C.W.P. Huibers

Alternative Relocation Routes in Ambulance Care

Master's thesis, November 1, 2017

Thesis advisors:
Prof. dr. S. Bhulai
M. van Buuren
Prof. dr. R.D. van der Mei
Dr. F. Spieksma

Universiteit Leiden

Centrum Wiskunde & Informatica

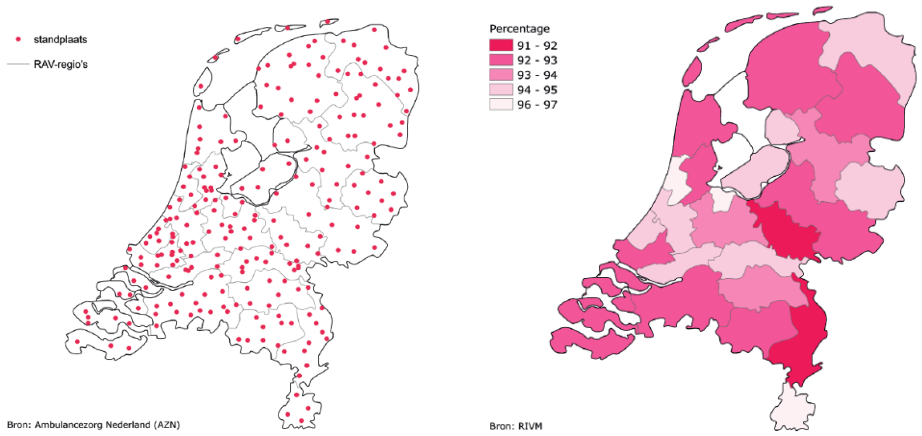
Contents

The notation used throughout this thesis, in alphabetical order.

A	Set of arcs between demand nodes.
C_r	Coverage assigned to route r.
C_{combi}	Coverage of the region using the combi coverage measure.
C_{MEXCLP}	Coverage of the region using the MEXCLP coverage measure.
C_{single}	Coverage of the region using the single coverage measure.
$C_i(S)$	Coverage of the region for configuration S except that the relocating ambulance is in node i instead of the origin.
$C(t, S, r)$	Coverage of route r at time t with initial configuration S.
c	Scaling parameter in the weighted multiple path evaluation method.
D	Destination of a relocation.
d_i	Probability that an arriving incident occurs in node i.
$f(t)$	Penalty function.
$f(\tau_{i,j})$	Contribution to the coverage of node j if the relocating ambulance is in node i.
$H(S)$	Set of all possible configurations that can be attained after one relocation from initial configuration S.
k_i	Number of other ambulances that can reach node i in time.
L	Time threshold for late arrivals.
N	Set of demand nodes.
N_S	Set of nodes that can be reached in time in configuration S.
n	Number of ambulances.
n_i	Number of idle ambulances that have node i as their destination.
O	Origin of a relocation.
$P_{S,i}$	Number of ambulances that can reach node i in time in configuration S.
q	Busy fraction.
R	Routeset containing routes to be evaluated.
S	Configuration, i.e. the set of locations of idle and relocating ambulances.
$S^{(O,D)}$	Configuration after relocating an ambulance from O to D, i.e. $(S \cup \{D\}) \setminus \{O\}$.
T_R	Longest travel time of the routes in routeset R.
$T_{S,i}$	Non-decreasing travel times for all ambulances in configuration S to node i.
$t_{i,j}$	Driving time from i to j when relocating.
t_i	Travel time from the origin to node i, i.e. $t_{O,i}$.
t_i^r	Time route r is in node i.
V	Set of Route Points.
V_r	Nodes along route r.
W	Set of base locations and waiting sites.
$x(t)$	Position of relocating ambulance at time t.
α	Scaling parameter for the convex combination method.
β	Scaling parameter in penalty functions.
γ	Scaling parameter in route evaluation in TIFAR.
δ	Maximum overlap of nodes for routes in a routeset.
ζ	Scaling parameter in penalty functions.
λ	Poisson arrival rate of incidents.
$\tau_{i,j}$	Driving time from i to j when responding to an incident.

1 Introduction

Every second counts in the world of ambulance care. Providers of emergency care need to respond quickly to requests. However, the resources and budget are limited. That's why emergency service providers need to think carefully about how they spend them. Mathematical models have been developed to obtain better efficiency in ambulance care. In 2015, 93.4% of the most urgent calls in the Netherlands were reached within 15 minutes [1]. Although this percentage is increasing, it is required by law in the Netherlands that 95% of the most urgent calls be answered within 15 minutes. There are 24 EMS regions in the Netherlands. Figure 1*a* shows the base locations in the Netherlands for each Emergency Medical Service (EMS) region, while Figure 1*b* shows the percentage of $A1$-calls that have been answered within 15 minutes in 2015. The research project REPRO (from REactive to PROactive planning of ambulance services), aimed at developing mathematical models in cooperation with ambulance service providers from the Netherlands, focuses on this issue. The project strives to make new and better ambulance planning methods.



(a) (b)

Figure 1: *Map of the EMS regions in the Netherlands. The pink dots in* (*a*) *indicate a base location for ambulances. The color in* (*b*) *indicates the percentage of* $A1$*-calls that were reached within 15 minutes in 2015 per region.*

In this thesis we take the developed models for ambulance relocation and extend them by looking at different routes a relocating ambulance can take. An ambulance responding to an incident will still take the fastest route. We test our methods on various graphs, either square grids or graphs representing a real EMS region in the Netherlands. When an ambulance is relocating, it will always take the fastest path in the current models, but it might improve coverage when the ambulance takes a different route. This brings us to the main question of our thesis:

How can we evaluate routes depending on coverage and travel time and how does relocating over different routes influence the performance of the model?

The idea to look at routes of relocating ambulances came from the EMS region Amsterdam. There it was observed that some relocating ambulances did not reach

their base location, because they had to respond to an incident before reaching its destination. This implies that the route a relocating ambulance takes to the base location influences whether it can respond to an incident in time.
We use two different simulation tools in this thesis. The first model is programmed in Matlab and will be used in the first half of this thesis. Here we look at different relocation policies and how changes in the test region affect the performance of the policies. Then we switch to the Testing Interface For Ambulance Research (TIFAR) when we analyse real EMS regions. Our contribution consists of developing relocation policies for the models and implementing them in Matlab and TIFAR. Furthermore, we tested our methods in various simulations to see how they perform.

In Section 2 we formulate the problem more extensively and give some background information about how the Emergency Medical Services are organized in the Netherlands. In Section 3 we introduce the basis model we use for our first set of simulations. Furthermore, we discuss several measures for the coverage of a region. In Section 4 we propose a way to extend the base model with our own relocation policies. In Section 5 we test our relocation policies on different grids using simulations and discuss the results for these grids. In Section 6 we test the policies for Flevoland using generated data based on historical data. We also run simulations on historical data from 2011.

In the remaining sections of our thesis we describe the second simulation tool in more detail. Section 7 is an introduction to the TIFAR. In Section 8 we present a way to implement dynamic routing in our simulations. Here we describe how we choose the routes we want to consider for each relocation, and how we evaluate them. The results of our simulations are shown in Section 9. We run our simulations for the EMS regions Gooi & Vechtstreek, Amsterdam and Utrecht. Finally, in Section 10 we discuss our conclusions and give some suggestions for further research.

2 Problem Description and Background

When an incident occurs, an ambulance needs to respond to it as soon as possible. To achieve this, it is important that available ambulances position themselves at strategic places throughout the region. Extensive research has already been done to develop new models for ambulance positioning and relocation. The article by Bélanger, Ruiz and Soriano [2] is a comprehensive survey of the different models that have been proposed to organise ambulance care. It mentions models to determine where the base locations should be in the region [4], [5], as well as ambulance relocation models [9].

To achieve a better coverage of the region, it can be necessary to move an ambulance to another base location. This is called a *relocation*. In the currently developed models, the route an ambulance takes when it relocates, is the shortest route with respect to the travel time. However, a relocating ambulance does not always arrive at its destination, because it has to go to an incident before it arrives. Thus the route an ambulance takes is important, since it also gives coverage to the region while driving. Driving along one of the alternative routes might decrease the number of late arrivals. Thus we want to take multiple routes of similar length into consideration. Note that choosing a certain route depends on the current state of the system, which includes the position of the other ambulances.

An example of multiple routes of similar lengths is in Flevoland, which uses the model presented in [3], where it was observed that if an ambulance relocates from Dronten to Lelystad, it always uses the highway. However, one can also drive through Biddinghuizen in the south, or use one of the secondary roads to reach Lelystad. These alternative routes have roughly the same length, but the ambulance covers different parts of Flevoland. Figure 2 shows possible routes the ambulance can take.

Figure 2: *Multiple routes from Dronten to Lelystad, indicated with black lines.*

This leads us to the main question of our thesis: How can we evaluate routes depending on coverage and travel time and how does relocating over different routes influence the performance of the model? What factors of the region influence the performance of dynamic routing? When we develop a method for dynamic routing, we need to consider how to determine the possible routes the ambulance can take as well and how we can evaluate these routes. When we test our methods, we are particularly interested in the effect on the number of late arrivals and the mean response time, with the goal of getting more insight in dynamic routing. To do this, we first need some background information on ambulance care.

2.1 Ambulance Care

We explain the basics of ambulance care on the basis of the Netherlands. In the Netherlands there are 24 Emergency Medical Service (EMS) regions, which in Dutch are called Regionale AmbulanceVoorzieningen (RAV). Each of these regions has its own ambulance service provider, who organises the emergency care. Each service provider manages the ambulance units in the region, and determines which ambulance responds to an emergency call.

There are three different call urgencies in the Netherlands:

- $A1$-calls: in this type of call there is an immediate threat to the health of the patient, or the threat can only be fixed after on-site care of ambulance staff. These calls have the highest priority, and the ambulance should be on-site within 15 minutes.
- $A2$-calls: in this type of call there is serious health damage to the patient, but there is no acute danger to life. The ambulance staff aim to be on-site within 30 minutes, but this is not enforced by law.
- B-calls: this type of call is for transport of a patient between hospitals and home addresses. The destination is known in advance. This type of call is also called planned transport.

The dispatcher determines urgency of the calls. In this thesis we assume that every call is of type $A1$, unless stated otherwise. In the Netherlands it is required by law that 95% of the $A1$-calls are answered in 15 minutes. The response time consists of three parts. These three parts are the dispatch time, chute time and travel time. If an ambulance is not on the scene within 15 minutes, then we call it a late arrival.

3 Current Model

In this section, we describe the model introduced in [3], which we used as a basis for our model. In Section 3.1 we describe the basis model and in Section 3.2 we discuss performance measures for this model and extended versions of the model. In Section 3.3 we explain how the model implements dynamic routing by explaining how we choose a relocation.

3.1 Basis Model

In this section we describe the basis model we used in this thesis, which is based on the model used in [3]. The region is modeled as a weighted directed graph, where N is the set of demand nodes, each representing a neighborhood or postal code area of the region. Some of these nodes are also base locations or waiting sites for ambulances. The set of base locations and waiting sites is $W \subset N$. Incidents arrive according to a Poisson process with arrival rate λ. The probability that this incoming incident occurs at node i is the demand of that node, which we denote by d_i. Thus the arrival rate of an incident at node i is λd_i. Note that $\sum_{i=1} d_i = 1$. The connections between nodes are modeled by arcs $(i, j) \in A$, where $i, j \in N$. The driving time between two nodes when the ambulance is relocating is $t_{i,j}$. When an ambulance responds to an incident, it drives faster while using optical signals and sirens. This changes the travel time between nodes. Let $\tau_{i,j}$ be the travel time between nodes i and j of an ambulance responding to an incident.

Now we look at the ambulances in the region. We denote the number of ambulances in the region by n. These ambulances can be idle, relocating or busy. If an ambulance is idle, it is at a waiting site and can respond to an incident. A relocating ambulance can also respond to an incident, but it is moving between two nodes. When an incident occurs, we use the node that a relocating ambulance is driving from as its location. The set of current locations of idle and relocating ambulances, which is tracked by a list of demand nodes, is called the current configuration of the region, which we denote by the set S. It is possible that there are multiple ambulances at one location, hence a node can occur more than once in S. Note that the number of available ambulances is $|S|$. In Figure 3 we show an example of a configuration for a small graph.

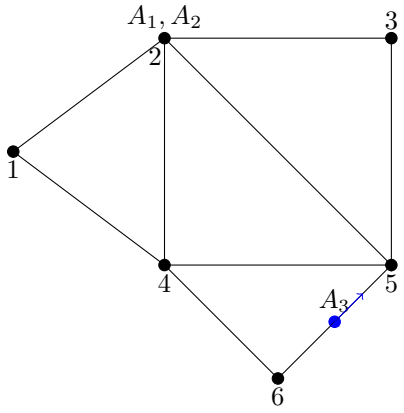


Figure 3: *A graph with idle ambulances in node* 2 *(A_1, A_2) and a relocating ambulance going from node 6 to node 5 (A_3). The configuration is thus* $S = \{2, 2, 6\}$.

Finally, if an ambulance is busy, then it is responding to an incident and therefore it cannot respond to another incident. The time that an ambulance is busy on average, is called the busy fraction, which we denote by q. An ambulance is late, if it arrives at an incident after time threshold L. This threshold is 15 minutes for $A1$-calls in the Netherlands. In reality there are different types of ambulances, but in the first part of this thesis we only consider one type of ambulance.

Table 1: Notation.

N	Set of demand nodes.
W	Set of base locations and waiting sites.
A	Set of arcs.
d_i	Probability that an arriving incident occurs in node i.
λ	Poisson arrival rate of incidents.
$t_{i,j}$	Driving time from i to j when relocating.
$\tau_{i,j}$	Driving time from i to j when responding to an incident.
n	Number of ambulances.
S	Configuration, i.e. the set of locations of idle and relocating ambulances.
q	Busy fraction.
L	Time threshold for late arrivals.

