EVALUATING DYNAMIC DISPATCH STRATEGIES FOR EMERGENCY MEDICAL SERVICES: TIFAR SIMULATION TOOL

Martin van Buuren

Center for Mathematics and Computer Science
Science Park 123
1098 XG Amsterdam, NETHERLANDS

Karen Aardal

Delft University of Technology
Mekelweg 4
2628 CD Delft, NETHERLANDS

Rob van der Mei

Center for Mathematics and Computer Science
Science Park 123
1098 XG Amsterdam, NETHERLANDS

Henk Post

Connexxion
Laapersveld 75
1213 VB Hilversum, NETHERLANDS

ABSTRACT

In life-threatening emergency situations, the ability of emergency medical service (EMS) providers to arrive at the emergency scene within a few minutes may make the difference between survival or death. To realize such extremely short response times at affordable cost, efficient planning of EMS systems is crucial. In this article we will discuss the Testing Interface For Ambulance Research (TIFAR) simulation tool that can be used by EMS managers and researchers to evaluate the effectiveness of different dispatch strategies. The accuracy of TIFAR is assessed by comparing the TIFAR-based performance indicators against a real EMS system in the Netherlands. The results show that TIFAR performs extremely well.

1 LITERATURE REVIEW

In this section we give a brief overview of the available literature on models for EMS systems.

Deterministic static models
In the context of EMS, deterministic static models are usually integer programming (IP) formulations for finding a static distribution of EMS vehicles over all potential base locations. The Location Set Covering Model (LSCM) by **?** is a binary integer programming (BIP) formulation that tries to minimize the number of bases that contains an EMS vehicle. This must be done in such a way that at most one EMS vehicle will be stationed at a potential base location and each demand point can be reached by an EMS vehicle within a given time r. LSCM provides a lower bound on the number of EMS vehicles needed. The Maximal Covering Location Problem (MCLP) by **?** is a BIP-formulation that tries to find a way to distribute a given number of EMS vehicles over the potential base locations such that as many demand points as possible can be reached by at least one EMS vehicle within the given time standard r. The major disadvantage of both LSCM and MCLP is that at the moment when an EMS vehicle departs to an incident location some demand points might not get reached within the given time standard, or rather said: these demand points might not be covered any more. The Backup Covering Problems (BACOP) in **?** address this by rewarding double covered demand points in the objective function. The Double Standard Model (DSM) in **?** can be seen as an extension of BACOP. This model maximizes the population covered by (at least) two EMS vehicles, and it is the first model that considers two different maximum allowed patient waiting times $r_1 < r_2$, from now on called response times, for high urgency and lower urgency calls, respectively. In case there are two types of vehicles: advanced life support (ALS) vehicles that may handle all types of calls,

and basic life support (BLS) vehicles that may only handle calls of a certain urgency level, the Tandem Equipment Allocation Model (TEAM) in **?** gives a way to model the situation. This model maximizes the population covered at least once by each type of vehicle, given the number of EMS vehicles of each type and the corresponding response times r_1 and r_2.

Probabilistic static models
Probabilistic static models are more realistic in the sense that they take the probability q into account by which an EMS vehicle is available for dispatch, independent of the status of all other EMS vehicles. The Maximum Expected Covering Location Problem (MEXCLP) in **?** gives a lower bound on the number of vehicles that is needed, like LSCM does for deterministic static models. In case one wants to assign a given number of EMS vehicles to potential base locations such that a maximal number of demand points can be reached with some predetermined probability β within a given time standard r, one can use the Maximum Availability Location Problem (MALP) in **?**. The objective of Rel-P in **?** is to impose an upper bound β_i on the probability that a call on demand point i does not receive immediate service. This model also allows an upper bound on the number of EMS vehicles at each base. The Two-Tiered Model (TTM) in **?** can be seen as the probabilistic static version of TEAM.

