Noname manuscript No.
(will be inserted by the editor)

Models for Ambulance Planning on the Strategic and the Tactical Level

J. Theresia van Essen · Johann L. Hurink ·
Stefan Nickel · Melanie Reuter

Received: date / Accepted: date

Abstract Ambulance planning involves decisions to be made on different levels. The decision for choosing base locations is usually made for a very long time (strategic level), but the number and location of used ambulances can be changed within a shorter time period (tactical level). We present possible formulations for the planning problems on these two levels and discuss solution approaches that solve both levels either simultaneously or separately. The models are set up such that different types of coverage constraints can be incorporated. Therefore, the models and approaches can be applied to different emergency medical services systems occurring all over the world. The approaches are tested on data based on the situation in the Netherlands and compared based on computation time and solution quality. The results show that the solution approach that solves both levels separately performs better when considering minimizing the number of bases. However, the solution approach that solves both levels simultaneously performs better when considering minimizing the number of ambulances. In addition, with the latter solution approach it is easier to make a good trade-off between minimizing the number of bases and ambulances because it considers a weighted objective function. However, the computation time of this approach increases exponentially with the input size whereas the computation time of the approach that solves both levels separately follows a more linear trend.

Keywords Ambulance Planning · Stochastic Programming · Integer Programming · Local Search · Simulation

J. Theresia van Essen · Johann L. Hurink
Center of Healthcare Operations Improvement and Research (CHOIR),
University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
E-mail: j.t.vanessen@utwente.nl

Stefan Nickel · Melanie Reuter
Discrete Optimization and Logistics, Karlsruhe Institute of Technology, Englerstr. 11, Building 11.40 2. OG, D-76128 Karlsruhe, Germany

1 Introduction

Emergency Medical Service (EMS) systems as they exist in the U.S., Canada or European countries like the Netherlands are very complex. When planning such a system, there are lots of different aspects that have to be considered and many questions have to be answered, for example legal regulations, regional distinctions or geographical characteristics. Usually, the problem of planning the EMS system can be divided into smaller subproblems as location planning, dispatching and so forth which are generally easier to solve than the overall problem. However, the subproblems depend on one another such that the solution of one subproblem forms the basis for solving the next one. Combining all the partial solutions then defines the planning of the EMS system. One of the subproblems within EMS systems is the question where to locate bases and ambulances throughout the considered region. This can either be done with a prefixed number of available ambulances or the location decisions can be made simultaneously while determining the number of needed vehicles. The problem of locating ambulances and ambulance bases can be divided into three phases: the strategic, the tactical and operational level. At the strategic level the locations of the ambulance bases are determined while considering constraints on the coverage. In the next step, the tactical level, the explicit number of ambulances needed per base to fulfill all demand is specified. At the operational level, the allocation of ambulances to emergencies and relocation of ambulances must be carried out in real-time. In this paper, we focus on the first two levels as the operational level differs significantly from the other two and should be covered separately.

In this work, we first present a solution approach for solving the strategic and tactical planning problem simultaneously. This approach is mainly used as a benchmark to be able to evaluate solutions and computation times, as we suggest to split the problem into the two levels mentioned above. The reason for this splitting is that solving the strategic and tactical level simultaneously leads to a complex problem with long computing times. However, as the location of (larger) bases is often fixed for years but the number and location of ambulances can change each year, it seems to be a logic decision to solve the two levels separately. Nevertheless, when a replanning of an EMS system is wanted and the location of bases together with number of ambulances should be determined simultaneously, the proposed simultaneous approach can be applied to tackle the problem.

The chosen solution approach for solving both levels at once is a stochastic programming formulation. The input instances for this formulation are quite large, and therefore, the problem has to be simplified more than is the case when we solve both levels separately. However, solving the problem in two stages may result in a suboptimal solution. Therefore, we compare the solutions of the stochastic programming formulation with the solutions of the approach that solves the problem in two stages.

At the strategic level, we determine the locations of the ambulance bases. When locating these ambulance bases, we also have to take into account the location of the emergency departments as not only the driving time between the patient location and the ambulance location should be minimized, but also the total driving time from an ambulance location to the hospital location via the patient location might be of interest. Therefore, when the patient location is far from the hospital location, it might be necessary to locate an ambulance base close to the patient to minimize this total driving time.

The number of ambulances needed at each base is a decision at the tactical level as this may change regularly according to changes in demand. The decision at the tactical level highly depends on the time-dependent demand and travel time, and therefore, we propose to use simulation to solve this problem. We start with an initial number of ambulances based on average demand and travel time. This number is adapted iteratively by simulating the current situation and suggesting moves to improve the current solution. In other words, we incorporate simulation in a local search approach. In the simulation, we also consider the covering constraints as mentioned in the model for the strategic level.

In contrast to many of the probabilistic approaches presented in literature like the maximum expected covering location problem introduced by Daskin [9], the maximum availability location problem introduced by ReVelle and Hogan [29] or the stochastic formulation by Beraldi et al. [3], we chose scenarios to model the uncertainty instead of defining busy fractions for the ambulances or making the calls occurring randomly. In addition, the models presented in this paper have a different way of modeling the coverage constraints. Based on the idea presented by Gendreau et al. [12] that simple coverage might not be sufficient, coverage constraints are modeled generic such that different levels of coverage can be included.

Concluding, the contribution of this paper is as follows: we decompose the ambulance planning problem into a strategic and tactical level, present a formal description of the problem at each level, and introduce and compare methods for solving both levels separately and in an integrated way.

The paper is structured as follows: Section 2 covers a literature review. In Section 3, we present the problems at the two stages, give corresponding formulations and discuss possible shortcomings. The proposed solution approaches are presented in Section 4. First, the overall approach and second, the approaches for the two separated levels are given. The results of comparing the models are discussed in Section 5. The paper closes with a summary and an outlook in Section 6.

2 Literature Review

In ambulance location planning, there already exists a large variety of literature. We do not aim to review all developed approaches, and therefore, we only discuss the approaches most relevant for our research. For a complete overview of related literature, the reader is referred to surveys as they can be found in Marianov and ReVelle [23], Owen and Daskin [27], Brotcorne and Laporte [6], Galvao et al. [11], and Li et al. [22].

Concerning the modeling of the coverage constraints, several formulations are discussed in literature which can be used for the problem on the strategic level. The first emergency base location covering model in literature is the location set covering model (LSCM) which was introduced by Toregas et al. [33]. Its objective is to find the minimum number of ambulance bases needed to cover all demand points. Several other covering models such as the maximal cover location problem (MCLP) introduced by Church and ReVelle [8], the double standard model (DSM) introduced by Gendreau et al. [12], the maximum expected covering location problem (MEXCLP) introduced by Daskin [9], and the maximum availability location problem (MALP) introduced by ReVelle and Hogan [29] all assume a fixed amount of available bases, and thereby can only indirectly be used to minimize the number of bases. In addition, the DSM and MEXCLP models already assign several vehicles to each base to guarantee the coverage, i.e. solve the strategic and tactical problem at the same time. An interesting covering model is the gradual coverage model introduced by Karasakal and Karasakal [19]. This model uses a sigmoid function to model the gradual decline of coverage along with an increase of the distance, and thereby relaxes the ‘all or nothing’ assumption when a fixed radius is specified. Berman et al. [5] proposed the cooperative coverage model which assumes that a demand point is covered when the total received ‘signal’ from several bases exceeds a certain threshold. This means that when a certain demand point lies further away from a base, a second base may be needed to fulfill this threshold.

All the models mentioned above are deterministic and static. The first probabilistic approach was presented by Chapman and White [7] in 1974. It was a probabilistic set covering model in which servers were not always available. Aly and White [1] published a formulation for the probabilistic set covering problem together with a variation of it in 1978. They assumed the location of incidents to be random variables. As mentioned above, Daskin [9] proposed in 1983 the maximum expected location covering problem (MEXCLP). There, he included the idea that an ambulance is busy for a fraction of time. He assumed that the number of ambulances that have been placed on the network was given. As an extension of the MALP, Marianov and ReVelle [24] developed the Queueing MALP or Q-MALP in 1996. They used results from queueing theory to relax the assumption that the busyness probabilities of different servers are independent. In addition, travel times were also considered to be random variables.

Among other probabilistic approaches that for example use reliability constraints and busy fractions for servers as done by ReVelle and Hogan [28], there are two main approaches for including stochasticity into the ambulance location problem, namely hypercube queueing models and stochastic programming. The first hypercube queueing model was introduced by Larson [21] in 1974. Based on that, different variations can be found for example in [13], [17], [18], [30], or [32]. Beraldi et al. [3] present a stochastic integer problem formulation under probabilistic constraints (SIPC) that determines where service sites must be located and how many emergency vehicles must be assigned to each site while randomness in the demand of emergency services is assumed. They give a deterministic equivalent formulation of the introduced constraints using the so-called p-efficient points of a joint probability distribution function. Beraldi and Bruni [2] propose a stochastic programming model under probabilistic constraints as a two-stage approach. They relax the assumption of server independence and assume randomness in the emergency requests instead of in the server availability. Noyan [26] developed two types of stochastic optimization models involving alternate risk measures, the first one including integrated chance constraints (ICC) and the second one incorporating ICCs and a stochastic dominance constraint. He modeled the random demands using the scenario approach and relaxed the assumptions that the service providers operate independently and that the demand sites are independent of each other. The stochastic program presented in Section 4 of this paper is based on the formulations by Beraldi and Bruni [2] and Noyan [26]. In contrast to Beraldi and Bruni [2] we do not enforce that each demand location has to be served by only one ambulance base. In addition, we extend the general two-stage formulation Noyan described [26] by generic coverage constraints.