3.2 Performance Measures

There are multiple ways to measure the coverage of a configuration of ambulances. We look at three measures. The first one is the single coverage measure. Using this measure, the objective is to maximise the amount of locations that are covered by at least one ambulance. Given a configuration S, let N_S be the collection of nodes that can be reached by an available ambulance within L seconds. Then the single coverage measure of the configuration is:

$$C_{Single}(S) = \sum_{i \in N_S} d_i.$$

This coverage measure was first used in the Maximal Covering Location Problem introduced by Church and Revelle [4].

The second coverage measure we use is the MEXCLP coverage, which has been introduced by Daskin in the Maximum Expected Covering Location Problem [5]. In this problem we also take the probability that an ambulance is busy into account, and we consider the multiple ambulances that can cover a node. The formula for this measure is:

$$C_{MEXCLP}(S) = \sum_{i \in N} d_i \sum_{j=1}^{P_{S,i}} (1-q)q^{j-1}.$$

Here $P_{S,i}$ is the number of ambulances that can reach node i in configuration S on time. The term $(1-q)q^{j-1}$ is the probability that the j-th closest ambulance is available, while the $(j-1)$ closer ambulances are occupied. The MEXCLP coverage of a node then takes the summation of all ambulances that can reach the node within the target time, and the MEXCLP coverage of the configuration is the sum of the coverages of all nodes.

Finally we look at a third measure, which is based on the MEXCLP coverage. Note that for each demand node $i \in N$, we know that the travel time from each other demand node $j \in S$ to i is $t_{j,i}$. We can sort these travel times non-decreasing in the vector $T_{S,i} = (t_{s_1,i}; \ldots; t_{s_{|S|},i})$ such that $s_1, \ldots, s_{|S|} \in S$. We also use a penalty function f, which is non-decreasing as a function of the response time t. We now have the following measure:

$$C_{combi}(S) = \sum_{i \in N} d_i \Big(\sum_{j=1}^{|S|} (1-q)q^{j-1} f(t_{s_j,i}) \Big), \tag{1}$$

with $s_j \in T_{S,i}$. We want to minimise this value, in contrast to the single and MEXCLP coverage, since the penalty function is non-decreasing. Note that the contribution of an ambulance decreases the farther away it is from node i. This is because of the term q^{j-1} from the MEXCLP coverage in the equation.

There is a variety of functions that one can use for the penalty function f. The first one we discuss, is the average response time to a request

$$f(t) = t, t \geq 0.$$

Here, each extra time unit that one is late, gives the same additional penalty, see Figure 4*a*. Another penalty function is the maximum allowed response time

$$f(t) = \begin{cases} 0 & \text{if } t \leq L, \\ 1 & \text{if } t > L, \end{cases}$$

where L is the threshold for late arrivals. Thus, this function induces a penalty, when arriving after L time units. This function is displayed in Figure 4*c*. A drawback of this method is that it does not differentiate between a short response time and a response time barely below L. The following penalty function overcomes that problem:

$$f(t) = \left(1 + e^{-\beta(t-L)}\right)^{-1}, t \geq 0,$$

which gives us a smoother version of the previous function. Here β is a scaling parameter. Figure 4*b* shows this penalty function. Lastly, we consider a penalty function that is a combination of the two previous functions and is displayed in Figure 4*d*. The function is given by

$$f(t) = \begin{cases} \left(\zeta \cdot (1 + e^{-\beta(t-L)})\right)^{-1} & \text{if } t \leq L, \\ \frac{\zeta - 1}{\zeta} + \left(\zeta \cdot (1 + e^{-\beta(t-L)})\right)^{-1} & \text{if } \ t > L. \end{cases}$$

This function focuses on minimising the number of late arrivals, since the penalty increases drastically when the response time is larger than L. Here β and ζ are scaling parameters.

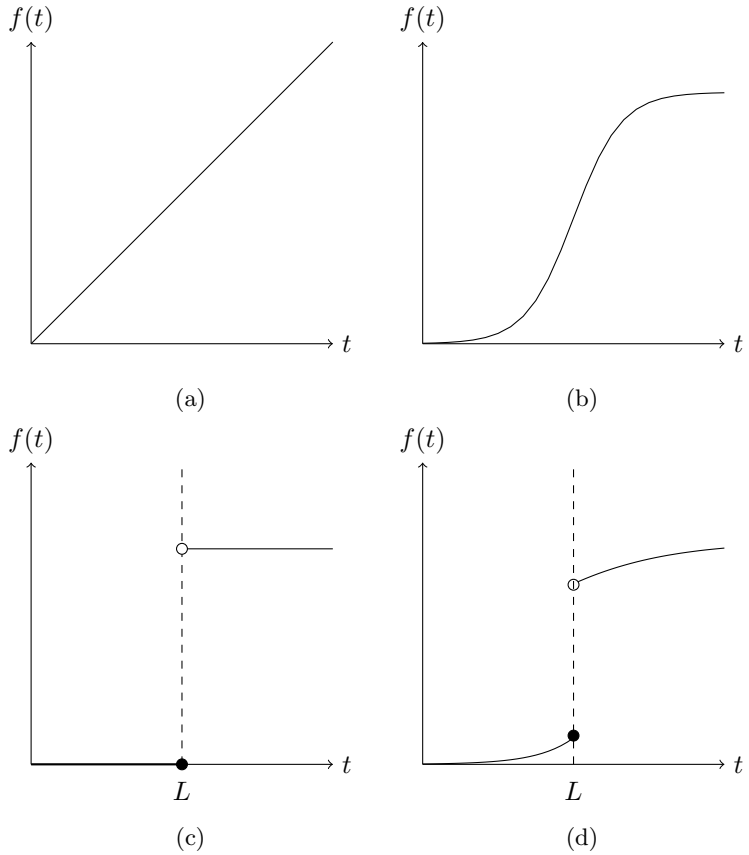


Figure 4: *Examples of penalty functions.*

3.3 Relocation Policy

In this section we describe how the base model chooses a relocation. There are four possible situations for a relocation to occur. We call these situations *decision moments*, which are moments that:

1. an incident occurs and an ambulance has been sent to that incident. There is now one less ambulance in the new configuration.
2. an ambulance has finished its service at the hospital or the incident scene and has become idle. Thus there is an extra ambulance available.
3. an ambulance starts its shift at a base location. There is now one more ambulance available.
4. an ambulance ends its shift. Fewer ambulances are available now.

We call these decision moments of types 1, 2, 3 and 4 respectively. At these decision moments, the dispatcher has to decide whether a relocation is necessary. It is possible that a relocation is not necessary at a decision moment of type 1, 3 and 4. The configuration may still be satisfactory in terms of expected response time to future incidents. At decision moments of type 2 there will always be a relocation, since the new idle ambulance has to be relocated. If a relocation is necessary to improve the configuration, then the dispatcher has to select a pair of base locations (O, D), meaning that an ambulance is sent from origin $O \in S$ to destination $D \in W$. Choosing a relocation pair depends on the current configuration. Each possible pair (O, D) gives a new configuration $S^{(O,D)} = (S \cup \{D\}) \setminus \{O\}$. We want to choose a relocation such that

$$S = \underset{S^{(O,D)} \in H(S)}{\operatorname{argmin}} \sum_{i \in N} d_i \Big(\sum_{j=1}^{|S|} (1-q) q^{j-1} f(t) \Big). \tag{2}$$

Here $H(S)$ is the set of all possible new configurations $S^{(O,D)}$ that can be attained after one relocation from the initial configuration. Thus we choose a relocation (O, D) that minimises Equation (2) and thus gives us the largest improvement in coverage using C_{combi} in Equation (1). We could also use the single coverage or the MEXCLP coverage measures, but we get a better performance if we use the Combi measure. For more information, see [3].

If a relocation pair (O, D) has been chosen, the dispatcher has to determine how the new configuration will be achieved. Originally, the dispatcher selects an ambulance at origin O that takes the shortest path to destination D. We are interested in the effect on the performance of routes other than the shortest path. Thus the dispatcher can also choose other routes. In Section 4 we discuss how one can select and evaluate alternative routes.

4 Measures for Routes

In this section we describe two ways to evaluate a possible route for a relocating ambulance. We assume that the current configuration is S and that a relocation starts in origin O and has destination D. In Section 4.1 we discuss the Multiple Path Evaluation (MPE) method, where we generate multiple paths and evaluate the coverage of those paths. In Section 4.2 we explain our second method, which uses a shortest path algorithm with a convex combination of the distance and coverage on each arc.

We compare our methods to the shortest path method. The shortest path method chooses for each relocation (O, D) the shortest path as a route for the relocating ambulance.

4.1 Multiple Path Evaluation

This method evaluates multiple routes for each relocation. Ideally, the routes do not have much overlap with each other. There are multiple ways that one can choose routes. In our model we choose the routes as follows: let $V \subset N$ be a collection of predetermined nodes. These nodes have to be chosen manually. We call V the *route points*. For each node $v \in V$, we take the shortest route from O to D through v. We also look at the shortest route from O to D, which does not necessarily pass any nodes in V. This gives us routeset R. Note that $|R| \leq |V| + 1$, since it contains the shortest route and one route for each route point, assuming that the routes are unique. Figure 5 shows a routeset created using two route points.

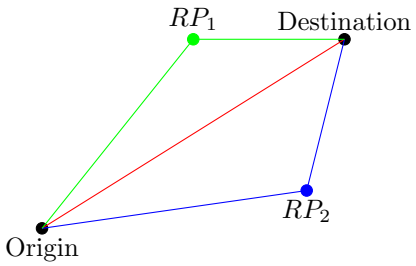


Figure 5: *Routeset consisting of three routes. The shortest route is red, and the other routes go through route points RP_1 and RP_2.*

Each route visits a set of nodes. Let V_r be the set of nodes on route $r, r \in R$. In order to evaluate a route, we give a value for the coverage for each node of the route. We determine the coverage of the configuration, when the relocating ambulance is in node $i \in V_r$ instead of in the origin O. Denote this coverage by $C_i(S)$. We assume that the position of the other ambulances does not change. Observe that $C_O(S)$ is the coverage of the initial configuration and $C_D(S)$ is the coverage after the relocation. For each node on the route we know the travel time to the next node, hence we know how long the relocating ambulance is in each area. The time interval that the relocating ambulance is in node i of route r is denoted by t_i^r. These time intervals are right open and left closed. Thus, for each route r we have a corresponding set of time intervals indicating the area where the relocating

ambulance is at time t. See Figure 6 for an example of a route r starting in O, and traveling through nodes i, j and k and arriving at destination D.

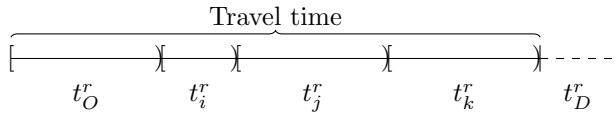


Figure 6: *Timeline indicating at which node a relocating ambulance is.*

We define $C(t, S, r)$ as follows:

$$C(t, S, r) = \begin{cases} C_i(S) & \text{if } t \in t_i^r. \\ 0 & \text{otherwise.} \end{cases}$$

$C(t, S, r)$ is the coverage of route r at time t with initial configuration S.
Now we can evaluate the route r by virtue of:

$$\int_{t=0}^{T_R} C(t, S, r)dt. \tag{3}$$

Here, T_R is the longest travel time of any of the generated routes. We use the single coverage or the MEXCLP coverage for $C_i(S)$. In this thesis we use the MEXCLP coverage, unless explicitly stated otherwise. In Figure 7 we map the coverage of two possible routes. The red line is the shortest path, while the blue line is an alternative path.

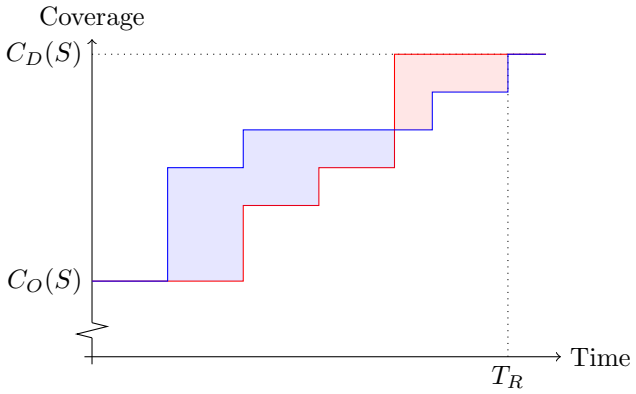


Figure 7: *The coverage of the region during two routes.*

Figure 7 shows that the blue route has a better coverage early on, but it also arrives later at the destination. Hence the red route has a better coverage at the end while the blue route is still relocating. The blue area represents when the blue route is better with respect to the coverage, and similarly, the red area represents when the red route is better. The size of the area shows how much better a route is. We can see which route has a higher value for the coverage by subtracting the blue area from the red area.

It is possible that an ambulance has to respond to an incident before it has completed its relocation. To account for this, we want that the later nodes on the routes contribute relatively little to the value of the route in comparison to the earlier nodes, since we are less likely to reach those nodes. We can modify Equation (3) to achieve this. We add the decreasing function e^{-ct}, were c is a constant, which gives us

$$\int_{t=0}^{T_R} C(t, S, r)e^{-ct}dt.$$

We call this method the Weighted Multiple Path Evaluation (WMPE).

4.2 Convex Combination

The previous method chooses multiple routes for each relocation and assigns a value to each of these. Our second method determines one route, by utilising Dijkstra's Algorithm [6]. We model the road network as a directed graph, where the value of arc (i, j) depends on the distance or driving time between i and j and the coverage of the configuration, if the relocating ambulance is at node j instead of the origin O. When we determine the coverage of a node, we assume again that the positions of the other ambulances does not change. Thus the length of the arc (i, j) is

$$\alpha t_{i,j} + t_{i,j} \cdot (1 - \alpha)(1 - C_j(S)). \tag{4}$$

Here $\alpha \in [0, 1]$ denotes the weight of the distance compared to the coverage. We have that $\alpha = 1$ gives us the graph with the normal travel times. We use the term $(1 - C_j(S))$ as opposed to $C_j(S)$ because Dijkstra's algorithm minimalises the shortest path. We want to travel through the nodes that give us a higher coverage, thus we want the arcs pointing to these nodes to have a relatively lower value, so that Dijkstra's algorithm is more likely to choose these arcs. We multiply the term $(1 - \alpha)(1 - C_j(S))$ with $t_{i,j}$ to normalise the two terms in the sum, because the coverage is smaller than one, while the driving time is generally larger than one second. Note that the driving time between nodes is known. Also observe that $(1 - C_j(S)) \geq 0$ since $C_j(S) \leq 1$. Furthermore, we use the MEXCLP measure for $C_j(S)$, since this gave us the best results in the simulations.

