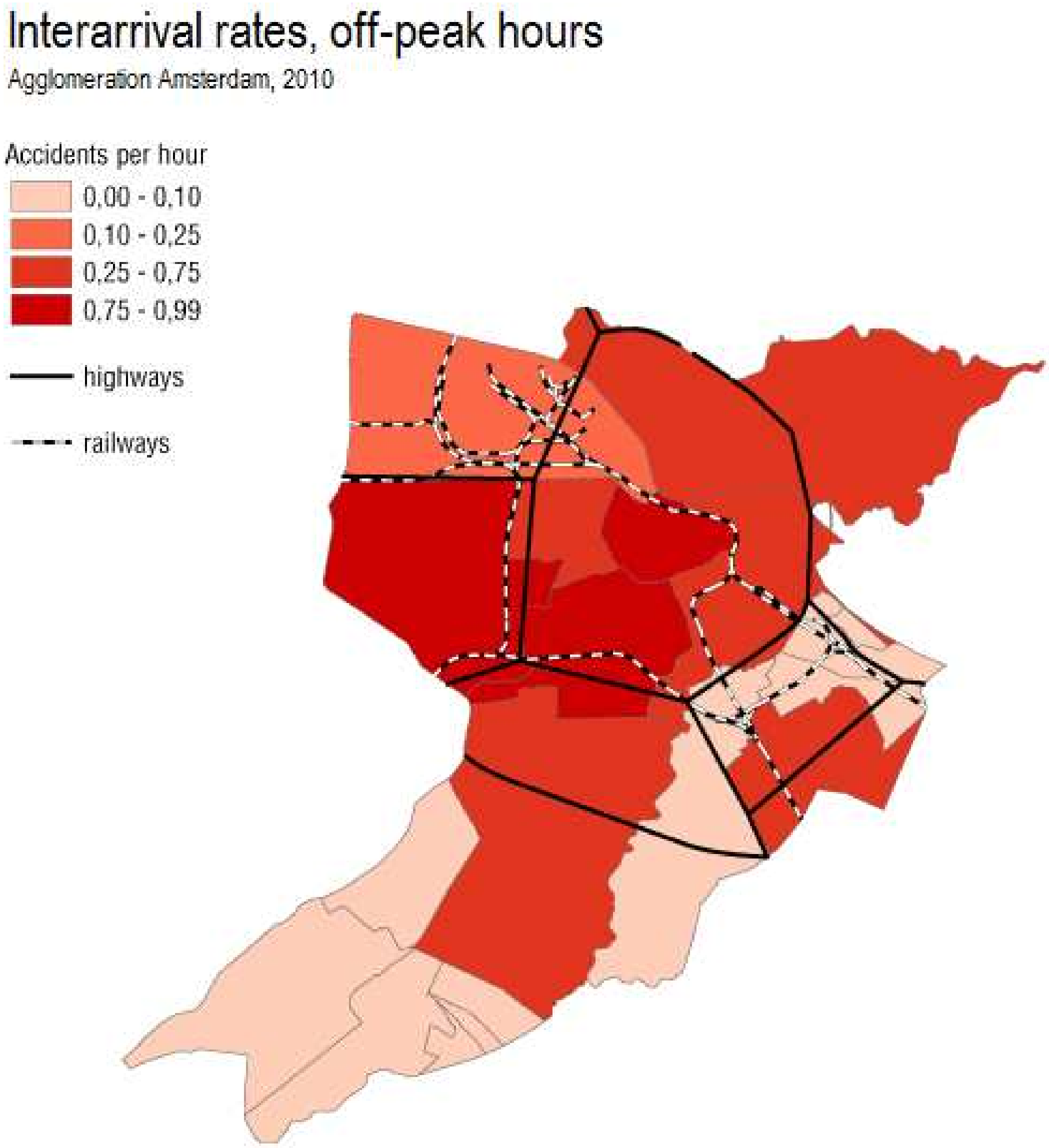

Figure B.3: Graphical illustration of interarrival rates in off-peak hours.