There is also quite some literature considering simulation for ambulance planning because determining the number of needed ambulances highly depends on the time-dependent demand and travel time. Several of the simulation studies, e.g. [10], [14], [15], and [16] use simulation to evaluate location policies determined by optimization models. Swoveland et al. [31] use simulation in combination with a form of branch-and-bound to determine the location of ambulances. Berlin and Liebman [4] use simulation to evaluate location policies, but also to determine the number of ambulances needed. The assigned number of ambulances per base is set sufficiently high to serve all requests, and after the simulation it is determined how many ambulances are needed to guarantee availability in, for example, 95% of the time. This idea is incorporated in the simulation approach developed in this paper to determine the number of needed ambulances at the tactical level. Zaki et al. [34] use simulation to determine the effect of using an ambulance from a different region on the average response time and overall coverage. The simulation shows that the average response time increases, but also the overall coverage increases.

To the best of our knowledge, existing literature lacks an explicit integration of the different levels for ambulance planning in EMS systems together with a comparison of separated and combined solution approaches (for the strategic and the tactical level). In addition, existing literature tends to present formulations (and algorithms) specified only for certain EMS systems. General and generic formulations and approaches are needed to enable a comparison between different systems, for example. This work is supposed to start filling these gaps.

3 Problem Formulation

In this section, formulations for the planning problems at the strategic and the tactical level are presented. We first discuss the problem of locating ambulance bases and after that, the problem of determining the number of ambulances per base is introduced.

3.1 Strategic Level

When an accident happens, an ambulance is sent to the location of the accident to provide first aid and to transport the patient to the hospital. To limit the risk of medical complications, the patient should arrive at the hospital as soon as possible. Because the locations of hospitals are fixed, the time until a patient arrives at the hospital can only be influenced by the time it takes for an ambulance to arrive at the patient's location as the treatment time needed at the scene cannot be influenced. In addition, the sooner an ambulance arrives at the accident location, the sooner the medical treatment can start.

In general, ambulances are stationed at ambulance bases. An ambulance drives from its base location to the patient's location and after transporting the patient to the hospital, the ambulance returns to its base. Most countries have some coverage requirements for the locations of the ambulance bases. For example, time limits can be given for the maximum time an ambulance is allowed to need to arrive at the patient's location or for the time period till a patient arrives at the hospital for further treatment.

At the strategic level, the aim is to minimize the number of ambulance base locations while fulfilling the given coverage requirements as show in Figure 1. To formalize this problem, we use the following notation of which an overview can be found in the appendix. The set of demand locations, i.e., the set of locations where an accident might happen, is given by set I. The demand for each demand location $i \in I$ is given by d_i and can either be stochastic or deterministic. For now, we assume that the demand is fixed and given. The set of potential base locations is given by set J and from these potential base locations, we have to select a subset such that all coverage requirements are fulfilled. We model this by introducing a binary variable X_j which is one when ambulance base location $j \in J$ is selected and zero otherwise.

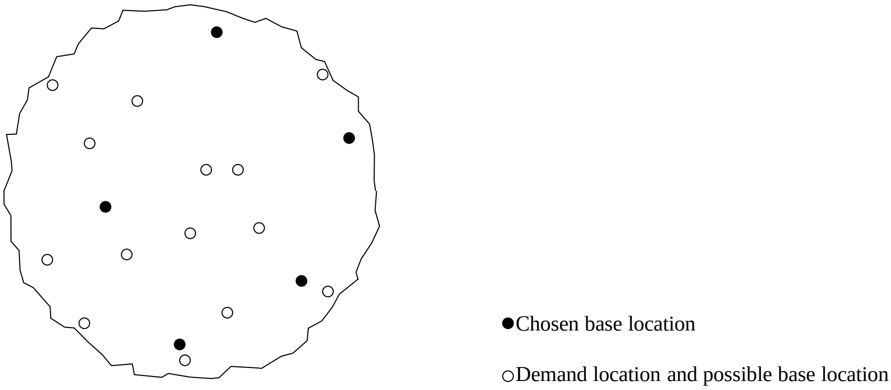

Fig. 1 The locations of bases are determined on the strategic level

To specify the coverage requirements, for each demand location $i \in I$, we specify a subset of base locations that lie within the range of the considered coverage requirement. As the coverage mostly only depends on the driving time between the demand and ambulance base location, these subsets can easily be determined beforehand. When, for example, the driving time from the base location via a demand location to the hospital location is limited by a coverage requirement, the coverage for a certain demand location only depends on the driving time between the demand and base location. Because the driving time from the demand location to the hospital location is fixed, this can easily be subtracted from the total driving time. Often, more than only one coverage requirement is considered. For this, we introduce a set K of coverage requirements. For example, there might be different maximal allowed driving times for varying severity of the incidents, resulting in several different coverage requirements, or double coverage requirements as proposed by Gendreau et al. [12] might be given. The subset of base locations which fulfill coverage requirement $k \in K$ for demand location $i \in I$ is denoted by $J_{ki} \subseteq J$. To determine whether demand location $i \in I$ is covered according to coverage requirement $k \in K$, we introduce binary variables Y_{ki} that take value one if for demand location $i \in I$ the coverage constraint $k \in K$ is fulfilled and zero otherwise. The following constraint ensures that these binary variables Y_{ki} take value zero if the coverage requirement is not fulfilled:

$$\sum_{j \in J_{ki}} X_j \geq Y_{ki}, \ \forall i \in I, \ \forall k \in K. \tag{1}$$

In general, for most coverage constraints $k \in K$, not 100% of the demand but only a (large) fraction of the demand has to be covered. In addition, this fraction could be specified for all locations $i \in I$ together (the entire country) or specific subsets of locations (regions). The latter ensures that all regions in a country have the same coverage, while in the first situation, one region could be covered less than another region. Furthermore, different coverage requirements may focus on different levels. For example, a first coverage requirement should hold for the entire country, a second coverage requirement for each state in a country, and a third coverage requirement for each municipality in a country. Therefore, we introduce for each coverage requirement $k \in K$ a set of regions R_k for which the coverage requirement must hold. More precisely, for each coverage requirement $k \in K$, we partition the set of demand locations I into $|R_k|$ subsets denoted by I_{kr} with $r \in R_k$, i.e., for each region $r \in R_k$ we specify the demand locations $i \in I$ that lie within this region. As the fraction of demand covered according to coverage requirement $k \in K$ does not have to be the same for each region $r \in R$, we introduce α_{kr} as the fraction of demand to be covered in region $r \in R$ according to coverage requirement $k \in K$. This fraction can, for example, be smaller for a region in which certain demand locations are hard to reach because they lie on a mountain top or on an island. In addition, in case that coverage requirements differ for urban and rural areas (e.g., by law), this also can be modeled by the introduced constraints. More formally, the following constraint ensures that a fraction α_{kr} of the demand in region $r \in R_k$ is covered according to coverage requirement $k \in K$:

$$\sum_{i \in I_{kr}} d_i Y_{ki} \geq \alpha_{kr} \sum_{i \in I_{kr}} d_i, \ \forall k \in K, \ \forall r \in R_k. \tag{2}$$

Note that constraint (2) may become irrelevant for some coverage requirements $k \in K$ as it may be dominated by one of the other coverage requirements. To illustrate this, let us consider two coverage requirements $k_1, k_2 \in K$. When $J_{k_1 i} \subseteq J_{k_2 i}$ for all $i \in I$ and $\alpha_{k_1 r} \geq \alpha_{k_2 r}$, then coverage requirement k_2 is dominated by coverage requirement k_1. However, for most practical instances this situation will not occur for all demand locations $i \in I$, but only for a subset of the demand locations, because the base locations included in J_{ki} may e.g. depend on varying driving times per demand location.

As the aim at the strategic level is to minimize the number of chosen ambulance base locations, our objective function can be formulated as follows:

$$\min \sum_j X_j. \tag{3}$$

Summarizing, the problem for locating bases at the strategic level looks as follows:

$$
\begin{aligned}
\min & \sum_j X_j && && (4)\\
\text{s.t.} & \sum_{j\in J_{ki}} X_j \geq Y_{ki} && \forall i \in I,\ \forall k \in K\\
& \sum_{i\in I_{kr}} d_i Y_{ki} \geq \alpha_{kr} \sum_{i\in I_{kr}} d_i && \forall k \in K,\ \forall r \in R_k\\
& X_j, Y_{ki} \in \{0,1\} && \forall i \in I,\ \forall j \in J,\ \forall k \in K
\end{aligned}
$$

It is easy to see that in practice there are some shortcomings of the model. Note that if demand location $i \in I$ is covered by one of the selected ambulance base locations, this does not necessarily mean that this demand location is always covered in practice, because the coverage also depends on the ambulance availability at this base location. In addition, it may happen that the nearest hospital to a demand location has insufficient capacity, and thus, the patient has to be transported to another hospital. In other words, the strategic level does not consider the varying ambulance and hospital capacity. Therefore, we have to extend our model to also include these aspects within the tactical level.