After all the arcs have been given a value, we apply Dijkstra's algorithm which gives us the shortest path from O to D, which now takes the coverage into consideration. We let the relocating ambulance drive along the resulting route.

5 Grid Simulations

In this section we discuss the experimental set-up. We test our methods on various 10×10 grids and on the EMS region of Flevoland. We discuss the grids in the corresponding subsections. We aim to test our methods with respect to the sensitivity of the arrival rate λ, the distribution of the demand and the size of the grid. To accomplish this, we have done a number of simulations testing these factors. The simulation in Section 5.1 is used as a basis to which we compare the other simulations. In Section 5.2 we look at the effects of an increased arrival rate. In Sections 5.3 and 5.4 we use a grid with a different demand distribution. In Section 5.5 we look at a larger grid. In Section 5.6 we look at the effect of adding a hole in the middle of the grid. Finally, we look at the effect of our methods on grids with random lengths in Section 5.7. We also look at the EMS region of Flevoland in Section 6.

For every simulation we generate incidents for a period of 200 days. The incidents arrive according to a Poisson process of rate λ. We create a table with incidents containing the information when an incident occurs, where it happens and if the ambulance has to go to the hospital afterwards. This table is called an *instance*. For each grid we generate one instance, and we test all our methods on this instance. Thus the time and place an incident occurs is the same in each simulation, only the way we relocate ambulances is different.

For each of the simulations we have to calculate the busy fraction, since it is different for each grid. To do so, we first simulate the *static policy* on the grid. In the static policy, an ambulance returns to its home location, the base location at which it started, at decision moments of type 2. There are no other relocations in the static policy. After we simulate the static policy, we calculate the busy fraction q by dividing the average time an ambulance was unavailable by the total model time. The busy fraction of Flevoland is known.

Furthermore, we assume that the number of ambulances in the grid stays constant. Thus none of the ambulances ever go off shift in our simulations. This changes for the real EMS regions in Sections 6 till 9, where we use the actual ambulance schedules.

We use the methods in Sections 4.1 and 4.2. For the multiple path evaluation method and the convex combination method we use the MEXCLP coverage, since this gave us the best results in our simulations. In the weighted multiple path evaluation method, we use $c = 0.0003$.

We consider the following performance measures for our methods.

- The percentage of late arrivals.
- The mean response time.
- The mean single coverage.
- The mean MEXCLP coverage.

In these simulations, the response time consists only of the travel time. The dispatch and chute time are not taken into account. To compensate for this, we deduct the expected time of the dispatch and chute time from the time threshold L. This expected time is estimated to be 3 minutes. Thus for the simulation of Flevoland we have that a time threshold of $L = 12$ minutes instead of 15 minutes. For the 10×10 grids we have a time threshold of $L = 8$ minutes, since the nodes in the grid are closer to each other than in Flevoland. The single and MEXCLP coverage of the region are determined at two instances in the simulation. The first instance is when an incident occurs and the second instance is when a busy ambulance becomes idle. We use the coverage of the region at these instances to determine the mean coverage of the region. See Section 3.2 for how we calculate these coverages. Furthermore, we calculate the 95% Student's t-confidence intervals (CI) for the percentage of late arrivals.

We discuss the results of our simulations in the following sections.

5.1 Basis Grid

In this section we first discuss the graph that we use as the basis of our simulations. It is the 10×10 grid shown in Figure 8. The nodes on this grid are connected to their horizontal and vertical neighbors. We set the travel distance between two nodes to be 100 seconds when relocating and 90 seconds when responding to an incident. There are also base locations, hospitals and route points. These are indicated with red, blue and green nodes respectively. The base locations are chosen randomly, except that they are not too close to each other. At the start of the simulation there are two idle ambulances at each waiting site, thus there are ten ambulances in total in the region. The hospitals are placed in the corners of the map and not in the same place as the base locations. This is to ensure that there is always a relocation when an ambulance becomes available at a hospital. The route points are chosen in the center of the grid, because nodes in the center cover a larger area than the nodes on the border.

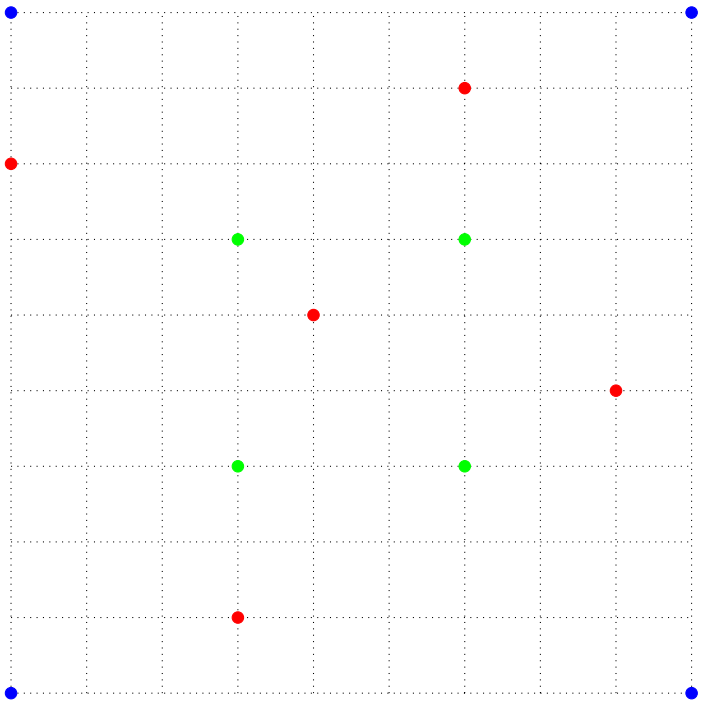


Figure 8: *Our basis grid for the simulations. Blue nodes represent hospitals, red nodes are base locations, and green nodes are route points for the MPE method. Every node is connected to its horizontal and vertical neighbor, but not to its diagonal neighbor.*

We equip the grid in Figure 8 with a uniform demand, i.e. $d_i = \frac{1}{100}$ for each node i. We also have an arrival rate of $\lambda = 0.0015$. The busy fraction of the grid is $q = 0.288$. We use the same values for the variables in the other simulations, unless stated otherwise. Table 2 shows the result of our simulation. We tested the Shortest Path method (SP), the MPE method, the WMPE method and the Convex Combination method (CC). We assume that an ambulance is late to an incident if it arrives after 8 minutes, i.e. $L = 8$ minutes.

Table 2: Simulation results.

	SP	MPE	WMPE	CC
Late arrivals	3.65%	3.50%	3.49%	3.24%
95% CI Lower bound	3.42%	3.27%	3.23%	3.03%
95% CI Upper bound	3.87%	3.72%	3.71%	3.46%
Mean response time	226.13 s	225.34 s	225.30 s	224.12 s
Mean single coverage	98.33%	98.34%	98.34%	98.35%
Mean MEXCLP coverage	91.10%	91.11%	91.11%	91.12%

We see that all our methods perform better than the shortest path method with respect to every performance measure. The MPE and WMPE method are 0.15 and 0.16 percentage points better respectively than the shortest path method with respect to late arrivals. The convex combination method performs the best with a 0.41 percentage point decrease in late arrivals compared to the shortest path method. Furthermore, observe that the 95% confidence interval of the convex combination method has only a small overlap with the 95% confidence interval of the shortest path method.

Thus, there is some merit in using different relocation routes for ambulances, since our methods show improvement over the shortest path method. An explanation for this is that there are usually multiple shortest paths between two nodes in the grid, since the length of each arc is the same. The shortest path method picks one of these shortest path at random, which might not necessarily give the best coverage while driving. The MPE method has more routes to choose from, and makes a decision based on Equation (3). Thus having only a small set of routes to choose from can already give an improvement. On the other hand, the convex combination method considers all the arcs in the graph and determines which path over these arcs gives the highest contribution to the coverage.

We did not need to test the convex combination method for other values than $\alpha = \frac{1}{2}$, since the results are the same for all values of $\alpha \in [0, 1)$. The reason for this is that the distance between two neighboring nodes is the same. Thus the contribution of $\alpha t_{i,j}$ in Equation (4) is the same for each arc. Routes that travel over the same number of arcs have the same contribution of $\alpha t_{i,j}$ to their value. Thus the difference in their value is determined solely by the $(1-\alpha)(1-C_j(S))$ term on the arcs. Since $(1-\alpha)$ is now a constant for a given α, the value of the route depends on the term $(1-C_j(S))$. Furthermore, a route over more nodes generally has a larger value for Equation (4), since it sums over more arcs, so the convex combination method rarely considers routes longer than the shortest path. Thus, which route is chosen is unaffected by α. Hence, we did not vary α for most of the grids. However, we did change it for the simulation of the grid in Section 5.7 and for Flevoland in Section 6, since the distance between two neighboring nodes is not constant there.

5.2 Increased Arrival Rate

In this section we look at the influence of an increased arrival rate, since we are interested if our methods still work better when ambulances are busier. To do this, we changed the arrival rate to $\lambda = 0.003$. This results in twice as many incidents on average. The busy fraction also increases because of this change. We have $q = 0.803$. Since there are more incidents in this simulation, we set $L = 12$ minutes. The results of the simulation are in Table 3.

Table 3: Simulation results for $L = 12$ minutes.

	SP	MPE	WMPE	CC
Late arrivals	41.22%	40.44%	39.95%	40.32%
95% CI Lower bound	40.80%	40.02%	39.53%	39.90%
95% CI Upper bound	41.65%	40.86%	40.37%	40.74%
Mean response time	1081.9s	1074.7 s	1046.7 s	1054.7 s
Mean single coverage	55.92%	56.34%	56.78%	56.49%
Mean MEXCLP coverage	28.24%	28.50%	28.87%	28.54%

We see that the WMPE method outperforms all other methods. This is to be expected when an ambulance is busy 80% of the time, since now the first part of the relocation is more important than the end of the relocation, as the ambulance is likely to be interrupted during the relocation. The MPE and convex combination methods also perform better than the shortest path method. The largest improvement is in the mean response time.

If we compare it to our basis grid, we see that the new methods still perform better than the shortest path, although the convex combination method performs worse than the WMPE method. Furthermore, there are roughly 30,000 relocations of ambulances during the 200 days, as opposed to the roughly 45,000 relocations from the method in Section 5.1. The reason for this is that when an ambulance becomes available, it has a higher probability that there is already an incident waiting for a response. Thus a larger decrease in percentage points of late arrivals is achieved while the number of relocations is lower.

5.3 Corner Demand

In the following two subsections we want to test the influence of the demand pattern on the performance. To do this, we made two more grids with a different distribution of the demand. In this section we discuss the first of these grids. This grid is shown in Figure 9. This grid has a higher demand in the corners. We have done this to simulate a more rural area with a few villages. The base locations, hospitals and route points are in the same place as in Figure 8.

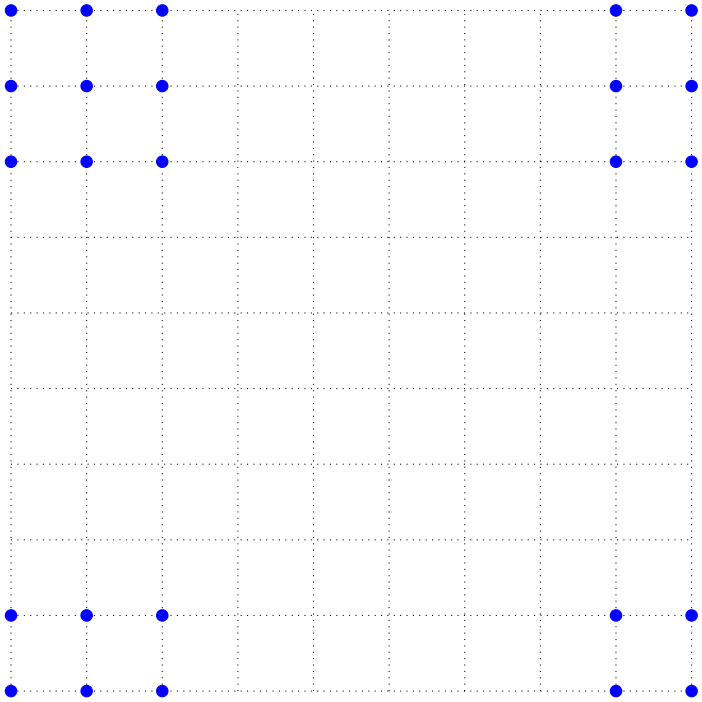


Figure 9: *A grid with different demands on each node. Blue nodes have a demand of* $\frac{1}{40}$ *and the rest of the nodes have a demand of* $\frac{1}{200}$.

We use the following parameters for this grid. The arrival rate is set to $\lambda = 0.0015$ for this grid. This gives us a busy fraction of $q = 0.290$, which is almost the same busy fraction as in Section 5.1. Table 4 shows the results for $L = 8$ minutes. This gives us the following results.

Table 4: Simulation results for $L = 8$ minutes.

	SP	MPE	WMPE	CC
Late arrivals	4.68%	4.63%	4.61%	4.10%
95% CI Lower bound	4.42%	4.37%	4.36%	3.86%
95% CI Upper bound	4.94%	4.89%	4.87%	4.34%
Mean response time	251.87 s	251.13 s	251.13 s	250.64 s
Mean single coverage	98.43%	98.42%	98.43%	98.44%
Mean MEXCLP coverage	89.02%	89.02%	89.02%	89.03%

We see again that the convex combination method performs better on this grid than the other methods. The performance with respect to late arrivals is improved with 0.58 percentage points, when compared to the shortest path method. The 95% confidence interval of the number of late arrivals also has no overlap with the 95% confidence intervals of any of the other methods. We further have that the WMPE method outperforms the MPE method. Note that the MPE method has a lower mean single and MEXCLP coverage than the shortest path method, as well as the WMPE and convex combination method.