Dynamic relocation models
Modern research in EMS deployment is mainly focused on dynamic ambulance management (DAM). Dynamic models help to relocate idle EMS vehicles such that a maximum of calls can be reached within a time threshold. In relation to the previous models, dynamic models are not searching for a static equilibrium but rather contribute to real life relocation systems. **?** gives a dynamic extension of MEXCLP. **?** shows using an IP-formulation that for rural communities with large areas to cover, dynamic ambulance modeling can be beneficial. The Dynamic Double Standard Model (DDSM^t) in **?** allows relocations with an induced cost of available EMS vehicles between bases, and its objective function maximizes double coverage whilst relocation costs act as a penalty. The Approximate Dynamic Programming (ADP) formulation as described in **?** and **?** captures the random evolution of a system and uses simulation to try to obtain the best dispatch policy such that the total costs are minimized. Both DDSM^t and ADP search for possible relocations at the moment that an EMS vehicle becomes available or unavailable for dispatch to a new call. Recent work on the ADP approach is done by **?**.

Early work on dynamic relocations using Markov Decision Processes (MDPs) is done by **?** and **?**. **?** applied a tree-search heuristic for approaching optimal relocations to the Stockholm region in Sweden. **?** proposes a way to optimize relocations by a Markov chain model. **?** compares multiple relocation policies using MDPs.

For a more thorough discussion on how these static and dynamic models work we refer the reader to Van **?** and **?**.

Simulation models
Simulation models are extremely powerful, because of their high flexibility, e.g., with respect to modifications to the model assumptions. Moreover, they are appreciated by EMS managers due to their graphical/informative character. Also, the effect of certain decisions is easy to understand and explain when a simulation is used. Simulations provide a way to see the effect of certain decisions on a real time basis, see **?**. The effect of all previously mentioned models are tested on real life environments by using some kind of real data such as historical call record data. Two simulation packages are worth mentioning: BartSim and SIREN. BartSim is a simulation package developed in **?** for St. Johns Ambulance Service in Auckland, New Zealand, to assist during policy making. EMS vehicles have a computer-aided dispatch system that logs all call data such as travel times, treatment time and transfer time. This simulation engine

is the first of its kind that uses real data for modeling the calls. In this way, data does not have to be recorded manually as in EMS studies before; Henderson and Mason state that during the survey in **?** data was gathered manually for a period of two weeks. They also point out that using a GIS-system is relatively new in EMS planning. Auckland was described by a graph containing 2 200 vertices and 5 000 directed arcs. Some of these vertices are neither intersections nor dead ends. Leaving them out of the graph results in 765 vertices called decision vertices. Incidents are generated on the vertices by a bootstrapping procedure. The travel speed of the EMS vehicles is time dependent in BartSim. During pre-processing, the all-to-all shortest path between these 765 decision vertices is calculated with the Floyd-Warshall algorithm at times 8:00, 12:00 and 17:00, where a heuristic is used to estimate the travel speed at these times. See **?** for more information about the algorithm. The shortest path calculation between two vertices are done by a heuristic that is claimed to have a good level of accuracy. The BartSim simulator works as a discrete-event simulator. SIREN is the successor of BartSim and it simulates EMS movements as well. This software package was used for the Auckland and Melbourne areas. Nowadays the simulation package is integrated in commercial packages. Some information about SIREN can be found in **?** and **?**. SIREN uses better base locations and it considers to move bases for improved response times. SIREN includes real data and simulations yield an improvement up to 9% on the previous strategy Melbourne used. SIREN also includes stochastic travel times and non-homogeneous call generation. The simulation package can dispatch more than one vehicle to the same call. SIREN can handle up to 6 000 vertices and 14 000 arcs, and contains arc specific travel times for EMS vehicles that drive with optical and auditory signals.