3.2 Tactical Level

In constraint (2) it is ensured that the full demand d_i of a location $i \in I$ is covered if at least one base location $j \in J$ with $j \in J_{ki}$ is opened. Thus, it is somehow assumed that always enough ambulances are available at the opened base locations. However, as this is in general not the case, we determine at the tactical level the number of ambulances needed to fulfill the considered coverage requirements as shown in Figure 2. Note that also the hospital capacity can be a restriction, but as this capacity depends on much more than only the emergency calls, we do not include this at this level.

Before we introduce the model to determine the number of needed ambulances, we first clarify the situation at the tactical level by providing the following sketch of the handling of an emergency call. When an emergency call occurs, we first might have to wait until one of the ambulances becomes available. When an ambulance is or becomes available, it drives from the ambulance base location to the location of the emergency call. Note that for the tactical level we assume that ambulances always start at their bases when driving to an emergency and that they always go back there afterwards. Of course, this assumption is only valid for the tactical level and would not be applied on the operational level where the actual location of the ambulance has to be taken into account. The driving time from the base location to the call location depends on the traffic intensity at that moment and has influence on the coverage.

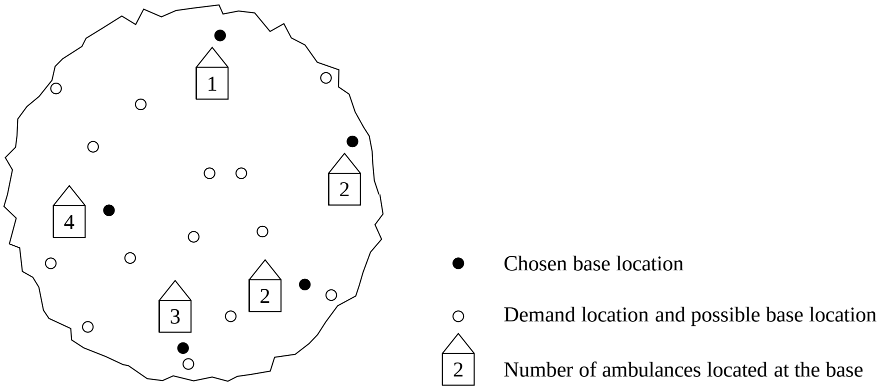
1
2
4
2
3
● Chosen base location
○ Demand location and possible base location
2
Number of ambulances located at the base

Fig. 2 The numbers of ambulances per base location are determined on the tactical level

An example for a coverage requirement could be that an ambulance should arrive at the emergency location within 15 minutes after the emergency call occurred. This means that the call is covered when the waiting time plus the stochastic driving time from base to call location is less than or equal to 15 minutes. When the ambulance arrives at the emergency call location, the patient is first treated at the scene and, if necessary, placed in the ambulance. This time between the arrival and departure of the ambulance at the emergency call location is called the treatment time and depends on the injuries of the patient. After the treatment time, the patient is transported to the hospital if necessary. This driving time also depends on the traffic intensity at that moment. Another coverage requirement could be that a patient should arrive at the hospital within 45 minutes after the call occurred. This means that an emergency call is covered according to this coverage requirement when the waiting time plus the driving time from base to call location plus the treatment time and the driving time from call to hospital location is less than or equal to 45 minutes. Note that this time includes much uncertainty. When the ambulance arrives at the hospital, the patient is transferred to the hospital personnel. This also takes some time which differs per patient and injury. After the transfer, the ambulance can return to its base location and becomes available, possibly after a cleaning process, for the next emergency call. Figure 3 visualizes this example.

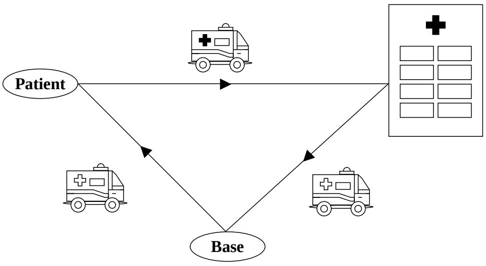

Fig. 3 A typical workflow on the tactical level

To model the sketched situation at the tactical level, we first have to generate incoming emergency calls. The arrivals of emergency calls can be modeled in several ways, but for now we assume that the arriving emergency calls are known beforehand. We denote this set of emergency calls by E, and each emergency call $e \in E$ has an arrival time given by t_e and a location denoted by l_e. Recall that the location l_e of emergency call $e \in E$ is an element of the set of demand locations I. Then, each emergency call should be served by one of the base locations chosen at the strategic level. Note that for the strategic level, set J represented the set of potential ambulance base locations, while for the tactical level, set J denotes the set of selected ambulance base locations. To assign emergency call $e \in E$ to one of the available base locations $j \in J$, we introduce binary variables Z_{ej} which are one when emergency call $e \in E$ is served by ambulance base $j \in J$. The following constraints ensure that each emergency call $e \in E$ is assigned to exactly one ambulance base location:

$$\sum_{j \in J} Z_{ej} = 1, \; \forall e \in E. \tag{5}$$

An ambulance base location can only be assigned to an emergency call when there is an ambulance available to serve it. For this, it sometimes might be necessary to let a patient wait until an ambulance becomes available. This waiting time for emergency call $e \in E$ is denoted by a variable W_e. This waiting time W_e has to be chosen such that an ambulance becomes available at base location $j \in J$ after this waiting time. The ambulance then has to drive to the emergency call location, pick up the patient, and transport the patient to the hospital. Finally, the ambulance returns to its base location. As this total driving time depends on the varying traffic intensity, the varying treatment time at the emergency location, and the varying transfer time at the hospital, we represent this driving time by a stochastic parameter v_{ej} which denotes the time an ambulance from base $j \in J$ is occupied when it is assigned to emergency call $e \in E$. For now, we assume that this value can be determined beforehand for each pair of emergency call $e \in E$ and ambulance base location $j \in J$. When all emergency calls are assigned to an ambulance base location and the waiting times for emergency calls $e \in E$ are determined, we can calculate how many ambulances from base location $j \in J$ are occupied at each moment in time. To formalize this, we discretize the considered planning horizon in a set of time points denoted by T. The number of ambulances at base location $j \in J$ occupied at time $t \in T$ is now represented by an integer variable A_{jt} for which the correct value is determined by the following constraint:

$$A_{jt} = \sum_{e \in E} Z_{ej} \mathbb{1}_{\{t_e + W_e \leq t \text{ and } t_e + W_e + v_{ej} \geq t\}}, \; \forall j \in J, \; \forall t \in T. \tag{6}$$

At this tactical level, it is not straightforward to determine whether emergency call $e \in E$ is covered according to coverage requirement $k \in K$, because this depends on the assigned waiting time W_e and the varying travel time. To determine the coverage, we introduce a variable o_{ejk} denoting the time required to serve emergency call $e \in E$ according to coverage requirement $k \in K$ when assigned to ambulance base location $j \in J$. For example, when an ambulance must arrive at the incident location within 13 minutes, we have to consider the travel time from the assigned base location $j \in J$

to the location l_e of emergency call $e \in E$. Emergency call $e \in E$ is then covered according to coverage requirement $k \in K$ when $W_e + o_{ejk}$ is less than c_k, whereby c_k represents the time limit for coverage requirement $k \in K$. Again, binary variables Y_{ke} are used to specify whether emergency call $e \in E$ is covered according to coverage requirement $k \in K$. The correct value of Y_{ke} is determined by the following constraint:

$$W_e + \sum_{j \in J} Z_{ej} o_{ejk} \leq c_k + M(1 - Y_{ke}),\ \forall e \in E,\ \forall k \in K, \tag{7}$$

where M is a sufficiently large number. To make sure the correct fraction α_{kr} of all emergency calls in region $r \in R_k$ is covered according to coverage requirement $k \in K$, we get the following constraint:

$$\sum_{e \in E} Y_{ke} \mathbb{1}_{\{l_e \in I_{kr}\}} \geq \alpha_{kr} \sum_{e \in E} \mathbb{1}_{\{l_e \in I_{kr}\}},\ \forall k \in K, \forall r \in R_k. \tag{8}$$

The objective to minimize the number of ambulances needed is included as follows:

$$\min \sum_{j \in J} \max_{t \in T} A_{jt}. \tag{9}$$

Summarizing, the complete formulation of the problem at the tactical level is given by:

$$\begin{aligned} \min\ & \sum_{j \in J} \max_{t \in T} A_{jt} && && (10)\\ \text{s.t.}\ & \sum_{j \in J} Z_{ej} = 1, && \forall e \in E \\ & A_{jt} = \sum_{e \in E} Z_{ej} \mathbb{1}_{\{t_e + W_e \leq t \text{ and } t_e + W_e + v_{ej} \geq t\}}, && \forall j \in J,\ \forall t \in T \\ & W_e + \textstyle\sum_{j \in J} Z_{ej} o_{ejk} \leq c_k + M(1 - Y_{ke}), && \forall e \in E,\ \forall k \in K \\ & \sum_{e \in E} Y_{ke} \mathbb{1}_{\{l_e \in I_{kr}\}} \geq \alpha_{kr} \sum_{e \in E} \mathbb{1}_{\{l_e \in I_{kr}\}}, && \forall k \in K, \forall r \in R_k \\ & X_j, Y_{ki} \in \{0,1\} && \forall i \in I,\ \forall j \in J,\ \forall k \in K \end{aligned}$$

The formulations introduced in this section cannot be used in their present form to solve the problems on the tactical and strategic level simultaneously or separately in reasonable computation time. Therefore, some modifications and simplifications are introduced in the next section such that the problems can be solved within a reasonable amount of time.