What stands out is that the performance of the MPE method is much closer to the performance of the shortest path method when compared to our basis grid, while the convex combination method still performs better. This could be because we did not change the route points, which are still in the center of the grid. Thus the ambulances drive more often through the center, where the demand is lower, and thus they do not contribute much extra to the coverage. The convex combination method takes all arcs in consideration when determining the best path, thus one gets more diverse routes. We conclude that the location of the route points is very important for the performances of the MPE and WMPE methods.

5.4 Bubble Demand

In this section we look at our second grid where we changed the demand pattern. Previously, we put a higher demand at the corners of the grid. This time we put higher demand in the middle. This is shown in Figure 10. We moved the hospitals in the grid, since it makes more sense that not all hospitals are in the corners, were the demand is lower than in the middle. The hospitals are indicated by a blue outline. The base locations and route points are still in the same place.

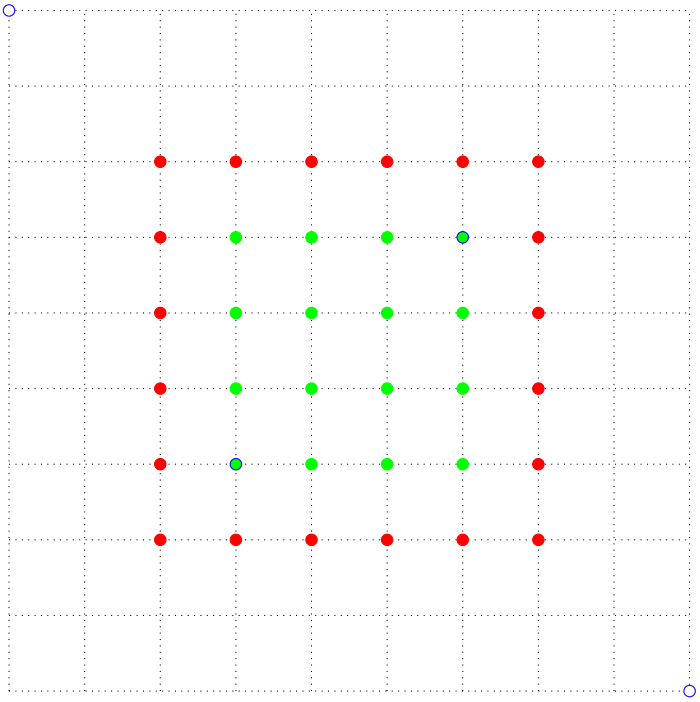


Figure 10: *A grid where the demand is concentrated in the middle. The green nodes have a demand of* $\frac{3}{100}$*, the red nodes have a demand of* $\frac{1}{100}$ *and the other nodes have a demand of* $\frac{1}{200}$*. The nodes with a blue outline also contain hospitals.*

We keep the same arrival rate for the incidents. The busy fraction for this grid is $q = 0.29$. Table 5 shows the result for $L = 8$ minutes.

Table 5: Simulation results for $L = 8$ minutes.

	SP	MPE	WMPE	CC
Late arrivals	1.47%	1.47%	1.40%	1.35%
95% CI Lower bound	1.32%	1.32%	1.26%	1.20%
95% CI Upper bound	1.61%	1.62%	1.54%	1.49%
Mean response time	197.76 s	192.25 s	193.94 s	195.80 s
Mean single coverage	99.26%	99.27%	99.27%	99.27%
Mean MEXCLP coverage	94.86%	94.88%	94.88%	94.87%

What is striking is that the MPE method has a similar performance to the shortest path method with respect to late arrivals. On the other hand the WMPE method performs better than the shortest path and the MPE method. This can be explained by the fact that the WMPE method has a bigger focus on the beginning of a relocation route, which might prevent it from taking longer routes than necessary. The convex combination method still performs the best of all methods with respect to late arrivals, although the relative increase is smaller, 0.12 percentage points, when compared to our base grid, where the improvement was 0.41 percentage points. A reason for this is that the number of late arrivals in these simulations is lower than in the base simulations.

When we compare the results of this grid to our basis grid, we see that the number of late arrivals has decreased for all relocation methods. The reason for this is that the incidents are much more concentrated in a central spot. This makes the area easier to cover with the available ambulances, since the base location in the middle provides good coverage of the area with higher demand.

5.5 Larger Grid

In Figure 11 we increased the size of the grid by a factor four, in order to see how our methods perform on larger graphs. When increasing the size of the grid, the hospitals, base locations and route points are also relocated. We increased the number of base locations and ambulances in the grid as well. There are now eighteen ambulances in the grid, with two starting at each base location.

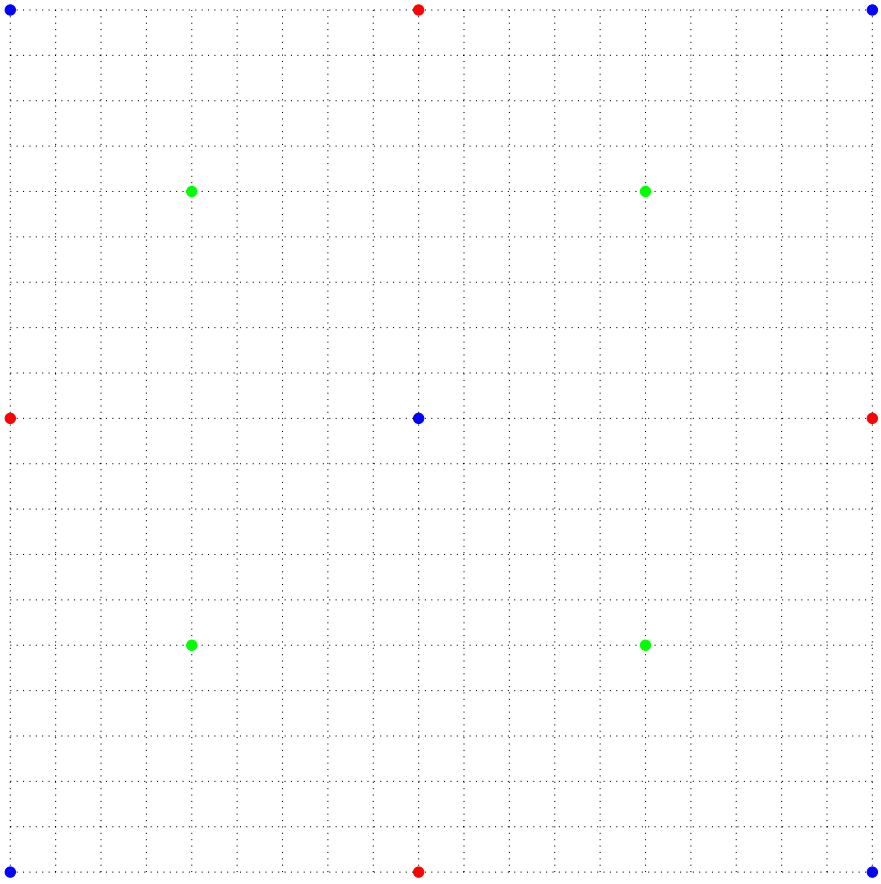


Figure 11: *A* 20×20 *version of our grid. There are hospitals and base locations in blue nodes. Red nodes only have base locations, and green nodes are route points for the MPE method.*

In this larger grid with a new configuration we have to calculate the busy fraction again. We get a busy fraction of $q = 0.333$ when we use the static policy, thus we use this as the busy fraction for our simulations. Since we increased the size of the grid, we also increased the time threshold for late arrivals to $L = 12$ minutes. See Table 6 for the results.

Table 6: Simulation results.

	SP	MPE	WMPE	CC
Late arrivals	2.85%	3.23%	3.15%	2.33%
95% CI Lower bound	2.64%	3.01%	2.93%	2.15%
95% CI Upper bound	3.05%	3.44%	3.36%	2.51%
Mean response time	388.64 s	361.31 s	360.63 s	371.12 s
Mean single coverage	97.58%	97.59%	97.58%	97.58%
Mean MEXCLP coverage	88.46%	88.53%	88.52%	88.50%

The MPE and WMPE method do not perform better than the shortest path method in terms of late arrivals. This could be, because the grid changed too much when we upscaled it. The length of the routes in general is longer than in the base grid. When an ambulance drives over one of the alternative routes, it arrives even later at its destination. Furthermore, the base locations are further away from each other, and have to cover a larger area. Thus it could be more important that the ambulance reaches it destination on time, since the base location has to cover a larger area. Another possibility is that when deciding a route for a relocating ambulance, we do not take into consideration that another ambulance might become available during its relocation. We also do not take into consideration where the other relocating ambulances are going, since we freeze them in place when determining the route. This might result in routes that cover the same area multiple times. Finally, it is also possible that the route points have been chosen poorly.

However, these methods have the best performance with respect to mean response time. The WMPE method does outperform the MPE method except in coverage, although the difference is small for both coverage measures.

On the other hand, the convex combination method outperforms the shortest path method in every performance measure. There is no overlap between the 95% confidence intervals for the late arrivals. This could be because the convex combination method prefers shorter routes over longer routes, thus an ambulance does not arrive much later at its destination when compared to the shortest path method.

5.6 Grid with Hole

In this section we tested our methods on a grid with a hole in it, inspired by Central Park in Manhattan. The grid is shown in Figure 12. We tested two situations. In the first situation the route along the black nodes was inaccessible. In the second situation, ambulances could drive over the route along black nodes, creating a short-cut. The demand of the black nodes is zero, hence there can be no incidents there. We also changed the location of the route points, since the original points are no longer on the grid. We placed them in the corners of the hole. One base location is also gone, since the number of nodes in the grid is decreased. Thus there are only eight ambulances instead of ten.

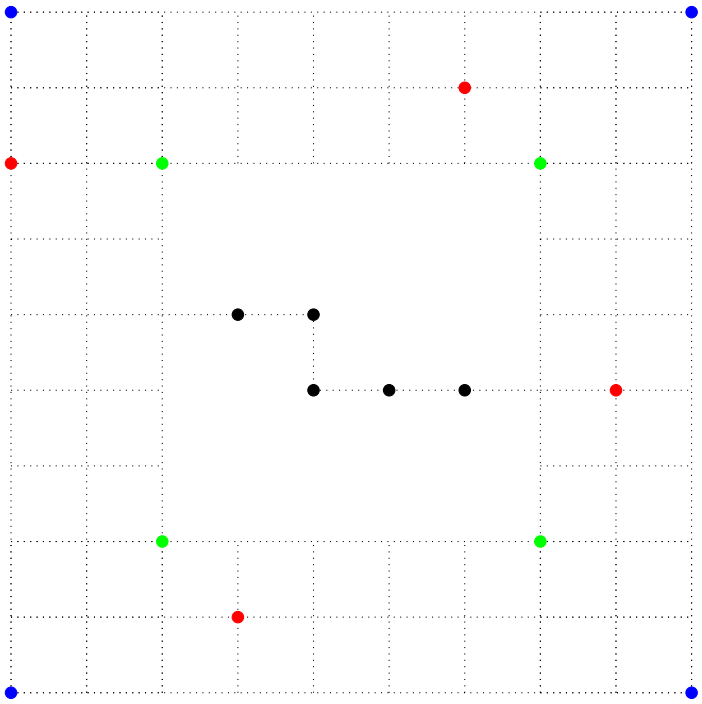


Figure 12: *A grid with a hole in it. Blue nodes represent hospitals, red nodes are base locations, and green nodes are route points for the MPE method. The black nodes indicate the additional route.*

In this grid we have that the busy fraction is $q = 0.291$. Table 7 shows us the results for the grid without the shortcut along the black nodes, while Table 8 shows the results with the shortcut.

Table 7: Simulation results without shortcut.

	SP	MPE	WMPE	CC
Late arrivals	13.27%	13.54%	13.52%	8.70%
95% CI Lower bound	12.85%	13.13%	13.10%	8.35%
95% CI Upper bound	13.68%	13.96%	13.94%	9.04%
Mean response time	311.63 s	310.94 s	311.17 s	201.75 s
Mean single coverage	90.19%	90.23%	90.22%	91.63%
Mean MEXCLP coverage	75.44%	75.48%	75.47%	77.07%

Table 8: Simulation results with shortcut.

	SP	MPE	WMPE	CC
Late arrivals	13.34%	13.10%	13.11%	9.26%
95% CI Lower bound	12.92%	12.69%	12.70%	8.90%
95% CI Upper bound	13.75%	13.51%	13.52%	9.61%
Mean response time	309.68 s	305.72 s	305.95 s	217.10 s
Mean single coverage	90.20%	90.23%	90.23%	91.35 %
Mean MEXCLP coverage	75.48%	75.51%	75.52%	76.80 %

We see that the convex combination method performs better than all other methods in both cases. There is an improvement of 4.57 percentage point in the number of late arrivals without the shortcut, and an improvement of 4.08 percentage point with the shortcut. The 95% confidence interval of the number of late arrivals also has no overlap with the 95% confidence intervals of the other methods. There is also a decrease in mean response time of 109.88 seconds and 92.58 seconds respectively.

On the other hand, the MPE and WMPE method do not perform better than the shortest path method with respect to late arrivals when there is no shortcut. The mean response times and mean coverages on the other hand are better. When there is a shortcut, the MPE and WMPE method improve on the shortest path method for all performance measures. A reason for this is that the shortest path is more likely to pick the shortcut, while taking another route would improve the coverage of the system.

It is striking that the shortest path and convex combination method perform worse when there is a shortcut. A possible reason for this is that the shortcut itself has demand zero, i.e. there are no incidents on the shortcut. Thus while taking the shortcut ensures that one reaches their destination earlier, the coverage of the grid during the travel time is lower. If an incident happens while the ambulance is on the shortcut, it is less likely to respond in time.