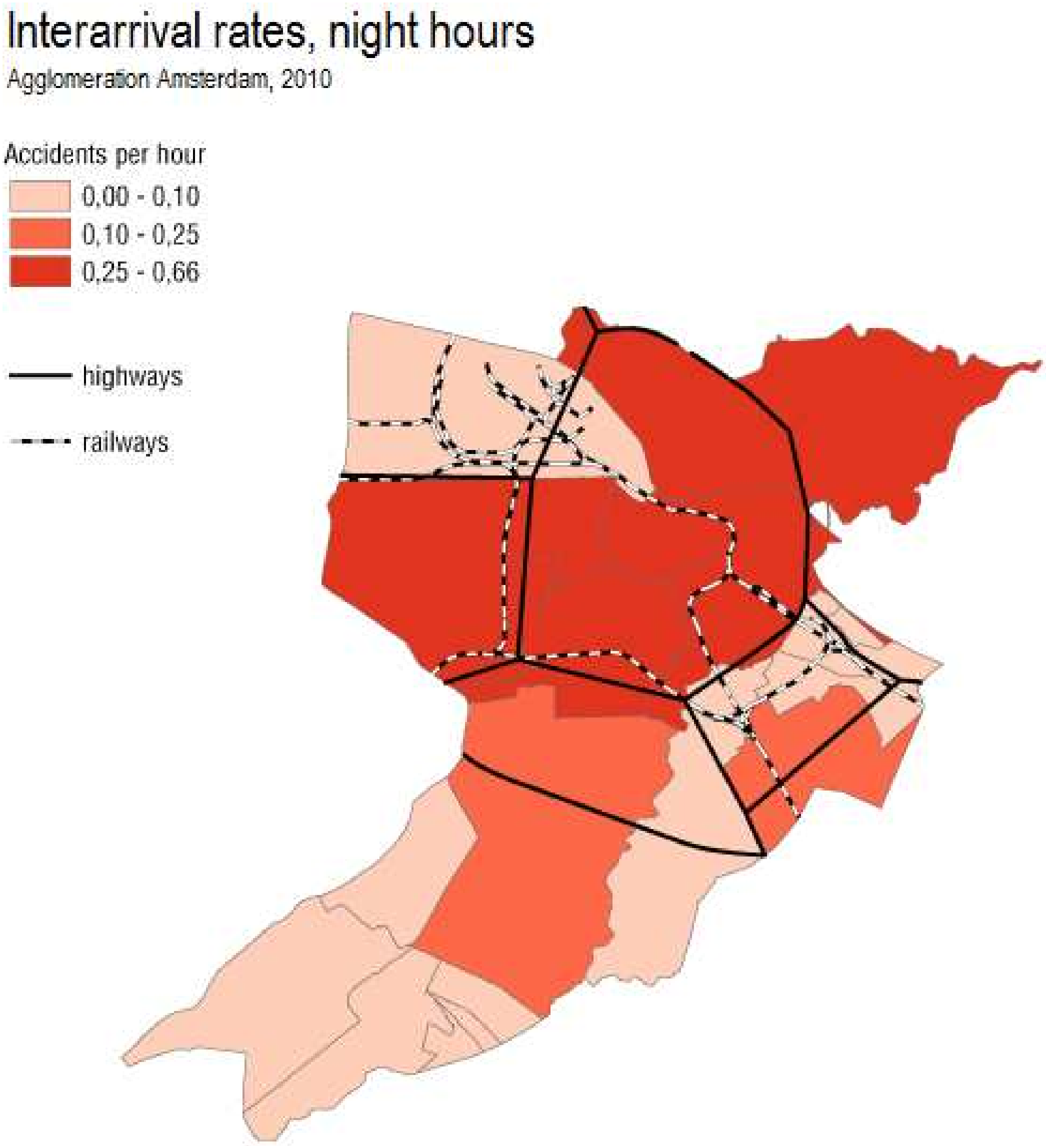

Figure B.4: Graphical illustration of interarrival rates in night hours.

Bibliography

[1] Adel, A., White, J. A. (1978). Probabilistic formulation of the emergency service location problem. *The Journal of the Operational Research Society, 29(12)*, 1167–1179.

[2] Batta, R., Dolan, J. M., Krishnamurthy, N. N. (1989). The maximal expected covering location problem: revisited. *Transportation Science, 23*, 277-287.

[3] Beraldi, P., Bruni, M. E. (2009). A probabilistic model applied to emergency service vehicle location. *European Journal of Operational Research, 196*, 323-331.

[4] Beraldi, P., Bruni, M. E., Conforti, D. (2004). Designing robust emergency medical service via stochastic programming. *European Journal of Operational Research, 158*, 183-193.

[5] Brotcorne, L., Laporte, G., Semet, F. (2003). Ambulance location and relocation models. *European Journal of Operational Research, 147*, 451-463.

[6] Chapman, S., White, J. (1974). Probabilistic formulations of emergency service facilities location problems. In Paper presented at the 1974 ORSA -TIMS Conference, San Juan, Puerto Rico.

[7] Daskin, M. S. (1983). A maximum expected covering location models: formulation, properties and heuristic solution. *Transportation Science, 17(1)*, 48-70.

[8] Daskin, M. S. (1987). Location, dispatching and routing model for emergency services with stochastic travel times. *Spatial Analysis and Location-Allocation Models*, eds. Ghosh, A. and Rushton, G. Van Nostrang Reinhold Company, New York, 224-265.

[9] Goldberg, J. B., Paz, L. (1991). Locating emergency vehicles bases when service time depends on call location. *Transportation Science, 25*, 264-280.

[10] Goldberg, J. B. (2004). Operations research models for the deployment of emergency services vehicles. *EMS Management Journal, 1(1)*, 20-39.

[11] Ingolfsson, A., Budge, S., Erkut, E. (2008). Optimal ambulance location with random delays and travel times. *Health Care Management Science, 11(3)*, 262-274.

[12] Larson, R. C. (1974). A hypercube queueing model for facility and redistricting in urban emergency services. *Computers and Operations Research, 1*, 67-95.

[13] Marianov, V., ReVelle, C. (1995). Siting emergency services. In Z. Drezner (Ed.), *Facility location: a survey of applications and methods*. Berlin: Springer.

[14] Marianov, V., ReVelle, C. (1996). The queueing maximal availability location problem: a model for the siting of emergency vehicles. *European Journal of Operational Research, 93*, 110-120.

[15] Noyan, N. (2010). Alternate risk measures for emergency medical service system design. *Annals of Operations Research, 181*, 559-589.

[16] Owen, S. H., Daskin, M. S. (1998). Strategic facility location: a review. *European Journal of Operational Research, 111(3)*, 423–447.

[17] ReVelle, C., Hogan, K. (1989). The maximum availability location problem. *Transportation Science, 23*, 192-199.

[18] Snyder, L. V. (2006). Facility location under uncertainty: a review. *IIE Transactions, 38(7)*, 537–554.

[19] Toregas, C., Swain, R., ReVelle, C., Bergman, L. (1971). The location of emergency service facilities. *Operations Research, 19*, 1363-1373.