4 Solution Approach

Solving the problem in two stages as described in the previous section may result in a suboptimal solution. However, when combining the strategic and tactical level, the size of the input instance increases and extra simplifications of the problem might be needed. Therefore, in the following we first present a possible solution approach for solving both levels simultaneously and, after that, present a solution approach for solving the two described problems hierarchically.

4.1 Strategic and Tactical Level Combined

As we want to take the uncertainty of emergencies into account, we chose stochastic programming to model the overall problem. A stochastic program in general is a mathematical program where stochastic elements are present in the data, which can influence the objective and/or the constraints. In practice, the detail in which uncertainty is implemented in the model can range from a few scenarios (as the possible outcomes of the data) to specific and precise joint probability distributions.

The formulation presented in this section is a two-stage stochastic program, with we further refer to as SPAB (Stochastic Planning of Ambulances and Basis), using scenarios that resulted from simulating a stochastic arrival process. We chose to model the occurrences of emergencies as a Poisson arrival process and simulated the arrival of emergencies at defined locations within a defined time interval. To simplify the problem we suggest one hour time intervals for the generation of scenarios as an ambulance is often occupied for about one hour when serving an emergency. This means that we do not have to incorporate time into our developed model as all ambulances are assigned to only one emergency. Therefore, we do not have to determine how long an ambulance is occupied.

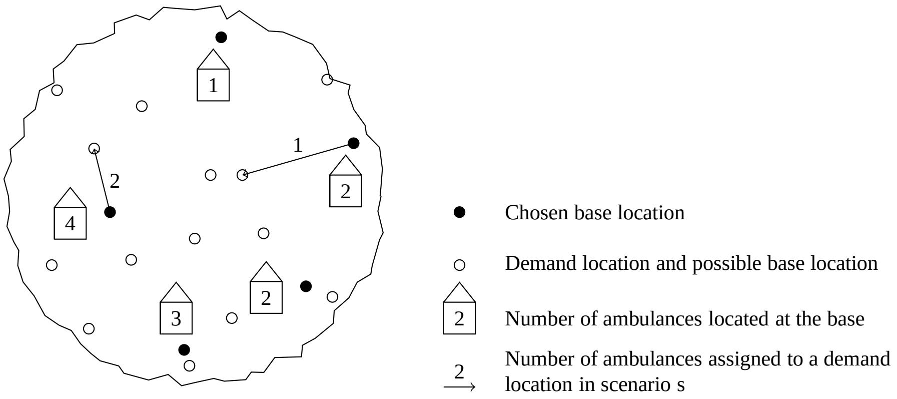
1
1
2
2
4
2
3
2
● Chosen base location
○ Demand location and possible base location
2 Number of ambulances located at the base
2 Number of ambulances assigned to a demand location in scenario s

Fig. 4 The location of the bases and the number of ambulances are determined simultaneously; for each scenario ambulances are assigned from the bases to the emergency calls

We consider two decision stages in our model. At the first stage, we decide where to locate the ambulance bases and how many ambulances to locate at which ambulance base. In a second stage, we consider all the potential scenarios and decide which ambulance shall be allocated to which emergency, in order to compute the coverage resulting from the chosen location decisions and ambulance configuration. As a result, we combine the strategic and the tactical level and solve both levels as only one problem as shown in Figure 4.

Additionally to the formulations of the previous section, we need a set S of scenarios. Each scenario $s \in S$ occurs with a given probability p^s. As stochastic parameters we have the number of emergencies occurring in demand location $i \in I$; for scenario $s \in S$ this number is denoted by d_i^s. Furthermore, we introduce decision variables G_j representing the number of ambulances at base location $j \in J$ and B_{ij}^s denoting the number of ambulances at base location $j \in J$ that are allocated to emergencies in demand location $i \in I$ when considering scenario $s \in S$. Recall that the variables X_j state whether base location $j \in J$ is opened. This leads to the following formulation, where M denotes a sufficiently large number and C being a constant greater than zero.

$$\min C \cdot \sum_{j \in J} G_j + \sum_{j \in J} X_j \tag{11}$$

$$\text{s.t.} \quad \sum_{i \in I} B_{ij}^s \leq G_j \qquad \forall j \in J,\ \forall s \in S \tag{12}$$

$$\sum_{j \in J} B_{ij}^s \leq d_i^s \qquad \forall i \in I,\ \forall s \in S \tag{13}$$

$$\sum_{i \in I_{kr}} \sum_{j \in J_{ik}} \sum_{s \in S} p^s B_{ij}^s \geq \alpha_{kr} \cdot \sum_{i \in I_{kr}} \sum_{s \in S} p^s d_i^s \qquad \forall k \in K,\ \forall r \in R_k \tag{14}$$

$$M \cdot X_j \geq G_j \qquad \forall j \in J \tag{15}$$

$$G_j \in \mathbb{N} \qquad \forall j \in J$$

$$X_j \in \{0,1\} \qquad \forall j \in J$$

$$B_{ij}^s \in \mathbb{N} \qquad \forall i \in I,\ \forall j \in J,\ \forall s \in S$$

The objective function (11) minimizes a weighted combination of the number of located ambulances and the number of bases opened. We cannot assign more ambulances from one node to emergencies than we have allocated to the node which is expressed by constraints (12). In addition, we are not allowed to assign more ambulances than needed for covering the emergencies at a node j which is stated by constraints (13). The constraints (14) assure that a fraction α_{kr} of the emergencies in region $r \in R_k$ is definitely covered by ambulances. Finally, the last constraints (15) only enable locating ambulances at base location $j \in J$ if the base is opened. Besides, the decision variables must be integer or binary.

When compared to the introduction of the problem in Section 3.2, we do not include waiting times and the stochastic driving times. As we use hourly intervals, an ambulance can only serve one call within this hour, and therefore, an ambulance is not used to serve multiple calls. Because of this, there is no waiting time for the arriving calls. In addition, we do not include stochastic driving times, but only use a fixed driving time which is used to determine whether demand location $i \in I$ is covered or not by one of the chosen base locations. Beforehand, a confidence interval for the driving times can be determined such that, for example, the driving time is met in 95% of the cases.

The approach proposed in this section solves the strategic and tactical level for the ambulance planning problem simultaneously. However, when considering the problem from a practical point of view it might also be natural to solve the problem in two stages. In the next sections, we discuss solution approaches for the strategic and tactical level that can be used to solve the two levels in two steps.

4.2 Strategic Level

In this section, we introduce a solution approach to solve the strategic level, further referred to as SAP (Strategic Ambulance Planning), without considering the tactical level at the same time. The ILP (4) formulated in Section 3.1 can directly be used to solve the strategic level and can be handled by a standard solver. The only question remaining is how to model the demand used in the coverage constraints for each demand location. As the demand fluctuates per day, it is hard to define a fixed value which results in a feasible solution every day. However, the coverage requirements usually only have to hold per year instead of each day, and therefore, it often is sufficient to include the average demand. This way it is ensured that most demand locations can be reached within the given time limit, and further alterations can be made on the tactical level by determining the correct number of ambulances per base.