2 EMS IN THE NETHERLANDS

In this section we give a brief outline of the EMS system in The Netherlands. The Netherlands is partitioned into 25 EMS regions. Each of these regions is served by a single ambulance service provider, called RAV (Dutch: regionale ambulancevoorziening). Emergency medical call centers (EMCCs) handle the incoming medical emergency calls by civilians, medical specialists and other emergency services. It also coordinates EMS vehicle movements within the RAV. For a complete overview of demographical data we refer to **?**.

Urgency levels
Each incoming call is assigned an urgency level. The Dutch EMS distinguishes three levels: A1, A2 and B.

- A1 An urgent call with an acute threat to the patient's life. Vital functions of the patient are not or rarely present, or cannot be determined through the telephone. The EMS vehicle uses optical and visual signals and tries to get to the patient as soon as possible. Examples: Heart attack, reanimation or serious traffic incidents.
- A2 The patients life is not under direct threat, but there might be serious injuries. The EMS vehicle may use optical and visual signals if the EMS personnel has discussed this with the EMCC, but this only happens on rare occasions. Examples: A broken leg or a general practitioner asks for transportation to a hospital.
- B A call without urgency A1 or A2 in which the patient must be transported within a given predetermined time interval. A typical B call exists of transferring a seriously ill person from one hospital to another, because this hospital is specialized in the patient's condition. When a seriously ill person receives a scheduled transport from an EMS vehicle to his or her home, it will be classified as a B call as well.

A major difference between calls that are labelled with urgencies A1 and A2 on one hand and calls with urgency B on the other, is that calls with A1 or A2 urgency are not known beforehand, whilst calls with urgency B can be planned in advance.

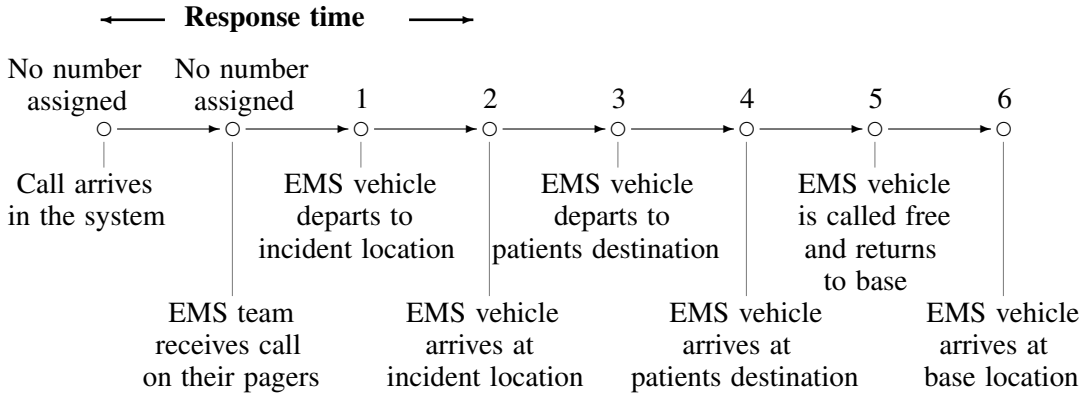

Figure 1: Overview of system statuses.

Monitoring the status of calls
For each incoming call, status information is time-stamped and logged, see Figure 1. The *response time* is defined as the length of the time interval between the moment that the call center receives the phone call until the moment the EMS vehicle arrives at the incident location. The maximum response time of calls with urgency A1 is 15 minutes, and for calls with urgency A2 30 minutes. A key performance indicator is the fraction of calls that meets these maximum response-time thresholds. For B-calls there is no maximum response time defined, since they are usually planned in advance.