5.7 Random Lengths on Edges

During the simulations we noticed that the convex combination method performed better than any other method. One of the explanations why it works so well on the grids is that there are many shortest paths between two nodes, since the length of the edges is the same. Thus the shortest path method takes a random route, while the convex combination method picks the path with the best coverage among these shortest paths. This made us wonder if the method would still improve the result if we changed the lengths of the arcs so that there are fewer shortest paths. To do this we took the basis grid in Figure 8 and picked the length of the arcs uniform random between 75 and 125. We picked these values so that the graph still satisfies the triangle inequality. The arcs can only be integer values. We equip the grid with uniform demand distribution. Our new grid has a busy fraction of $q = 0.293$.

The results are shown in Table 9. Since the length between two nodes is different, we used multiple values of α in the convex combination method. Note that $\alpha = 1$ results in the shortest path method.

Table 9: Simulation results.

	SP	MPE	CC $\alpha = \frac{1}{2}$	CC $\alpha = 0$
Late arrivals	3.85%	4.39%	3.77%	3.65%
95% CI Lower Bound	3.61%	4.14%	3.53%	3.42%
95% CI Upper Bound	4.08%	4.64%	4.00%	3.87%
Mean response time	239.63 s	238.39 s	239.28 s	238.32 s
Mean single coverage	98.12%	98.32%	98.14%	98.14%
Mean MEXCLP coverage	89.90%	91.08%	89.90%	89.90%

At first we see that the MPE method performs considerably worse when compared to the shortest path. This is most likely caused by the fact that the route points have not changed, so they might be in unfavorable positions. Although an improvement on the shortest path method is still possible. The convex combination method performs better for both values of α. However, if we compare it to our basis grid, the improvement is relatively smaller. This is probably because there are fewer shortest paths.
It is also striking that $\alpha = 0$ outperforms $\alpha = \frac{1}{2}$. This implies that the coverage part in Equation (4) is more important than the distance part. Note that the coverage part of the equation also takes the distance into account, to normalise the equation. Thus if one takes $\alpha = 0$, one does not lose the influence of the distance completely.

6 Flevoland Simulations

Ultimately, we want to test our methods on real EMS regions to see if we get possible improvements. Thus we made simulations for the EMS region Flevoland. We also look at other regions in Section 9, where we use TIFAR for our simulations. First, we talk about the region itself and we explain how the ambulance care is organised in Flevoland in Section 6.1. In Section 6.2 we discuss the results of the simulations.

6.1 EMS Region Flevoland

Flevoland is a rural area, with nearly 400,000 citizens. Almost half of these citizens live in Almere, the biggest city of Flevoland. The rest of the people are mainly living in one of the five other towns. The base locations in Flevoland are located in these six towns, and there are hospitals in Almere and Lelystad. Figure 13 shows the region and its base locations, as well as the route points we use.

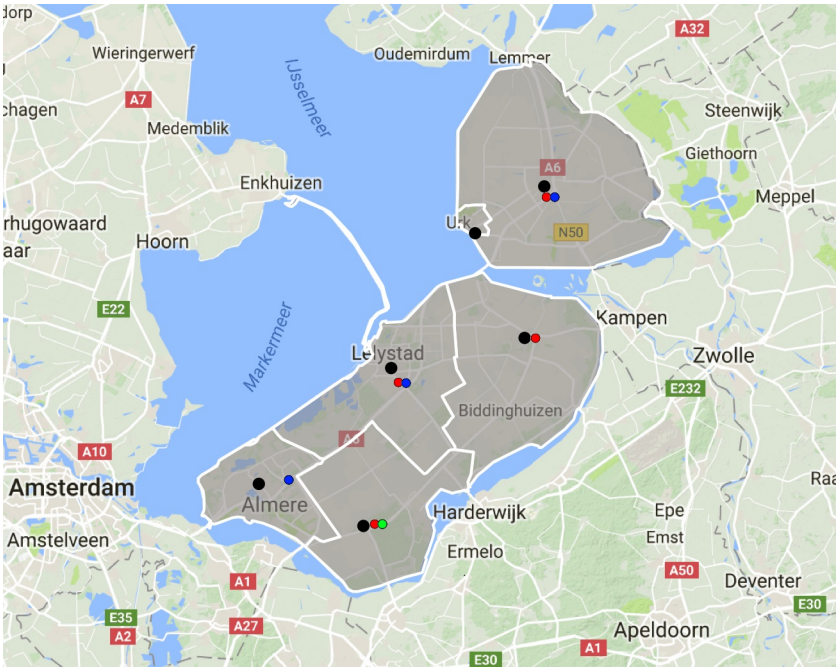


Figure 13: *Map of the EMS region Flevoland. The grey areas are the municipals of Flevoland, and the base locations are represented by black dots. The route point of RP_1 is colored green, the route points of RP_2 are colored red and the route points of RP_3 are colored blue.*

We model the region of Flevoland using the zip code areas as nodes in a directed graph. Flevoland has 91 zip code areas. The driving time between each pair of zip code areas are estimated by the Rijksinstituut Volksgezondheid en Milieu (National Institute for Public Health and Environment). See [7] for more information of how these times were estimated. These driving times are for ambulances with their sirens and optical signals turned on. To obtain the travel time of a relocating ambulance, we multiplied the driving times with $\frac{10}{9}$.

GGD Flevoland, the ambulance service provider of the region Flevoland, has provided us with historical data of emergency requests in the year 2011. We used this data to estimate the demand d of each area and the arrival rate λ of incoming incidents. For this simulation we divided a day into 24 time blocks of one hour each. We calculated the demand of an area and the arrival rate for each of these time blocks. The arrival rate is based on $A1$ and $A2$-calls. However, we assume that every incoming incident has priority $A1$, since it becomes too complicated to work with two time thresholds. If we neglect the $A2$ priority incidents on the other hand, we do not get a realistic simulation, since 27% of the incidents have priority $A2$ [1].

We also take the schedule of the ambulances in Flevoland into account. For the previous grids, we assumed that the number of ambulances was constant over time and that their shift never ends. Now an ambulance has to return to its base to end its shift, and the number of ambulances changes throughout the day. The ambulance schedules are provided by GGD Flevoland.

To apply the MPE method, we need a set of route points. We use the following set of route points in our simulation.

1. RP_1: one route point in Zeewolde.
2. RP_2: four route points in Zeewolde, Dronten, Lelystad and the Noordoostpolder.
3. RP_3: three route points in Almere, Lelystad and the Noordoostpolder.

We chose RP_1 because we are interested in alternative routes from Almere to Dronten through Zeewolde. We chose RP_2 to see the effects of adding more than one route point. Finally we chose RP_3 since it passes the major cities in Flevoland. See Figure 13 for the location of these route points.

6.2 Results

In this section, we discuss the results of the MPE, WMPE and CC method for Flevoland. We also look at the effect of the methods on different times of the day, since the number of incidents and the number of ambulances is different during the day, the evening and the night. Furthermore, we used real data from 2011 to see how well our methods perform.

First we tested the MPE method. We used the route points shown in Figure 13. Using this method, we get the results as shown in Table 10.

Table 10: Simulation results for the MPE method.

	SP	MPE (RP_1)	MPE (RP_2)	MPE (RP_3)
Late arrivals	6.06%	5.78%	5.71%	5.90%
95% CI Lower bound	5.61%	5.34%	5.28%	5.46%
95% CI Upper bound	6.50%	6.21%	6.16%	6.34%
Mean response time	279.18 s	279.18 s	278.01 s	279.08 s
Mean single coverage	96.29%	96.30%	96.27%	96.26%
Mean MEXCLP coverage	90.19%	90.19%	90.20%	90.15%

The MPE method is effective in decreasing the number of late arrivals, since all collections of route points give an improvement. Even the set RP_1 that only contained one route point gives an improvement of 0.28 percentage point. Adding more route points like in RP_2 decreases the number of late arrivals even more, although the extra effect of the three extra route points is small compared to just adding the initial route point. We also see that more route points does not necessarily mean that one gets better results, since RP_1 has less late arrivals than RP_3. Thus one has to choose the route points carefully, since taking more route points leads to a longer computation time.

We now look at the WMPE method for these sets of route points. They are shown in Table 11.

Table 11: Simulation results for the WMPE method.

	WMPE (RP_1)	WMPE (RP_2)	WMPE (RP_3)
Late arrivals	6.48%	6.50%	6.31%
95% CI Lower bound	6.02%	6.04%	5.88%
95% CI Upper bound	6.94%	6.97%	6.77%
Mean response time	285.86 s	284.56 s	281.49 s
Mean single coverage	96.29%	96.24%	96.24%
Mean MEXCLP coverage	90.18%	90.13%	90.15%

The WMPE method performs worse than the MPE method, while it performed better on the grids. It also performs worse than the shortest path method. An explanation could be that the ambulances are not interrupted as often as on the grid, since the arrival rate for incidents is lower in Flevoland than in the grids. Thus, if the start of a relocation has higher weight, then the ambulance might take a less favorable route and arrive later at an incident. This could also be a result of

a badly chosen value for c, but we did not have time to test it for any other values of c.

We saw that the MPE method performs better in Flevoland than the shortest path method, but we also want to know during which part of the day it performs better, and during which part it performs worse. To accomplish this, we divided the day into three parts. These three parts are the night from 0:00 to 8:00, the day from 8:00 to 16:00 and the evening from 16:00 to 0:00. Figure 14 shows the number of late arrivals during each of these time periods for the MPE method. We omitted the WMPE method since it did not give an improvement on the shortest path method.

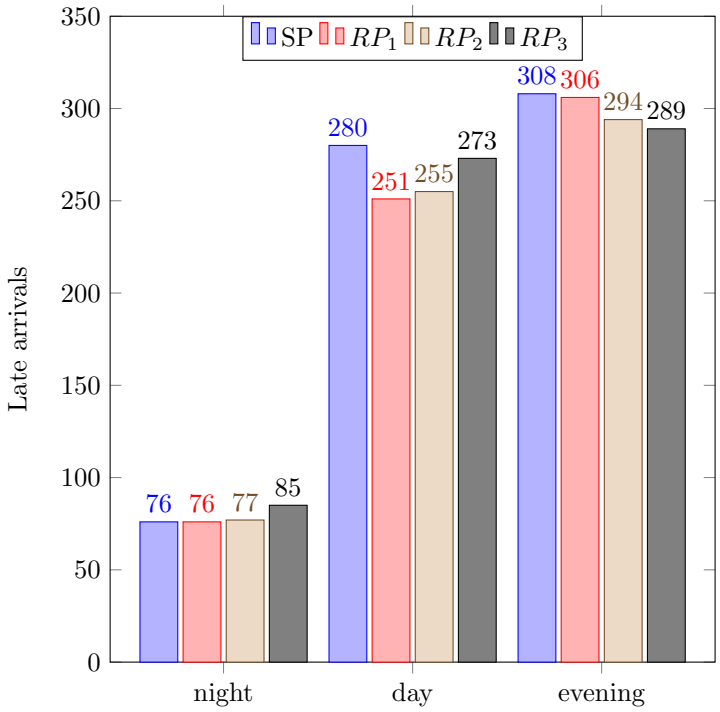


Figure 14: *The number of late arrivals during the morning, afternoon and night for various methods.*

It turns out that the MPE method does not perform better in the night. This could be because there are fewer incidents during the night, thus taking longer routes when relocating can be less efficient, since you are less likely to be interrupted during your relocation. Most of the improvement happens during the day and evening, where the number of incidents, and therefore late arrivals, are higher in general. This leads us to take a closer look at the time during the day and see how our methods perform during this period. We took real life data from the year 2011 in Flevoland. This data contains only incidents that happened during 7:00 and 20:00. We do not take the shifts of the ambulances into account, and we assume that there are ten ambulances in the region. In Table 12 we show the results during this period for various methods.

Table 12: Simulation results with data from 2011.

	SP	MPE	CC $\alpha = \frac{1}{2}$	CC $\alpha = 0$
Late arrivals	4.27%	4.14%	4.14%	4.10%
95% CI Lower Bound	3.82%	3.69%	3.69%	3.66%
95% CI Upper Bound	4.73%	4.59%	4.59%	4.55%
Mean response time	290.8 s	292.6 s	290.1 s	290.4 s
Mean single coverage	97.39%	97.39%	97.39%	97.39%
Mean MEXCLP coverage	93.45%	93.44%	93.45%	93.45%

We see that the MPE method and the convex combination method perform better during the day than the shortest path method. The late arrivals decrease by 0.13 percentage points for the MPE method and 0.17 percentage points for the CC method with $\alpha = 0$. Furthermore, we have that $\alpha = 0$ performs better than $\alpha = \frac{1}{2}$ for the convex combination method, just like in Section 5.7. We also see that the mean response time does not improve when compared to the shortest path method. Note that the percentage of late arrivals is lower than in Table 10, because we only consider the daytime, and the number of late arrivals is relatively smaller there.

7 Introduction to TIFAR

In this part of the thesis we discuss Testing Interface For Ambulance Research (TIFAR), a simulation tool developed by Stokhos which can be used to evaluate different ambulance dispatch strategies and help dispatchers when and where to relocate an ambulance to. Figure 15 shows the interface of TIFAR. Our goal is to add a relocation policy to the TIFAR simulation tool and see what the effects are on real Emergency Medical Service Regions in the Netherlands. In this section we give a brief explanation of how TIFAR works. See [8] for a more detailed explanation.