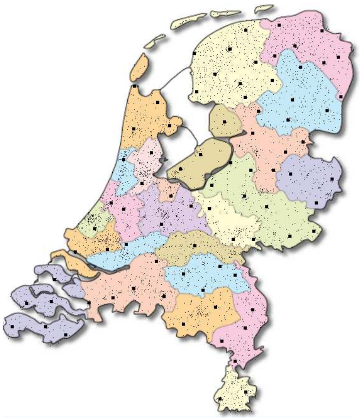

Fig. 5 Solution for average demand

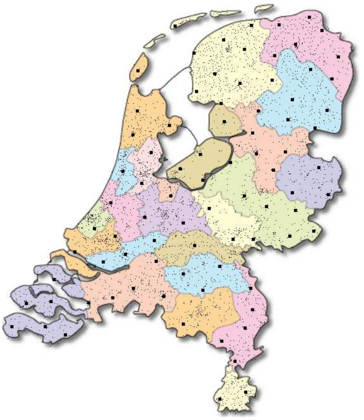

Fig. 6 Solution for equal demand

Because the base locations are chosen for a longer period, for example five years, we believe that for the strategic level it is sufficient to only consider the average demand for each demand location $i \in I$. The fluctuations of the demand can then be accounted for at the tactical or operational level. When the average demand is used, the base locations are situated such that demand locations with high demand are covered and locations with less demand might not be covered. However, the coverage requirements ensure that these differences are not too large. An alternative approach would be to set all demand equal to one (while keeping the α-values). In this way, each demand location is equally important. Figures 5 and 6 show that applying these two possibilities for the demand can result in different solutions. Using the Netherlands as an example and generating the two scenarios, we see that different base locations are chosen. The chosen base locations are depicted by the black squares and the small black dots are the demand locations. When all demand is set to one, 84 bases are needed, while only 81 bases are opened when the average demand is considered. A drawback when setting all demand to one, is that high demand locations might not be covered which leads to one or more of the coverage requirements not being fulfilled in practice. Therefore, we choose the average demand as input for the ILP at the strategic level.

4.3 Tactical Level

The formulation introduced in Section 3.2 cannot be used immediately to solve the problem at the tactical level. In practice, the emergency calls are not known beforehand and the number of emergency calls is too large to be included in an IP. We choose to model the arrival of emergency calls as a stochastic arrival process, and therefore, to solve this problem by means of simulation. The stochastic arrival process is modeled by the random demand scenarios S as defined in the stochastic programming approach. The solution approach developed in this section is further referred to as TAP (Tactical Ambulance Planning).

At the tactical level, we determine the number of ambulances needed at each opened base location based on the opening decisions on the strategic level to fulfill the considered coverage constraints. By using simulation, we can only determine whether a given configuration of the number of ambulances satisfies the given coverage constraints. To minimize the total number of ambulances needed, we combine the simulation with a local search heuristic.

Due to the stochastic arrival process of the emergency calls, the demand at the demand locations fluctuates per time interval. Therefore, the solution obtained by the solution approach used at the strategic level might not lead to a feasible solution in practice. Thus, it might be necessary to open one or more additional base locations to make sure that the coverage requirements are fulfilled. In practice, the ambulance base locations opened in the solution of the strategic level may be assumed to be used for several years. The base locations opened at the tactical level may not be real base locations in that sense, but, for example, a parking lot where the ambulance waits until a call arrives. This means that these base locations may change over the years.

Thus, in the first step in our simulation approach we start with the bases opened at the strategic level and we aim at opening one or more extra bases to make the solution robust for fluctuating demand (see Algorithm 1). In this step, we do not yet consider ambulances as first enough bases must be opened such that the coverage requirements given by constraints (2) are fulfilled for the considered scenarios. The addition of one or more ambulance bases is done using a greedy strategy by considering the coverage requirements one by one. The coverage requirements can be sorted in an hierarchical way by the number of regions considered. The coverage requirement with the maximum number of regions, i.e., with maximal $|R_k|$ is considered to be the lowest coverage level. The coverage requirements on lower levels are usually stricter and fulfilling these requirements will often already improve the coverage on higher levels. Therefore, we start at the lowest coverage level and check for each region

Algorithm 1 Add bases to fulfill coverage requirements

```
for all $k \in K$ do
    for all $r \in R_k$ with $\sum_{e \in E} Y_{ke} \mathbb{1}_{\{l_e \in I_{kr}\}} < \alpha_{kr} \sum_{e \in E} \mathbb{1}_{\{l_e \in I_{kr}\}}$ do
        repeat
            for all $j \in \bigcup_{l_e \in I_{kr}} J_{l_e k}$ do
                $\bar{X}_j := 1$
                $\bar{Y}_{kl_e} := \max_{j \in J_{kl_e}} \bar{X}_j$
                $\Delta_j := \frac{\sum_{e \in E} \bar{Y}_{ke} \mathbb{1}_{\{l_e \in I_{kr}\}}}{\sum_{e \in E} \mathbb{1}_{\{l_e \in I_{kr}\}}} - \frac{\sum_{e \in E} Y_{ke} \mathbb{1}_{\{l_e \in I_{kr}\}}}{\sum_{e \in E} \mathbb{1}_{\{l_e \in I_{kr}\}}}$
            end for
            Add base location $j$ for which $\Delta_j$ is maximum, i.e., $X_j := 1$ for ambulance base location
            $\arg\max_j \Delta_j$.
            $Y_{kl_e} := \max_{j \in J_{kl_e}} \bar{X}_j$
        until $\sum_{e \in E} Y_{ke} \mathbb{1}_{\{l_e \in I_{kr}\}} \geq \alpha_{kr} \sum_{e \in E} \mathbb{1}_{\{l_e \in I_{kr}\}}$
    end for
end for
```

$r \in R_k$ whether the coverage requirements are met or not with the additional chosen bases for the considered scenarios. When for region $r \in R_k$ the coverage requirement is not fulfilled, we add ambulance bases in the following way until the requirement is fulfilled. For each possible base location able to serve demand in this region $r \in R_k$, i.e., ambulance base locations $j \in \bigcup_{i \in I_{kr}} J_{ik}$, we determine what the change in coverage is when this base is opened. Then, we open the base which results in the highest increase in coverage. This is repeated until the coverage requirements are met for this region $r \in R_k$. This procedure is repeated for each coverage level until all coverage requirements are fulfilled for the considered demand scenarios.

Algorithm 2 Determine starting solution local search

```
Solve the following ILP:
 min   $\sum_{j \in J} G_j$
 s. t. $\sum_{j \in J} B^s_{ej} = 1,$                          $\forall e \in E,$
       $\sum_{j \in \cap_{k \in K} J_{l_e k}} B^s_{ej} = 1,$            $\forall e \in E$ for which $\cap_{k \in K} J_{l_e k} \neq \emptyset,$
       $G_j \geq \sum_{e \in E} B^s_{ej},$                     $\forall s \in S, j \in J,$
       $B^s_{ej} \in \{0,1\},$                       $\forall e \in E, j \in J, s \in S,$
       $G_j \in \mathbb{N},$                            $\forall j \in J.$
$G_j := 0$
for all $s \in S$ do
    $A_{jt} := \sum_{e \in E} B^s_{ej} \mathbb{1}_{\{t_e \leq t \text{ and } t_e + v_{ej} \geq t\}}$
    $G_j := \max\left(G_j, \max_{t \in T} A_{jt}\right)$
end for
```

After several ambulance base locations are added such that all coverage requirements are fulfilled, a local search procedure is used to minimize the number of ambulances. The first step in this local search heuristic is to determine an initial solution (see Algorithm 2) which determines a sufficient number of ambulances per base. To determine this initial number of ambulances per base, we assign each call to one of the opened bases in the set of opened bases that fulfil all coverage requirements for the considered call location. In this way, each call has its own ambulance. To be more precise, we solve a simple ILP which assigns all generated calls $e \in E$ to one of the opened bases $j \in \cap_{k \in K} J_{l_e k}$ that fulfills all coverage requirements if $\cap_{k \in K} J_{l_e k} \neq \emptyset$. If $\cap_{k \in K} J_{l_e k} = \emptyset$, then the call is assigned to one of the opened bases such that the total amount of ambulances needed to serve all calls is minimized. Note that although we consider several scenarios, each call $e \in E$ occurs only in one of the scenarios $s \in S$. Therefore, we can use binary variables B^s_{ej} to assign each call to one of the opened bases. Note that the number of ambulances G_j needed at base $j \in J$ equals $\max_s \sum_e B^s_{ej}$. As the emergency calls occur at different times in the considered time interval, the number of ambulances needed might be less as multiple calls might be served by the same ambulance. This is the case when there exist two or more emergency calls for which the busy time periods of the assigned ambulances do not overlap. By determining for each time $t \in T$ how many ambulances A_{jt} assigned to base $j \in J$ were simultaneously busy at time $t \in T$, the initial number of ambulances needed might be slightly reduced.

Algorithm 3 Local search to minimize number of ambulances needed

```
$F_j := \min\{G_j, 1\}$
**repeat**
    **for all** $j \in J$ for which $F_j = 1$ and $G_j > 0$ **do**
        $G_j := G_j - 1$
        **Algorithm 4:** Reassign calls
        **if** There exist a $k \in K$ and $r \in R_k$ for which $\sum_{e \in E} Y_{ke} \mathbb{1}_{\{l_e \in I_{kr}\}} < \alpha_{kr} \sum\limits_{e \in E} \mathbb{1}_{\{l_e \in I_{kr}\}}$ **then**
            $F_j := 0$
        **else**
            $\Delta_j := \max_{k \in K} \frac{\sum_{e \in E} \bar{Y}_{ke} \mathbb{1}_{\{l_e \in I_{kr}\}}}{\sum_{e \in E} \mathbb{1}_{\{l_e \in I_{kr}\}}} - \frac{\sum_{e \in E} Y_{ke} \mathbb{1}_{\{l_e \in I_{kr}\}}}{\sum_{e \in E} \mathbb{1}_{\{l_e \in I_{kr}\}}}$
        **end if**
        $G_j := G_j + 1$
    **end for**
    $\bar{j} = \arg\min_{j | F_j = 1} \Delta_j$
    $G_{\bar{j}} := G_{\bar{j}} - 1$
    **Algorithm 4:** Reassign calls
**until** $\sum_j F_j := 0$
```

In the simple approach sketched above, each call is served directly by one of the bases that fulfill the coverage requirements. It is expected that by adding some more intelligence in this choice, the number of ambulances needed per base can most likely be reduced. As a call might still be covered according to the coverage requirements when the patient waits until an occupied ambulance becomes available, we may not have to add an additional ambulance for such a call. Thus, we may further improve the initial solution in the next step by a local search procedure.