Classification of calls
Each call is classified as *declarable*, *EHGV* or *loss*. A call is classified as *declarable* if transportation is required: either a patient is brought to a hospital or to his or her home. An *EHGV* call (Dutch: Eerste hulp, geen vervoer) is a type of call where the EMS team can provide care locally, and the patient does not require transportation to a hospital. This status is determined by EMS personnel at the incident location. For EHGV calls, mobilophone status 5 will immediate follow after status 2 (see Figure 1 for a systematic overview of all statuses that a call can have). Sometimes an EMS team arrives at the scene and no patient is present. This might be the result of a patient who left after the call was made or a prank call to the emergency number. This type of call can be classified as a *loss* call. In case a patient in a hospital is not ready for transportation at the moment the EMS vehicle arrives, the call will be classified as a loss call as well.

3 THE TESTING INTERFACE FOR AMBULANCE RESEARCH (TIFAR) SIMULATION PACKAGE

In this section we give an outline of how TIFAR works. TIFAR is programmed in ansi C++, making it easily transferable to different platforms and easy to perform maintenance. Essentially, the program is based on three connected loops: (1) the main loop, (2) a renewal loop, and (3) an incident generator loop. Also, TIFAR can run in two modes: *visual mode* or *speed simulation mode*. The only two ways that these two modes differ are the moments on which they set the next time stamp where the (new) state of the system gets calculated, and how they perform output. The *visual mode* uses a GUI for output and the internal computer clock to determine the next time in which the current system state gets calculated. In this way one can see the EMS movements at real time, or with a speed factor. The *speed simulation mode* is a discrete event simulator that holds an ordered list with the end times of all currently ongoing events. For example: At the moment when an EMS vehicle departs, we can calculate the moment when the vehicle arrives. This gives a new end time. An end time enters the list when a call enters the system, when a vehicle departs from a location or when a vehicle arrives at a location. The next time at which the state of the system must be calculated then equals the first one in this ordered list. When a predetermined

Figure 2: Illustration of Amsterdam region with EMS bases and hospitals.

end time has passed, speed simulation mode terminates the program and displays statistics in the terminal. Speed simulation mode is faster than visual mode, because the renewal loop only gets called at necessary time stamps. For each of the two modes there is a separate main loop included in the program.

Call generation
The incident generator loop generates calls with the following properties:

1. The model time when the incident occurs, called the start time.
2. The location on the map where the incident occurs, called the origin.
3. The treatment time, i.e., the time the EMS personnel must spend at the origin.
4. The transfer time, i.e., the time the EMS personnel must spend at the hospital to transfer the patient into hospital care. Only if the call is not EHGV or loss.
5. The urgency of the call, being A1 or A2.
6. Whether the call is EHGV or loss or not. Recall that EHGV means that the patient does not require transportation to the hospital.

There are two ways to generate calls: one can choose to give each demand point its own distribution, or one can choose to have one distribution for the moment when a call in the RAV occurs and a separate distribution to determine the origin of a call within the RAV. TIFAR makes use of the latter. Calls are generated according to a Poisson process.

For ease of the model, it is assumed that each call is handled by exactly one EMS vehicle; note that this assumption can easily be relaxed. There are a couple of possible choices for generating the call's origin:

- *Using RD-coordinates:* The Netherlands has its own cartesian coordinate system called the "rijksdriehoekscoördinaten", where the unit is 1 meter. Every location in the country has an x- and y-coordinate. When knowing the border of the RAV in RD-coordinates, one can generate a call on a random position uniformly distributed within the RAV. The main advantage is that an incident can happen on every location within the RAV, even on water. One still has to keep in mind that this location must be mapped upon the road network which might lead to major granularity errors. The main disadvantage is that the population density is not considered in this distribution.