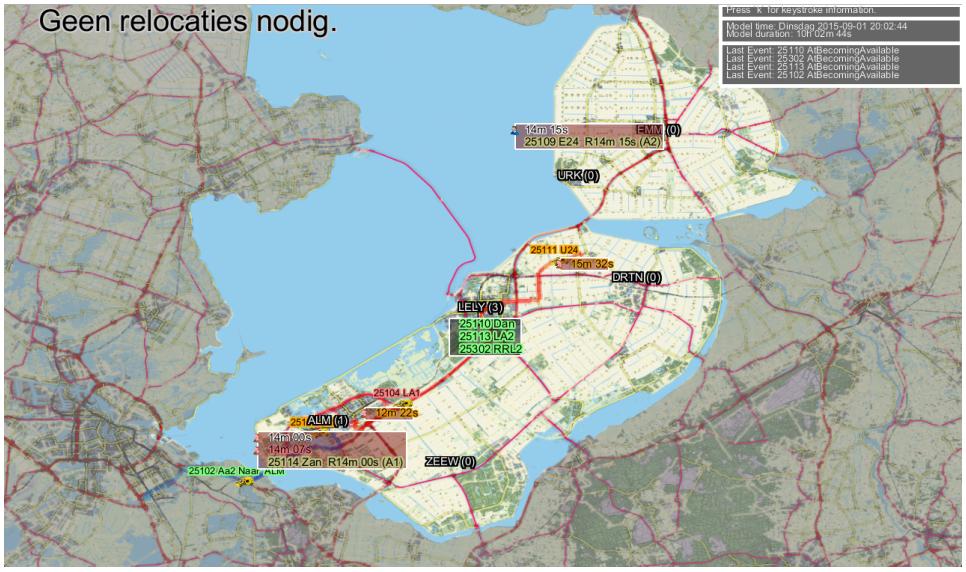


Figure 15: *The interface of TIFAR. The base locations are indicated with with names, and the number between the brackets are the number of ambulances at the base. Ambulances responding to an incident follow a red route. In the top left is shown which relocation should be taken, if any.*

We use real data of the EMS regions from 2015 in our simulations with TIFAR. This data contains:

- when the incident occurred,
- where it occurred,
- how long the ambulance has to be on-scene for treatment of the patient,
- if the patient has to go to the hospital, including which hospital,
- the time the ambulance has to spend at the hospital (if the patient needs to go to the hospital),
- the urgency of the call.

After an incident occurs, we have to dispatch an ambulance to the incident. All the calls are put into a queue. Handling of arriving calls is modelled as a priority

queueing system. Here, $A1$-calls are first-come-first-served, and $A2$-calls have lower priority and are handled when there are no calls with urgency $A1$ left or as soon as 15 minutes have passed, since we want $A2$-calls to be answered within 30 minutes. We assign the ambulance that can respond the quickest to the incoming call. Note that only one ambulance will be assigned to a call and that we do not reassign ambulances to the call.

The speed of an ambulance depends on the type of road and whether it uses its sirens and optical signals. If an ambulance responds to an $A1$-call, we assume that it uses its sirens and optical signals, and when it responds to an $A2$-call, we assume that it does not use them.

TIFAR can also use multiple relocation rules. We already discussed the Combi Algorithm in Section 3.3. However, for our simulations we use the Dynamic MEXCLP (DMEXCLP) algorithm. DMEXCLP uses the definition of coverage of the MEXCLP model discussed in Section 3.2. The MEXCLP coverage is computed when an ambulance becomes available for relocation purposes. When calculating the coverage of the system, we only consider the destination of the idle ambulances, since that will be the state of the system if no further incidents happen. Let n_i be the number of idle ambulances that have i as their destination. When an ambulance becomes idle and needs to be relocated, the DMEXCLP model sends the ambulance to the waiting site that results in the largest MEXCLP coverage. This is the same as choosing the base that results in the largest marginal coverage over all demand. This can be interpreted as the benefit of having an additional k-th ambulance near node i, given by $d_i(1-q)q^{k-1}$. The waiting site that gives us the biggest improvement in coverage can be calculated by

$$\underset{w \in W}{\operatorname{argmax}} \sum_{i \in V} d_i (1-q) q^{k(i,w,n_1,\ldots,n_{|W|})-1},$$

where

$$k(i, w, n_1, \ldots, n_{|W|}) = \sum_{j=1}^{|W|} n_j \mathbb{1}_{\{\tau_{i,j} \leq L\}} + \mathbb{1}_{\{\tau_{w,i} \leq L\}}$$

is the number of idle ambulances that can reach node i if we assume that the ambulance is relocated to waiting site w. The symbol $\mathbb{1}$ stands for the indicator function. With this method we determine the destination of our relocation. For more information about the DMEXCLP algorithm, we refer to the paper by Jagtenberg, Bhulai and van der Mei [9].

In Section 8 we add dynamic routing to TIFAR and in Section 9 we discuss the impact of this change.

8 Dynamic Routing in TIFAR

In this section we describe how we implement dynamic routing in TIFAR. Whenever a relocation occurs, we determine the origin and the destination using the DMEXCLP algorithm. Given this origin-destination pair, we propose a method that consists of two parts to determine which route the relocating ambulance should take. In the first part we determine multiple possible routes between the origin and the destination. We want these routes not to be too similar. In the second part we evaluate all the generated routes and return the route that gives the best coverage overall.

8.1 Generating routesets

In this section we explain how we create the set containing different routes that be evaluated. The first route we add is the shortest path between the origin and destination. We use *decision points* on the road network to help us generate more routes. Decision points are points on the road network where the route splits and the ambulance has to decide which route it takes. Figure 16 shows all the decision points in Flevoland.

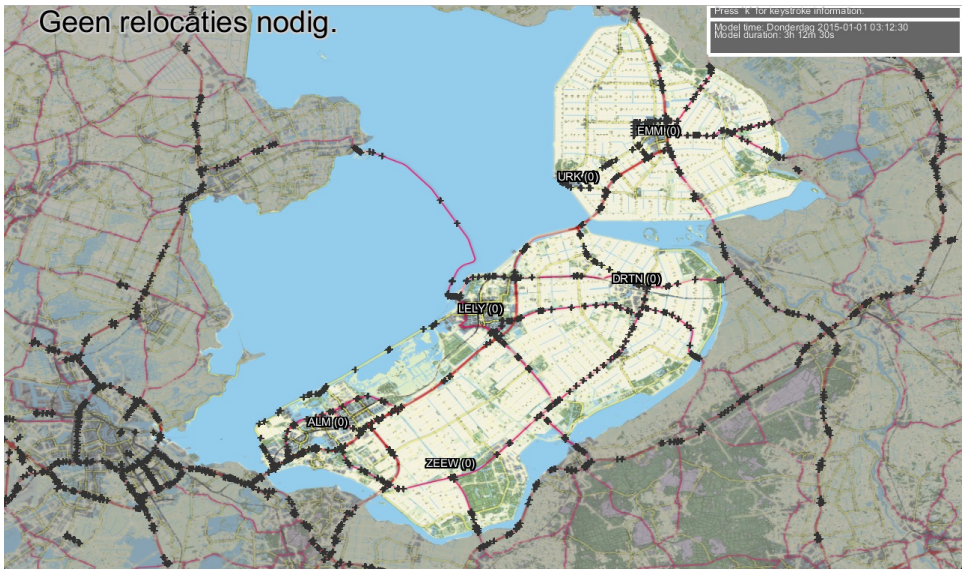


Figure 16: *The decision points of Flevoland are indicated with black crosses. We also take decision points on highways of nearby areas into account.*

When we generate a routeset, we want to look at fewer decision points, since we are only interested in the decision points accessible from the origin and destination within a reasonable amount of time. Let $t_{O,D}$ be the travel time between the origin and destination in seconds. We look at the decision points that can be reached from the origin and the destination in $t_{O,D}$ seconds. See Figure 17 for the decision points between Almere and Dronten.

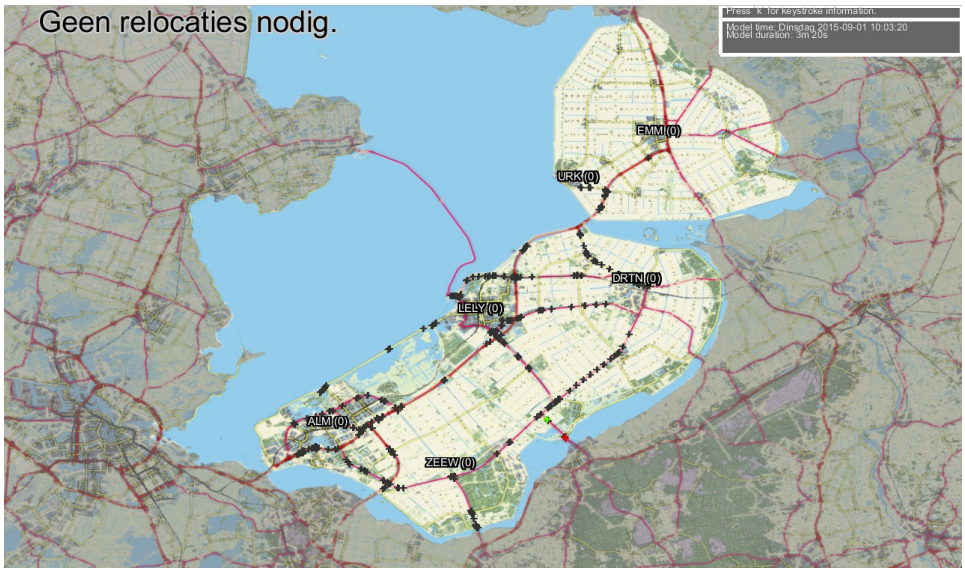
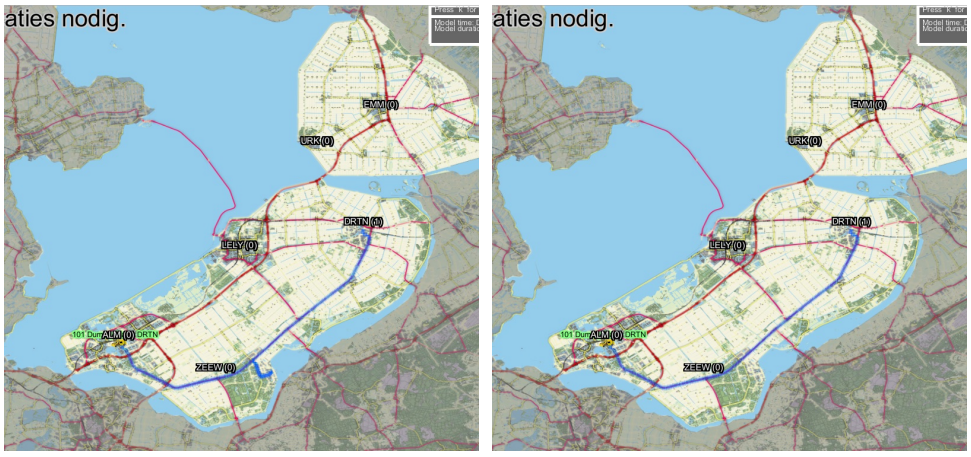


Figure 17: *The decision points reachable from Almere and Dronten within $t_{O,D}$ seconds, where O is Almere and D is Dronten.*

For all the resulting decision points we keep track whether there is already a route that visits the decision point. When we add a route to the routeset, we check all the decision points it travels trough and mark them as visited. After adding the shortest path and marking all the decision points along the way, we look at an unvisited decision point. We first choose the unvisited decision point with the lowest y-coordinate, and try to add the route through that decision point. We then repeat this for the unvisited decision point with the highest x-coordinate, then the highest y-coordinate and next the lowest x-coordinate. Then we look again at the unvisited decision point with the lowest y-coordinate, until all decision points are visited. Thus we mark the decision points in a counter clockwise fashion, starting from the outside. We use this method to speed up routeset generation, because it prevents picking multiple decision points on a road facing outward.
To illustrate this we look at Figure 17 again. If one first takes the route through the green decision point, one will not reach the red decision point, and thus we have to look at a route through the red decision point at a later time as well, since it is still unvisited. This route will be very similar to the route through the green decision point, and we want to avoid similar routes. Thus by picking the more outward decision points first, we mark more decision points and prevent looking at too many similar routes.

We also apply snapping to the route. This is a procedure designed to prevent going back and forth over the same path when we create the shortest path through the decision point by removing paths that we travel over twice. However, the decision point of the original route should remain covered. Thus we can only remove a path if one can reach the original coverage point within 15 minutes. See Figure 18 for an example of snapping. Here we have a route from Almere through a node in Zeewolde to Dronten. Figure 18a shows the route without snapping. We see that this route travels over the same road twice while entering Zeewolde. In Figure 18b

we applied snapping, and we do not go into Zeewolde, but we still cover the area because the ambulance drives past it.



(a) Route without snapping.

(b) Route with snapping.

Figure 18: *We have a route from Almere to Dronten through Zeewolde.*

When we create a route, we try to add it to the routeset. We compare the route to each other route in the set, since we want to generate multiple routes that do not have much overlap. If there is another route in the routeset that has at least δ% overlap with the new route, then we do not add the new route to the routeset. Thus lower values of δ result in smaller routesets. We can take multiple values for δ. Figures 19a, 19b and 19c show multiple routesets that are created between Almere and Dronten for different values of δ. It shows the different roads that are taken, but the routes have some overlap. If the black line is thicker, then there are more routes that travel over that road.

(a) $\delta = 10$, 2 routes were added.

(b) $\delta = 50$, 6 routes were added.

(c) $\delta = 90$, 15 routes were added.

Figure 19: *Routesets between Almere and Dronten with different values for* δ.

We want diverse routesets, but they should not be too large, since this increases the computation time. In the simulations we use $\delta = 50$.

8.2 Evaluating the routeset

After generating the routeset R for origin O and destination D, we need to evaluate each route in the routeset. The routes consist of a series of nodes. We introduce some notation that we use to give each route a value. Let t_i be the travel time from the origin to node i. Furthermore let $x(t)$ be the position of the ambulance at time t. The position of the ambulance can be determined much more precisely in TIFAR, as opposed to our first simulation model. We can now determine where the ambulance is between nodes. When we evaluate the routes, we assume that the other idle ambulances do not move and that no additional ambulances become available. Finally we use the function $f(\tau_{i,j})$ to determine whether node j is covered when the ambulance is in node i. If node j is covered, then $f(\tau_{i,j})$ is equal to the contribution to the coverage of node j, otherwise it is zero. We use the MEXCLP function, which gives us

$$f(\tau_{i,j}) = \mathbb{1}_{\{\tau_{i,j} \leq L\}} d_j (1-q) q^{k_j - 1},$$

where k_j is the number of other ambulances that can reach node j within the time threshold L. The approximation of the coverage of a route $r \in R$ consists of two parts. The first part is the contribution of driving along the route itself, which has the following formula:

$$\sum_{i \in r \setminus O} \sum_{j \in N} \int_{t_{i-1}}^{t_i} f(\tau_{x(t),j}) e^{-\gamma t_i} dt.$$

For each node $i \in r \setminus O$ we look at the contribution to the coverage of all demand nodes when the ambulance is in node i. The term $e^{-\gamma t_i}$ is used give higher weight to the start of the route, similar to the WMPE method in Section 4.2. Here γ is a scaling parameter that simulates uncertainty when time goes on. For γ we take the arrival rate between incidents in the region, taken from the data of 2015.