As we want to minimize the total number of ambulances needed, i.e. $\sum_{j \in J} G_j$, we reduce the number of ambulances until a further reduction leads to unfulfilled cov-

Algorithm 4 Reassign calls

```
$W_e := 0$ for all $e \in \bar{E}$.
$A_{jt} := 0$ for all $t \in T$ and the considered base $j \in J$.
for all $t \in T$ do
    for all $e \in \bar{E}$ for which $t_e + W_e = t$ do
        $\bar{J} := \{j \in \cap_{k \in K} J_{l_e k} | A_{jt} < G_j$ for $t \in [t_e + W_e, t_e + W_e + v_{ej}]\}$
        Add nearest base for $e \in E$ to set $\bar{J}$ if $A_{jt} < G_j$ for $t \in [t_e + W_e, t_e + W_e + v_{ej}]$ holds.
        if $\bar{J} = \emptyset$ then
            $W_e := W_e + 1$
        else
            Assign call $e \in E$ to nearest base $\bar{j} \in \bar{J}$, i.e., $Z_{e\bar{j}} := 1$.
            $A_{\bar{j}t} := A_{\bar{j}t} + 1$ for $t \in [t_e + W_e, t_e + W_e + v_{ej}]$.
        end if
    end for
end for
```

erage requirements (see Algorithm 3). We do this by reducing for each opened base $j \in J$ the number of ambulances G_j available by one and determining for which of the bases this reduction leads to the least decrease in coverage. For this case, the number of ambulances then definitely is reduced by one. By this approach, we aim to reduce the total number of ambulances as much as possible. Note that we only consider the bases for which the number of available ambulances G_j is at least one and for which the coverage requirements are still fulfilled after the reduction. To incorporate this in our local search approach, we introduce binary variables F_j which are one when base $j \in J$ can be considered for reduction and zero otherwise. For the base with the minimum reduction in coverage, the number of available ambulances G_j is then permanently reduced by one for the remainder of the local search approach. Note that after the reduction, not all calls can be served directly by the nearest base because not enough ambulances are available. Therefore, we possibly have to reassign the emergency calls that were assigned to the considered base $j \in J$ (see Algorithm 4). As not always an ambulance is available at the moment of arrival of the call, the call has to wait until an ambulance becomes available at the nearest base or is assigned to a base further away. By only reassigning the emergency calls assigned to the considered base, we give preference to the other calls as these are served immediately by the nearest base. This means that the reassigned calls have to wait until an ambulance becomes available for the entire duration of the transfer of the patient. A better solution can be achieved when all emergency calls are reassigned to a base, however, this will take too much time. The process of reducing and reassigning is repeated until none of the opened bases $j \in J$ is available any more for reduction, i.e. until $F_j = 0$ for all $j \in J$.

5 Computational Results

In this section, we test the developed approaches on data of the Netherlands. We determine values for the used parameters for both approaches and investigate which method performs the best. In addition, we determine the influence of the input data size on the computation time for both approaches.

5.1 Data

The discussed approaches are tested on real-world data of the Netherlands. The considered instances are based on data that is accessible on the internet. As the set of potential bases, we took the set of all 215 bases currently used in the Netherlands. The locations of these ambulance bases are given by the four digit zip code used in the Netherlands. As the set of hospital locations, we used all hospitals in the Netherlands that have an emergency department. These locations are also specified by a four digit zip code. The set of demand locations consists of all existing four digit zip codes in the Netherlands. Each four digit zip code can be linked to an RD-coordinate which is the coordinate system used in the Netherlands. These RD-coordinates can be used to determine the Euclidean distance between two locations. The Netherlands is divided into 24 regions, and for each of these regions the total demand for emergency calls is known. This total demand per region is divided over all demand locations in the region proportional to the number of inhabitants at the zip code of the considered demand location. To generate incoming calls, we assumed that the emergency calls can be modeled as a Poisson process.

In the Netherlands, an ambulance should arrive at the incident location within 13 minutes after the call came in for 97% of all incidents in a region (see [20]). In addition, all patients should arrive at the hospital within 45 minutes after the emergency call came in (see [25]). However, this last coverage requirement is not feasible for the used data as for some demand locations the nearest hospital is more than 45 minutes away. Therefore, we have introduce a coverage percentage of 99.5% for this constraint to guarantee that a feasible solution exists as we did not want to add additional possible base locations without having any information of where they could be located.

To generate scenarios, we generate incoming calls during one hour and do this for 10, 25, 50, 100, 250, 500, and 1000 independent hours. When more hours are considered the solution becomes more reliable, but also the computation time increases. Therefore, we investigate how many hours of incoming calls are needed to achieve a reliable solution within a reasonable amount of time. The generated scenarios are both used for the stochastic program and as input for the model on the tactical level. In the stochastic program, all scenarios have the same probability. The bases opened by the model at the strategic level are used as a starting point for the tactical level.

To determine the influence of the size of the input instance on the computation time, we also consider subinstances with only one, three, five, or ten of the 24 regions. These instances are depicted in Figure 7. The region with number one is used for the instance with one region. The regions with number two are added for the instance with three regions. The instance with five regions is obtained by adding the regions with number three and the instance with ten regions is formed by all the numbered regions.

For the objective function of the stochastic program we take $C = 1$.

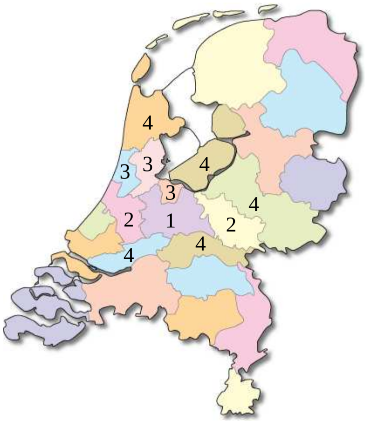

Fig. 7 Considered regions in the scenarios

5.2 Results

The results were generated on a PC with an AMD Phenom(tm) II X6 1100T Processor with 3.31 GHz and 16 GB RAM. All the approaches were implemented in AIMMS and the strategic and the stochastic models were solved with CPLEX 12.4. First of all, we investigate the number of scenarios needed (10, 25, 50, 100, 250, 500, and 1000 hours) to achieve reliable solutions within a reasonable amount of time. After this, we compare the two approaches of solving the strategic and tactical level combined and separately. The approach for the strategic and tactical level combined is SPAB and the approach for solving the strategic and tactical separately is a combination of SAP and TAP which is further referred to as STAP. For the achieved solutions, we give the number of bases opened, the number of ambulances needed and the required computation time for applying the approaches to the considered instances. In addition, we created smaller instances with only one, three, five, and ten regions to be able to analyze the computation times depending on the problem sizes. Due to an increasing computation time some of the instances were stopped after 24 hours. This is indicated by a $*$ in the tables. In addition, for the instances which were not solved after 24 hours, we give the integrality gap of the ILP for the strategic level.

First, we aim to determine the number of scenarios that have to be considered such that a reliable solution is found within a reasonable amount of time. To determine this, we first provide graphs which show the number of ambulances needed in the solutions of SPAB and STAP depending on the number of scenarios used. Figures 8 and 9 show that the number of ambulances increases when the number of scenarios increases, but stabilizes at a certain point. This is to be expected as the number of ambulances needed at a certain base is closely related to the maximum number of calls in the area covered by this bases over all scenarios. This maximum is likely to increase when more scenarios are considered, because a larger range of values is generated. However, at some point the maximum value is achieved and the number of ambulances needed stabilizes.