- *Using postal codes:* Each address with a mail box has one postal code assigned. Such a postal code consists of four digits and two letters, for example 1011 AA. There are buildings with the same postal codes, i.e., a part of the same street can share a postal code. However, the combination of a postal code and house number forms an unique combination. The four digits are forming the neighborhood, whilst the two letters specify the location within this neighborhood. People involved in route planning in The Netherlands therefore make the distinction between these so called 6PP- and 4PP-postal codes. A rule of thumb states that the cumulative amount of mail for all houses with the same 6PP postal code is the quantity that a postman can hold in his hands. One can map a postal code onto RD-coordinates. The main advantage with this way of generating call origins is that the population density is included, as one can estimate that the number of people located on each postal code is almost equal. One must however keep in mind that streets with only one mail box and not many inhabitants have their own postal code, while on the other hand nursing homes have one postal code and a high potential that an EMS vehicle should be called. The disadvantage is that forests, water and highways do not have postal codes because there are no mail boxes. Still, incidents can happen at these places.

- *Using bootstrapping:* Using EMS data from the past, one knows where an incident has happened, and thus where an incident can happen again. In The Netherlands, all EMS data from 2007 until now is stored. Incidents in the model can be generated using a bootstrap procedure from this data. The main advantage is that places where many incidents happen in reality, will be represented very accurately. The disadvantage is that there are a lot of places where no incident has happened before, but where new incidents might occur. Furthermore, new neighborhoods are not included by this way of incident generation. The use of historical EMS data can also involve privacy concerns, though if one can afford to lose some precision, the location can be made anonymous by mapping it onto the 4PP postal code.

- *Using trace driven*: When using the EMS's call record database, we know where and when every incident has happened. Using a simulation tool, we can simulate incidents with exactly the same characteristics and see how an alternative dispatch strategy performs. Comparing these statistics with the actual performance measured by EMS yields an excellent evaluation tool. An advantage is that it provides a good comparison between the simulated and actual data. However, we have the same privacy issues as with bootstrapping procedure.

TIFAR uses the postal codes method to simulate the incidents.

Dispatching policy

When a call occurs, we have to assign an EMS vehicle to the call. We will now describe how the call handling works. TIFAR has a queue that contains all calls. We assign EMS vehicles to calls *in order of*

priority: first we assign vehicles to calls with urgency A1 on a first-come-first-served basis, and if all calls with urgency A1 are served we start assigning vehicles to calls with urgency A2 on the same basis. The EMCC always assigns the nearest (in time) available EMS vehicle to a call. Note that a similar approach has been applied in **?**. Once an EMS vehicle has been assigned to a call, the call will not be handled by another EMS vehicle. Not even when the other EMS vehicle gets available for dispatch while being closer to the origin than the already assigned vehicle. This assumption is based on the fact that in practice, it rarely occurs that an EMS vehicle gets assigned to another call.

When driving to a call with urgency A1 we assume the vehicle has auditory and visual signals, and when driving to a call with urgency A2 we assume that the vehicle drives without them. Sometimes an EMS vehicle drives with these signals to a call with urgency A2, but since this rarely occurs we have not implemented this in TIFAR. When driving to a hospital, we assume that the EMS vehicle goes with A2 speeds. The speed at which a vehicle moves depends on the type of road it is on (e.g., highway, road in a city, road in the countryside) and whether or not auditory and visual signals are used. In reality, an EMS vehicle may drive with signals to an hospital.

When a call is declarable, the patient will be brought to the nearest hospital. We assume that this corresponds well with the real situation, although there are cases in which a specialized hospital should be chosen instead of the nearest one, see **?**. When an EMS vehicle departs from a hospital, it will head to the nearest base unless a relocation rule decides otherwise.

Relocation policies are confidential, and therefore we use a heuristic. We call an EMS vehicle *involved* with a base if one of the following two conditions hold:

1. The EMS vehicle is waiting at the base until a new call arrives or a relocation instruction sends it to another base location.
2. The EMS vehicle is driving to the base and is available for dispatch or relocation.

If a relocation rule states that an EMS vehicle must be sent from base j_1 to base j_2, TIFAR takes the set of all EMS vehicles involved with j_1, calculates each of their distances in time to j_2 and sends the nearest one to j_2. When there are zero EMS vehicles involved with j_1 the EMCC of TIFAR will not relocate an EMS vehicle from j_1. When relocating, we make a distinction between three regions (namely, North, Center and South, as can be seen in Figure 2).