The second part of the coverage is the contribution of staying at the destination. The ambulance arrives at the destination after t_D seconds. The term $e^{-\gamma t_i}$ is used to make sure the contribution of this term is finite. Thus for the second part we have:

$$\sum_{j \in N} \int_{t_D}^{\infty} f(\tau_{D,j}) e^{-\gamma t'} dt'.$$

Note that the contribution of this term is greater the sooner you arrive at the destination. Without this term, the ambulances would keep on driving and never arrive at their destination, since arriving at the destination would not give additional coverage in the equation.

When we combine these two parts, we get the following approximation of the coverage for routes $r \in R$:

$$C_r = \sum_{i \in r \setminus O} \sum_{j \in N} \int_{t_{i-1}}^{t_i} f(\tau_{x(t),j}) e^{-\gamma t_i} dt + \sum_{j \in N} \int_{t_D}^{\infty} f(\tau_{D,j}) e^{-\gamma t'} dt'. \tag{5}$$

9 Simulation Results

In this section we discuss the simulation results. We are interested in three regions, namely Gooi & Vechtstreek, Amsterdam and Utrecht. We used the historical data from 2015 for our simulations. We simulated the months September and October, since there are no major holidays in these months. We compare our dynamic routing method to the DMEXCLP method without dynamic routing. To compare both methods we look at the number of late arrivals and the mean response time. We also show where in the region we have the largest improvement.

9.1 Gooi & Vechtstreek

The EMS region Gooi & Vechtstreek is located north of Utrecht, and it is one of the smallest EMS regions in the Netherlands. It is a rural area with a population just over 250,000. Most of the people live in the towns Hilversum, Huizen, Blaricum, Bussum and Weesp. The ambulance service provider for the region is RAV Gooi & Vechtstreek. Yearly, there are more than 17,000 incidents. There are three base locations, situated in Hilversum, Blaricum and Weesp. Figure 20 shows a map of the region and its base locations.

Figure 20: *The Grey area is the EMS region Gooi & Vechtstreek. The three base locations are indicated with black dots.*

Table 13 shows the results of the simulations with and without dynamic routing.

Table 13: Simulation results with data from 2015.

	DMEXCLP	DMEXCLP with dynamic routing
Late arrivals	16.03%	15.87%
95% CI Lower bound	13.99%	13.84%
95% CI Upper bound	18.06%	17.90%
Mean response time	682.72s	688.31s

The high percentage of late arrivals is due to a shortage of ambulances in the evening, especially in the weekends. At that time there are barely enough ambulances available to handle the incoming calls. We also see that there is a small increase in mean response time. To get a better understanding of where and when we get the most improvement, we look at Figure 21. Nodes are colored green if the dynamic routing method performed better in that node than the DMEXCLP method, thus where dynamic routing had less late arrivals. The node is red if the dynamic routing method performed worse.

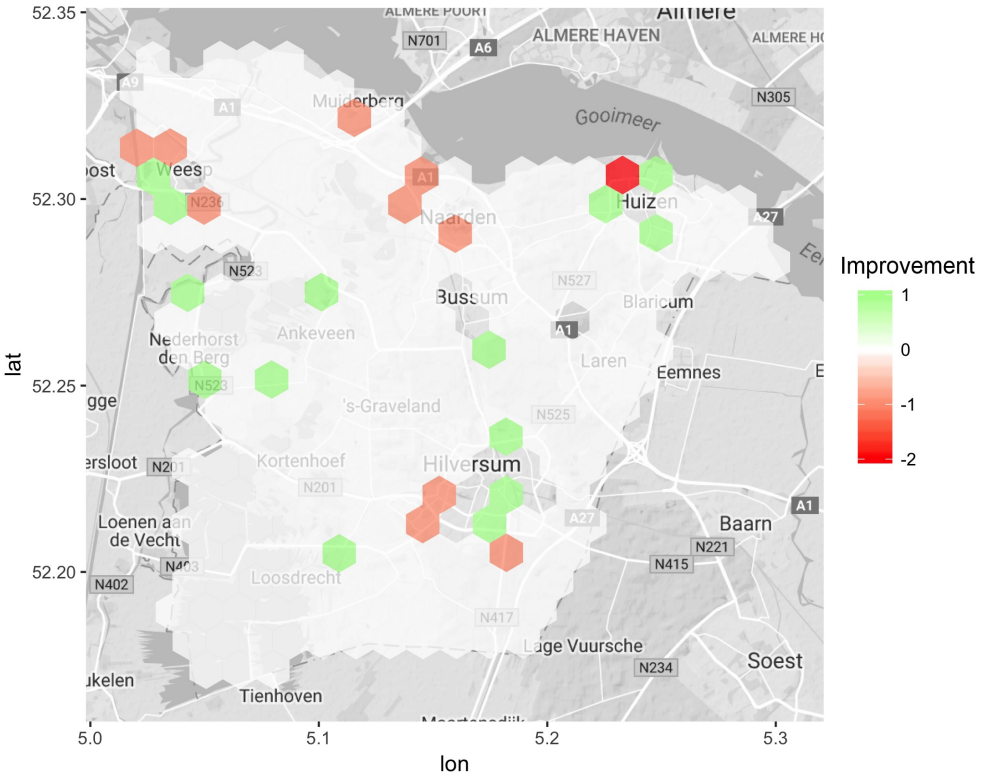
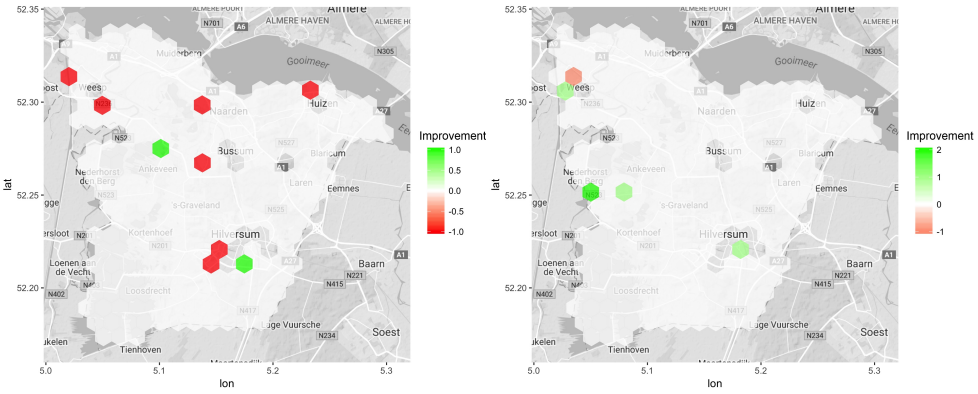


Figure 21: *Comparison of number of late arrivals in Gooi & Vechtstreek.*

We see that the most improvement is in Wijdemeren, the municipal in the south-west. This comes at the cost of more late arrivals in the municipal Gooise Meren, which is located in the North. Dynamic routing thus moves the location of late arrivals to more rural areas, since we now take the highway through Gooise Meren less and instead drive through Wijdemeren.
Since the region is understaffed in the evening, we are interested if our method performs better or worse in the evening compared to the rest of the day. Figure 22 shows the late arrivals in the evening compared to the rest of the day. Here the evening is from 16:00 to midnight. Dynamic routing performed better in green nodes and worse in red nodes.



(a) *Evening.* (b) *Rest of the day.*

Figure 22: *Comparison of number of late arrivals for different times of the day.*

We see that dynamic routing performs much worse during the evening than during the rest of the day. Sometimes there is a shortage in the number of available ambulances during the evening, which might be the reason that dynamic routing performs worse. Note that we used the same γ throughout the simulation. The results might improve if we take different γ for different times of the day.

Finally we looked at the relative improvement for the municipals in the region. Since there are less incidents in Wijdemeren, any extra incident that is reached on time gives a larger relative improvement than in more densely populated areas. Figure 23 shows the relative improvement.

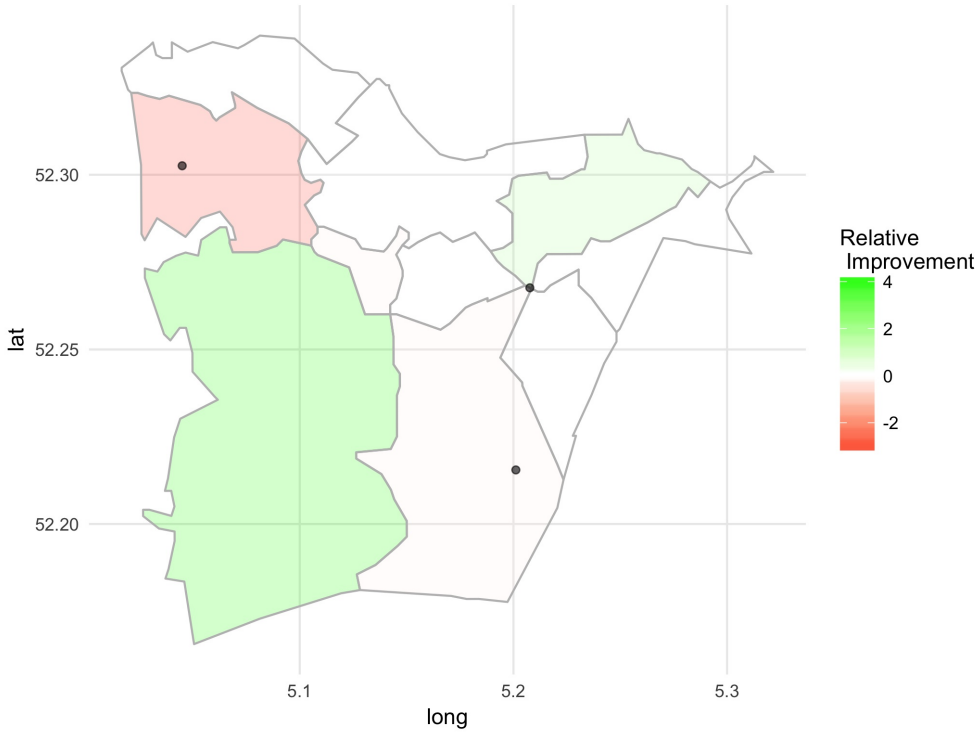


Figure 23: *The relative improvement in late arrivals for Gooi & Vechtstreek. The black dots indicate the base locations.*

We see that there is a relative improvement in Wijdemeren and Huizen, while dynamic routing performs worse in Weesp. Notice that both of these municipals do not have a base location. Wijdemeren normally has the lowest percentage on time arrivals. Dynamic routing redistributes the late arrivals so the percentage late arrivals of each municipal gets closer together.

9.2 Amsterdam

The EMS region of Amsterdam is a combination of two former EMS regions named Zaanstreek-Waterland in the north and Amsterdam-Amstelland in the South. The population of this region is 1,2 million people, with 68% of these people living in the city Amsterdam. This region is densely populated compared to Gooi & Vechtstreek. Figure 24 shows the region as well as the base locations.

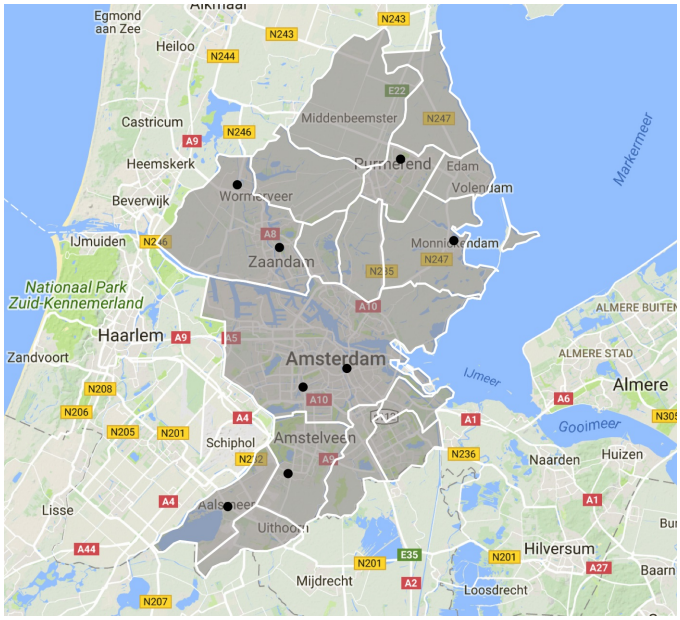


Figure 24: *The Grey area is the EMS region Amsterdam. The eight base locations are indicated with black dots.*

In Table 14 we show the simulation results for Amsterdam and in Figure 25 we show where dynamic routing gave a decrease in late arrivals and where it gave an increase. Green nodes indicate where dynamic routing performed better and red nodes where it performed worse. The base locations are also shown.

Table 14: Simulation results with data from 2015.