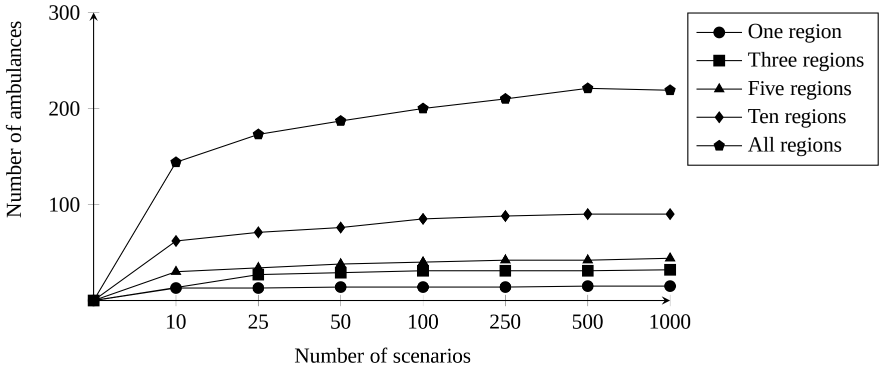

Fig. 8 SPAB: number of ambulances given for different number of regions for different number of scenarios

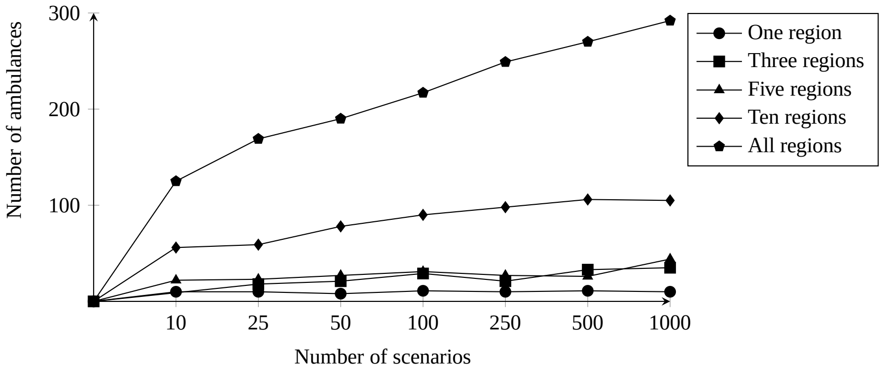

Fig. 9 STAP: number of ambulances given for different number of regions for different number of scenarios

Figures 8 and 9 also show that more scenarios are needed when the instance size increases. The total number of ambulances needed stabilizes when the maximum demand is achieved for each demand location in one of the considered scenarios. This is because the configuration for the number of ambulances per base determined by STAP or SPAB has to be feasible for all considered scenarios. When the instance size increases, more demand locations are considered, and thus, more scenarios are needed to achieve this maximum for each demand location.

However, not only the reliability of the solution is important when determining the base locations and number of ambulances needed but also the computation time. Figure 10 shows the computation times for STAP and SPAB when 10, 25, 50, 100, 250, 500, and 1000 scenarios are considered for the instance with ten regions. From this graph, we can conclude that the computation time for SPAB increases exponentially whereas the computation time for STAP follows a more linear trend. Based on the results for the computation time and the solution reliability, we suggest to use 100 scenarios to get a good trade-off.

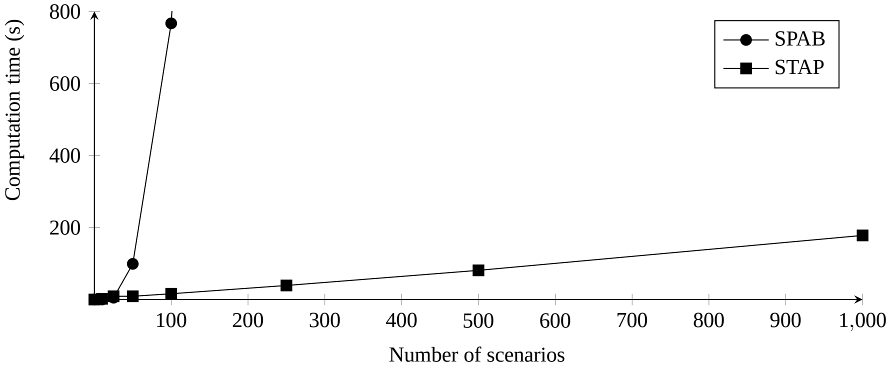

Fig. 10 Computation time for different number of scenarios and ten regions

Table 1 Comparison of solving STAP and SPAB (100 scenarios)

	# Regions	SAP	TAP	SPAB
# Bases	1	5	5	7
# Ambulances	1	–	11	14
# Bases	3	9	11	13
# Ambulances	3	–	29	31
# Bases	5	11	11	16
# Ambulances	5	–	31	40
# Bases	10	30	33	38
# Ambulances	10	–	90	85
# Bases	24	81	85	108
# Ambulances	24	–	217	200

To compare SPAB and STAP, we investigate the computation time and solution quality of both approaches. Table 1 presents the number of opened bases and assigned ambulances and Figure 11 shows the computation time for the considered instances when 100 scenarios are taken into account. Table 1 shows that for almost all instances additional bases are opened at the tactical level to be able to fulfill the coverage requirements for all scenarios. In addition, the number of opened bases for SPAB is higher than for STAP. An explanation for this is that for STAP more emphasis is put on minimizing the number of bases as this is done in the first step, while for SPAB minimizing the number of bases and ambulances is equally important. For the instances with one, three, and five regions the number of needed ambulances is lower for STAP than for SPAB. This is because in STAP an ambulance can be used for more than one emergency call whereas in SPAB only one call is assigned to each ambulance. However, for the instances with ten and all regions, the number of ambulances needed is lower for SPAB than for STAP. For these larger instances, the area covered by bases not near the boarders is larger. The calls in this area can easily be spread out over multiple bases as the locations of these calls are surrounded by bases, which is not the case in boarders regions. By spreading the calls over multiple bases, the number of ambulances needed can be reduced because the peak load on one base is spread out over other opened bases. Therefore, this effect is larger for SPAB than for STAP as more bases are opened in the solution of SPAB.

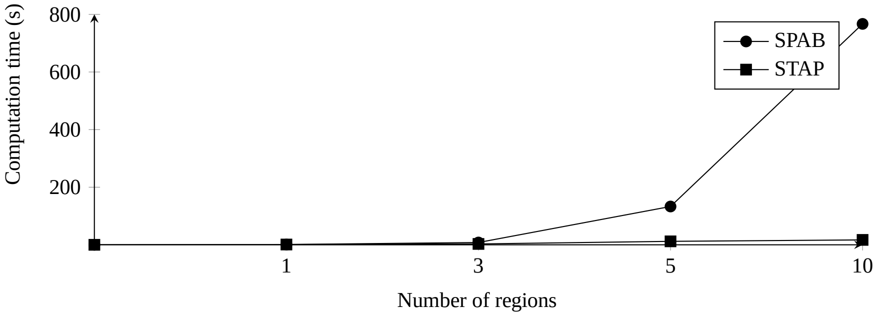

Fig. 11 Computation time for different number of regions with 100 scenarios

Figure 11 shows the computation time for SPAB and STAP for the instances with one, three, five, and ten regions when 100 scenarios are evaluated. It is evident that the computation time increases when the input size increases, however, the computation time of SPAB increases exponentially whereas the computation time of STAP follows a more linear trend. This linear trend for STAP does not continue for the instance with all 24 regions because the computation time for SAP takes more than 24 hours to prove that the found solution is optimal. However, this computation time can be reduced by interrupting the solver when a certain acceptable integrality gap is achieved.

Concluding, when looking at the solution quality, STAP performs better in minimizing the number of bases and SPAB in minimizing the number of ambulances. With SPAB it is also easier to make a good trade-off between the two objectives because of the weighted objective function. However, the computation time for SPAB explodes faster than the computation time of STAP.

As mentioned above, we have chosen $C = 1$ for the objective function of SPAB. In this case, minimizing the numbers of bases and ambulances is equally important. The results depicted in Table 2 show that by choosing a higher value for C the number of ambulances can be reduced by approximately 6%. However, this is at the cost of the number of bases opened as this increases with approximately 20%. Depending on the real-world situation, it can be the case that this is actually preferred. For example, if a large number of bases already exists or additional locations like parking lots can be used, it might be more important to avoid additional ambulances that would cause investments or need extra staff. Surprisingly, Table 2 also shows that the computation time can be significantly reduced when a higher value for C is chosen. For example, when considering the instance with ten regions and 100 scenarios, the computation times differ by nearly a factor of 16. Because the emphasis is on the number of ambulances, the options for opening bases are limited, and therefore, the solution space reduces significantly. When C is set to one, a trade-off has to be made between both objectives and this results in a larger solution space.

Table 2 The impact of C, having 10 regions and 100 scenarios

Value of C	# Base locations	# Ambulances	Computational time (s)
1	38	85	767
1000	46	80	48

6 Conclusions and Recommendations for Further Research

In this paper, we have considered the ambulance planning problem and discussed that it is implicitly treated on different levels. For the first two levels, the strategic and the tactical level, we have presented formulations and solution approaches. Furthermore, for the tactical level we have proposed a simulation combined with a local search. In addition, we gave an approach for solving both levels simultaneously by the means of stochastic programming. Using test instances derived for the Netherlands, we compared the presented approaches.