If an EMS vehicle is not asked to relocate, it gives mobilophone status 5 (see Figure 1) and goes to the nearest base to wait until it gets assigned to a new call, or until a relocation rule at a later time will send it to another base. This strategy has the disadvantage that EMS vehicles get drawn to the center. Let us illustrate that. In the North, there are two hospitals. When an EMS vehicle delivers a patient from an incident scene in certain locations in the Northern region to the BovenIJ hospital in Amsterdam North. The vehicle is now drawn to the center. In the south, there is a region where the nearest base is Amstelveen, and the nearest hospital is VU Medical Center (VUmc). After delivering a patient to the VUmc hospital the EMS vehicle will find that a base in the center is nearest, and will go to there to wait until a new call arrives. When a vehicle enters the center region there are two ways to leave it:

1. By a relocation rule.
2. When an incident occurs in another region and that region has no available EMS vehicles. If this is an A1 call it might not be served within the maximum allowed time due to the travel distance, and we want that to happen.

These observations again stress the importance of good relocation rules. Relocations are necessary to keep available EMS vehicles distributed well over the RAV.

TIFAR makes use of graphical user interface and route planning software. Let us mention the most important services. The map server has as input the RD-coordinates of the bottom left corner, output image dimensions and a scale factor, and returns the corresponding map in JPG format. The position planner has as input a postal code and house number, and returns the corresponding RD-coordinates. This is used to determine the location of the predetermined EMS bases and hospitals, and to display them on the map. The geo-projector maps RD-coordinates to the 'nearest' road. The geo-projector is used to determine the location of an available EMS vehicle that is not at a base.

The AB planner returns the shortest route and the corresponding time in seconds between two points in RD-coordinates. As input one gives the start point and the end point. This input can be an address, a postal code, or a geo-projection. TIFAR uses the latter two options. The EMS vehicles' travel speeds are obtained from historical data. The shortest path is calculated with the bidirectional A^*-algorithm, see **?**. The matrix planner has two modes: one-to-many and many-to-one. The many-to-one determines the distance and travel time from many start locations to only one end location. This is used to determine the nearest available EMS vehicle: from many vehicles to one call's origin. The one-to-many determines the distance and travel time from one location to many locations. This is used to determine the nearest hospital from a call location, or to determine the nearest base from a hospital when an EMS vehicle returns to a base. Input can be both a postal code or geo-projection. Tele Atlas data is used for road network information and geographic information.

4 RESULTS

We have run a variety of simulations with realistic scenarios based on the parameters obtained from historical data, that is mostly taken from the annual Dutch ambulance statistics overview **?**. Other detailed information about base locations, staffing levels, travel-time models and relocation policies that are implemented in an operational RAV in The Netherlands is partly taken from the national ambulance distribution plan **?** and partly from confidential resources.

The results show that the simulation-based performance of high-priority calls is extremely close to the actually realized performance by the RAV (the actual numbers are confidential). This shows that TIFAR is able to answer what-if scenarios for different configurations, which provides RAVs with a powerful tool to enhance the efficiency of their daily operation.

5 CONCLUSION

TIFAR is a decision support tool that is good in predicting high urgency calls. Our simulation results for calls with A1 urgency are very close to the actual statistics. TIFAR has several useful features: (1) it can handle large amounts of vertices (demand points, base locations and hospitals). Over 32 000 vertices are supported, which leads to very small granularity errors. This is much more than earlier models. (2) The graphical user interface can display effects of decisions in real time. By default simulations go by factor 60 (1 minute of simulation time equals one hour of ambulance movements), but faster settings are possible. (3) It is easy to implement relocation rules. (4) It is easy to vary the number of EMS vehicles. (5) It gives a clear overview of important statistics.