	DMEXCLP	DMEXCLP with dynamic routing
Late arrivals	0.69%	0.69%
95% CI Lower bound	0.53%	0.53%
95% CI Upper bound	0.85%	0.85%
Mean response time	469.60s	458.13s

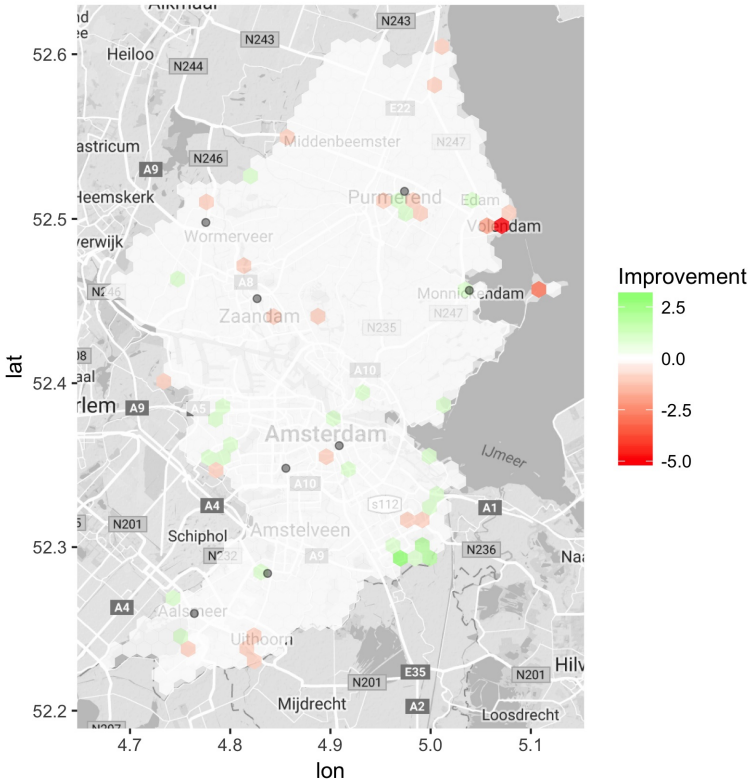


Figure 25: *Comparison of number of late arrivals in Amsterdam. The base locations are indicated with black dots.*

The number of late arrivals is so low because the ambulances that handle B-calls can also respond to A1-calls in this simulation, while they do not do in real life. The late arrivals in Amsterdam stay about the same, but the mean response time decreases when we use dynamic routing. This is mainly because dynamic routing has the biggest improvement in Amsterdam Zuid-Oost (south east Amsterdam), shown in Figure 25. This comes at a trade-off for more late arrivals in the semi-rural areas outside of the city Amsterdam, especially in Volendam. Observe that Amsterdam Zuid-Oost does not have a base location for ambulances, since we used older base locations. Thus dynamic routing sends an ambulance over Amsterdam Zuid-Oost to cover that part of region better.

Nowadays there is a post in Amsterdam Zuid-Oost. Thus dynamic routing can be used to cover an area where one would want a base location, and might even be used to search for appropriate base locations. There are other models to determine the optimal base location in a region, see [10] for more information.

9.3 Utrecht

Utrecht is a densely populated area with approximately 1,2 million inhabitants. It is one of the largest EMS regions in the Netherlands. The ambulance provider of the region is RAV Utrecht and they handle more than 85,000 incidents each year. The region along with its base locations are shown in Figure 26 .

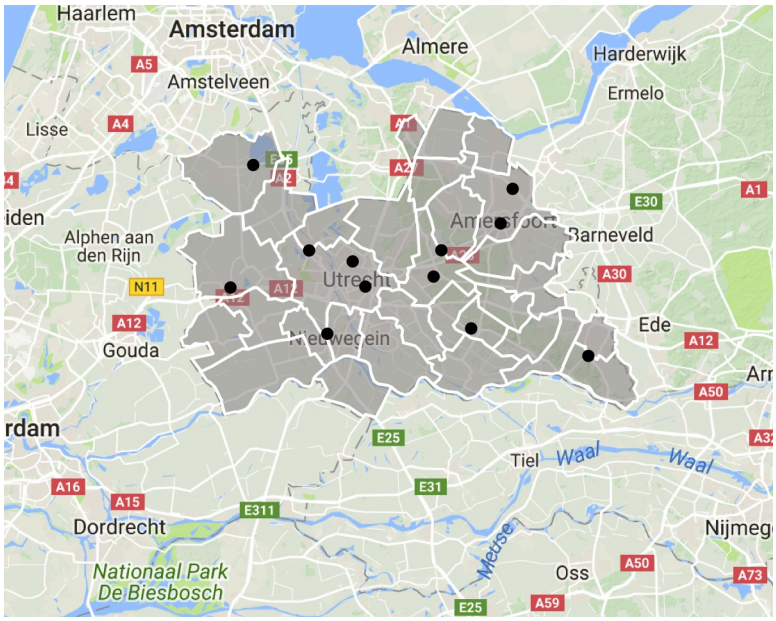


Figure 26: *The Grey area is the EMS region Utrecht. The base locations are indicated with black dots.*

Table 15 shows the late arrivals and the mean response time for Utrecht. Furthermore, in Figure 27 we compare both methods, similar to what we did for Gooi & Vechtstreek and Amsterdam.

Table 15: Simulation results with data from 2015.

	DMEXCLP	DMEXCLP with dynamic routing
Late arrivals	7.03%	7.15%
95% CI Lower bound	6.38%	6.50%
95% CI Upper bound	7.68%	7.81%
Mean response time	571.53s	568.47s

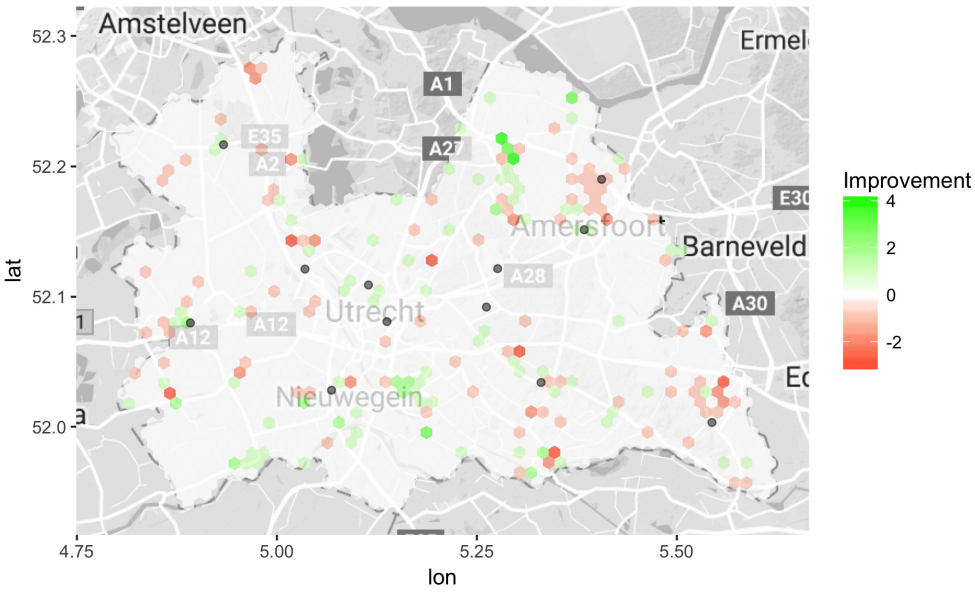


Figure 27: *Comparison of number of late arrivals in Utrecht. The base locations are indicated with black dots.*

Given the results in Table 15, we see a slight increase in the number of late arrivals. In Figure 27 we see that most of the deterioration happens in the cities, Amersfoort and Veenendaal in particular. There is a small decrease in the mean response time as well. This can be because the ambulances respond quicker to incidents farther away from base locations when we use dynamic routing. The most improvement is gained in semi-rural areas with no base location. Especially in Lopik in the south-west and Eemnes in the north-east have many green dots. This is because relocating ambulances drive through these regions more. However, there is a deterioration in the other corners of Utrecht. In both in the north-west and the south-east dynamic routing loses out to the standard DMEXCLP method. Both these regions have base locations, as opposed to Lopik and Eemnes. Thus because of dynamic routing, the ambulances arrive later at the base locations in the corners of the region, which results in more late arrivals.

Since we have the most improvement in more thinly populated areas, we are interested in the relative improvement of the region. In Figure 28 we show the relative improvement for each municipal in Utrecht.

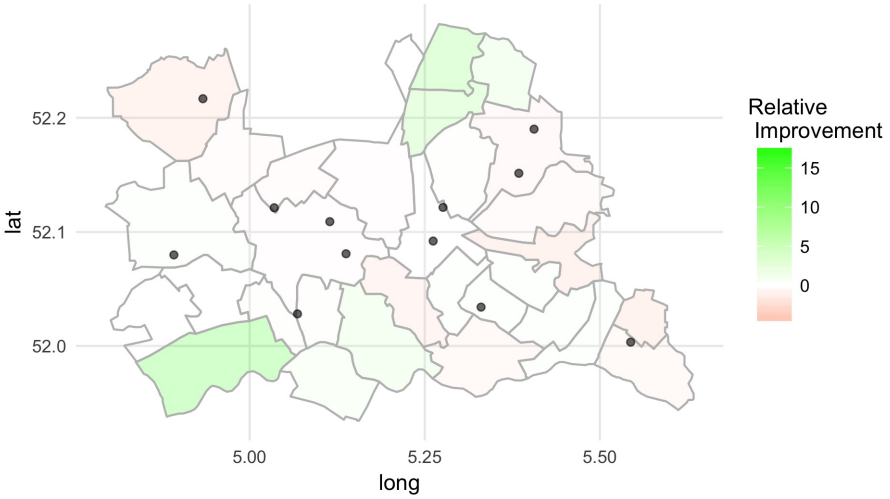


Figure 28: *Comparison of number of late arrivals in Utrecht. The base locations are indicated with black dots.*

The figure shows that there is a large relative improvement in Lopik and Eemnes. What stands out is that both of these regions are rural areas at the border of the regions without a base location. The deterioration is mostly in Veenendaal and de Ronde Venen in the north-west. Thus we see a redistribution of the late arrivals in the region, where the areas with a lower percentage on time arrivals improve.

10 Conclusion and Further Research

In this thesis we researched ambulance relocation models and developed algorithms to introduce dynamic routing to these models. We considered the following question:

How can we evaluate routes depending on coverage and travel time and how does relocating over different routes influence the performance of the model?

To answer this question, we developed multiple relocation methods. The first method was the Multiple Path Evaluation method, where we first generate a set of routes and we evaluate them using Equation (3). The second method changes the values on the arcs of the graph to depend on the distance and the coverage. We apply a shortest path algorithm on the newly obtained graph. To get a better understanding of how various parameters influence the performance of alternative routing for relocations, we ran simulations on small 10×10 grids. In the basic grid with uniform demand (see Figure 8) we saw that both methods, as well as the weighted variation of the Multiple Path Evaluation method, performed better than the shortest path method. The reason for this is that there are multiple shortest paths in the grid, and the shortest path method picks one randomly. Thus the MPE and CC methods that put more thought into picking a route perform better. We conclude that using an alternative routing policy can positively affect the performance of ambulance care.

Furthermore, the route points for the Multiple Path Evaluation method are key to its performance. We have observed that we can decrease the number of late arrivals by only considering a few extra routes. The Weighted Multiple Path Evaluation method performed best when the number of incidents is relatively high. The convex combination method performed better in most grids than any of the other methods. The improvements are especially apparent in the grid with higher demand in the corners (Figure 9) and the grid with the hole (Figure 12). These two grids have in common that the demand in the center of the grid was relatively small. The route points were also in the center of the grid, thus the MPE method took routes through the center, while the convex combination method took routes that stayed longer in the areas with higher demand.

We ran simulations for Flevoland using real life data from 2011. The data only contains incidents from 7:00 to 20:00. We observed that the number of late arrivals can be decreased by 0.17 percentage point.

For the real EMS regions that we simulated with TIFAR we saw mostly a shift in the late arrivals. In Gooi & Vechtstreek and Utrecht the ambulance arrived sooner in more remote areas, at the cost of the larger cities. This gave a large relative improvement in the remote areas with no base locations. Thus dynamic routing ensures a more even distribution of the ambulances.
In the simulation for Amsterdam most of the improvements where in Amsterdam Zuid-Oost. This is because this is a densely populated area without a base location, thus there were less late arrivals in that area and the mean response time decreased by 11.5 seconds. Thus dynamic routing can be used to cover for a potential missing base location.

A suggestion for further research is to look into the effect of the parameter γ in Equation (5). It is possible that there was too much emphasis on the beginning of the route, like with the WMPE method in Flevoland. This can lead to ambulances arriving later at their destination and potentially arrive later at an incident.

A possible extension of dynamic routing is to only use it when certain restrictions are met. For example, we can only consider dynamic routing

- during certain times of the day,
- when there are at least k ambulances available,
- for specific (O, D) pairs.

Further research is required to see what the effect of these restrictions is.

References

[1] H.H. Simons, Ambulance in-zicht 2015, Industry Report, Ambulancezorg Nederland, 2016.

[2] V. Bélanger, A. Ruiz, P. Soriano, Recent Advances in Emergency Medical Services Management, CIRRELT, 2015.

[3] T.C. van Barneveld, C.J. Jagtenberg, S. Bhulai, R.D. van der Mei, Real-Time Ambulance Relocation, Assessing real-time redeployment strategies for ambulance relocation, Submitted for publication, 2016.

[4] R. Church, C. Revelle, The Maximal Covering Location Problem, Papers of the Regional Science Association, 32:101-118, 1974.

[5] M.S. Daskin, A Maximum Expected Covering Location Model: Formulation, Properties and Heuristic Solution, Transportation Science, 17:48-70, 1983.

[6] E.W. Dijkstra, A Note on Two Problems in Connexion with Graphs, Numerische Mathematik, 1:269-271, 1959.

[7] G. Kommer, S. Zwakhals, Referentiekader spreiding en beschikbaarheid ambulancezorg 2008, 2008.

[8] M. van Buuren, K. Aardal, R. van der Mei, H. Post, Evaluating Dynamic Dispatch Strategies For Emergency Medical Services: TIFAR Simulation Toll, Proceedings of the 2012 Winter Simulation Conference, 2012.

[9] C. Jagtenberg, S. Bhulai, R.D. van der Mei, An efficient heuristic for real-time ambulance redeployment, Operations Research for Health Care, 4:27-35, 2015.

[10] P.L. van den Berg, Logistics of Emergency Response Vehicles: Facility Location, Routing, and Shift Scheduling, PhD thesis, Delft University of Technology, 2016.