6 ACKNOWLEDGEMENTS

The authors would like to thank Simon Visser, Maya Ronday, Suzanne van der Leeuw and Rob Bosman for their valuable input.

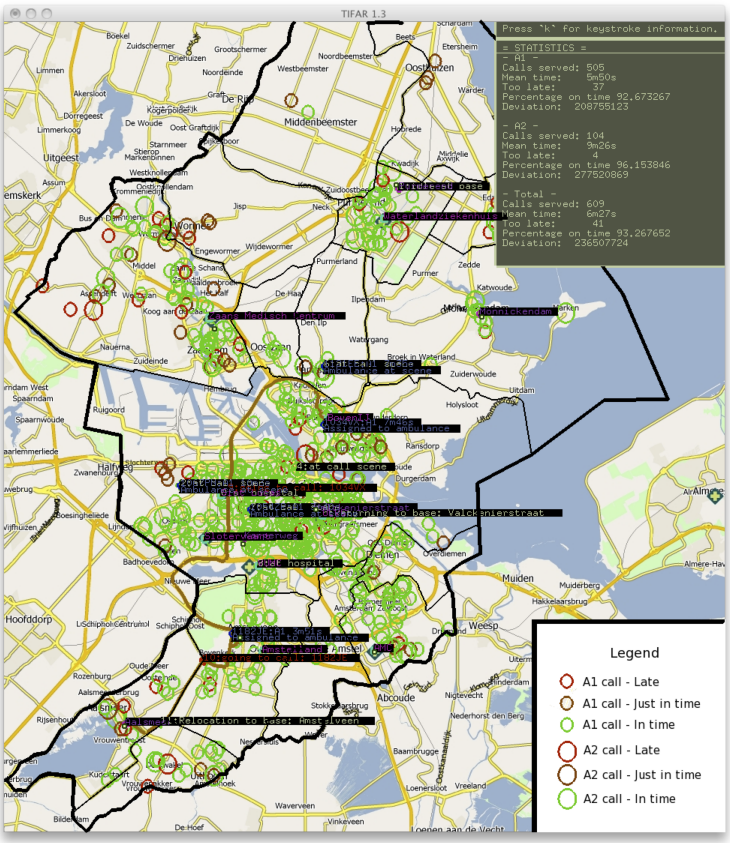

Figure 3: Illustration of the GUI of TIFAR.

AUTHOR BIOGRAPHIES

MARTIN VAN BUUREN is a scientific programmer and a Ph.D. candidate at Centrum Wiskunde & Informatica (CWI) and VU University Amsterdam. He develops strategic decision making software for emergency services. Having finished a mathematical studies at TU Delft, he now participates in the REPRO ambulance logistics project.

KAREN AARDAL is professor of Optimization at Delft University of Technology. She has also held positions in the UK and the USA, and at several other universities in The Netherlands. Her main research interest is integer programming and combinatorial optimization, in particular problems related to facility location. She has served on the Council and as Executive Committee Chair of the Mathematical Optimization Society, and on the board of INFORMS Computing Society. She is currently area editor of INFORMS Journal on Computing, editor of EURO Computational Optimization, and associate editor of Mathematical Programming B, Networks, and RAIRO – Operations Research.

ROB VAN DER MEI is professor in Operations Research at the VU University Amsterdam, and is heading the research cluster Probability, Networks and Algorithms, consisting of some 70 researchers, at the Centrum Wiskunde & Informatica (CWI). His research is mainly focused on the development and analysis of mathematical models for real-life problems related to capacity planning and performance analysis in many application areas. He has been involved in countless research projects with industrial partners, and is co-author of over 100 papers in the field.

HENK POST is a software architect at Connexxion, a large public transport company in The Netherlands. He develops software that is used to solve large-scale routing problems where buses, taxis and ambulances are involved. He has a special interest in the 'shortest path problem'. He currently focuses on all these subjects in his PhD thesis that he is working on at Delft University of Technology.