One main advantage of the formulations proposed in this paper is that the coverage constraints are modeled in a very generic way, and therefore, they can be easily adapted to different requirements within different countries. As a consequence, the formulations are not only suitable for the Netherlands but probably for the EMS systems of many countries. Based on the first results, it can be stated that the approaches are promising. However, there is still some further work that needs to be done. First of all, the approaches should be tested using real-world data. We are planning to do this in the form of a case study for the Netherlands. During this case study it would also be interesting to compare the current situation to the best solution found by our approach and check to which extent we are able to improve it. A further topic of research is to have a closer look at the different EMS systems in Europe (or even worldwide) to check how good the presented approaches fit to those systems and what further improvements are needed to make them even more generic. Going along with it, the tactical approach should be tested for larger instance sizes, i.e., bigger countries, to prove the applicability. For SAP and SPAB modifications should be examined that make the problems solvable for larger instance sizes. In the current approach, we chose a simple local search because it converges quickly to a local optimum. Other approaches like simulated annealing or tabu search might find better solutions, but will probably take longer. Nevertheless, it might be interesting to implement one of those heuristics and to compare computation times and solution qualities.

Concerning the simulation it would be of interest to implement it in a simulation software like AnyLogic to make solutions also ‘visible’. This would especially be helpful for practitioners when using the approaches in practice. The simulation could then be adapted to the operational level, too. In this paper, we left out the operational level since it differs significantly from the other two. Nevertheless, this level is very important for the quality of the EMS system and should therefore be incorporated into follow-up research. For example, it could be investigated how much the solution quality on the operational level depends on the solutions obtained at the tactical level when for example ambulances can be relocated to different locations throughout the day. For the operational level the number of ambulances needed together with their locations for different times of the day, week, month and year could be determined and possible allocation strategies could be tested. One of the major challenges at the operational level will be that for applying approaches in practice solutions are needed in real-time.

Another adaptation that is needed to make the developed approaches more applicable in practice is to implement different driving times for different situations, e.g., rush hours. This would make the subsets J_{ki} dependant of the considered scenario as some bases in subset J_{ki} might not cover demand location $i \in I$ anymore when the driving time increases. We could, for example, consider three different configurations of the driving time (worst, normal and best) to keep the computation times reasonable. We plan to incorporate this adaptation in our planned case study for the Netherlands.

Acknowledgements This research is supported by the Dutch Technology Foundation STW, applied science division of NWO and the Technology Program of the Ministry of Economic Affairs.

References

1. Aly, A.A., White, J.A.: Probabilistic formulation of the emergency service location problem. The Journal of the Operational Research Society **29**(12), 1167–1179 (1978)
2. Beraldi, P., Bruni, M.E.: A probabilistic model applied to emergency service vehicle location. European Journal of Operational Research **196**(1), 323–331 (2009)
3. Beraldi, P., Bruni, M.E., Conforti, D.: Designing robust emergency medical service via stochastic programming. European Journal of Operational Research **158**(1), 183–193 (2004)
4. Berlin, G.N., Liebman, J.C.: Mathematical analysis of emergency ambulance location. Socio-Economic Planning Sciences **8**(6), 323–328 (1974)
5. Berman, O., Drezner, Z., Krass, D.: Discrete cooperative covering problems. Journal of the Operational Research Society **62**(11), 2002–2012 (2011)
6. Brotcorne, L., Laporte, G., Semet, F.: Ambulance location and relocation models. European Journal of Operational Research **147**(3), 451–463 (2003)
7. Chapman, S.C., White, J.A.: Probabilistic formulations of emergency service facilities location problems. In: ORSA/TIMS Conference, San Juan, Puerto Rico (1974)
8. Church, R., ReVelle, C.: The maximal covering location problem. Papers in Regional Science **32**(1), 101–118 (1974)
9. Daskin, M.S.: A maximum expected covering location model: Formulation, properties and heuristic solution. Transportation Science **17**(1), 48–70 (1983)
10. Fujiwara, O., Makjamroen, T., Gupta, K.K.: Ambulance deployment analysis: A case study of bangkok. European Journal of Operational Research **31**(1), 9–18 (1987)
11. Galvão, R.D., Chiyoshi, F.Y., Morabito, R.: Towards unified formulations and extensions of two classical probabilistic location models. Computers and Operations Research **32**(1), 15–33 (2005)
12. Gendreau, M., Laporte, G., Semet, F.: Solving an ambulance location model by tabu search. Location Science **5**(2), 75–88 (1997)
13. Geroliminis, N., Karlaftis, M.G., Skabardonis, A.: A spatial queuing model for the emergency vehicle districting and location problem. Transportation Research Part B: Methodological **43**(7), 798–811 (2009)

14. Goldberg, J., Dietrich, R., Chen, J.M., Mitwasi, M., Valenzuela, T., Criss, E.: A simulation model for evaluating a set of emergency vehicle base locations: Development, validation, and usage. Socio-Economic Planning Sciences **24**(2), 125–141 (1990)
15. Harewood, S.I.: Emergency ambulance deployment in barbados: A multi-objective approach. The Journal of the Operational Research Society **53**(2), 185–192 (2002)
16. Henderson, S., Mason, A.: Ambulance service planning: Simulation and data visualisation. In: Operations Research and Health Care, *International Series in Operations Research and Management Science*, vol. 70, pp. 77–102. Springer US (2005)
17. Iannoni, A.P., Morabito, R.: A multiple dispatch and partial backup hypercube queuing model to analyze emergency medical systems on highways. Transportation Research Part E: Logistics and Transportation Review **43**(6), 755–771 (2007)
18. Iannoni, A.P., Morabito, R., Saydam, C.: Optimizing large-scale emergency medical system operations on highways using the hypercube queuing model. Socio-Economic Planning Sciences **45**(3), 105–117 (2011)
19. Karasakal, O., Karasakal, E.K.: A maximal covering location model in the presence of partial coverage. Computers and Operations Research **31**(9), 1515–1526 (2004)
20. Kommer, G.J., van der Veen, A.A., Botter, W.F., Tan, I.: Ambulances binnen bereik: Analyse van de spreiding en beschikbaarheid van de ambulancezorg in nederland. Tech. rep., National Institute for Public Health and the Environment (2003)
21. Larson, R.C.: A hypercube queuing model for facility location and redistricting in urban emergency services. Computers and Operations Research **1**(1), 67–95 (1974)
22. Li, X., Zhao, Z., Zhu, X., Wyatt, T.: Covering models and optimization techniques for emergency response facility location and planning: a review. Mathematical Methods of Operations Research **74**(3), 281–310 (2011)
23. Marianov, V., ReVelle, C.: Siting emergency services. Facility Location: a survey of applications and methods **1**, 199–223 (1995)
24. Marianov, V., ReVelle, C.: The queueing maximal availability location problem: A model for the siting of emergency vehicles. European Journal of Operational Research **93**(1), 110–120 (1996)
25. Nederland, A.: Ambulances in-zicht 2011. Tech. rep., National Institute for Public Health and the Environment (2011)
26. Noyan, N.: Alternate risk measures for emergency medical service system design. Annals of Operations Research **181**(1), 559–589 (2010)
27. Owen, S.H., Daskin, M.S.: Strategic facility location: A review. European Journal of Operational Research **111**(3), 423–447 (1998)
28. ReVelle, C., Hogan, K.: A reliability-constrained siting model with local estimates of busy fractions. Environment and Planning B: Planning and Design **15**(2), 143–152 (1988)
29. ReVelle, C., Hogan, K.: The maximum availability location problem. Transportation Science **23**(3), 192–200 (1989)
30. Silva, F., Serra, D.: Locating emergency services with different priorities: The priority queuing covering location problem. The Journal of the Operational Research Society **59**(9), 1229–1238 (2008)
31. Swoveland, C., Uyeno, D., Vertinsky, I., Vickson, R.: A simulation-based methodology for optimization of ambulance service policies. Socio-Economic Planning Sciences **7**(6), 697–703 (1973)
32. Takeda, R.A., Widmer, J.A., Morabito, R.: Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model. Computers and Operations Research **34**(3), 727–741 (2007)
33. Toregas, C., Swain, R., ReVelle, C., Bergman, L.: The location of emergency service facilities. Operations Research **19**(6), 1363–1373 (1971)
34. Zaki, A.S., Cheng, H.K., Parker, B.R.: A simulation model for the analysis and management of an emergency service system. Socio-Economic Planning Sciences **31**(3), 173–189 (1997)

Appendix

Table 3 Notation used in the given problem formulations

Notation	Description
Sets	
I	set of demand locations
J	set of potential base locations
K	set of coverage requirements
J_{ik}	subset of potential base locations that fulfill coverage requirement $k \in K$ for demand location $i \in I$
R_k	set of regions to be considered for coverage requirement $k \in K$
I_{kr}	subset of demand locations that lie within region $r \in R_k$
E	set of emergency calls
T	set of time points in the considered planning horizon
S	set of scenarios
Parameters	
d_i	demand at demand location $i \in I$
α_{kr}	fraction of demand that should fulfill coverage requirement $k \in K$ for region $r \in R$
t_e	arrival time of emergency call $e \in E$
l_e	location of emergency call $e \in E$
v_{ej}	time an ambulance from base $j \in J$ is occupied when it is assigned to emergency call $e \in E$
o_{ejk}	time emergency call $e \in E$ is occupied according to coverage requirement $k \in K$ when assigned to ambulance base location $j \in J$
c_k	time limit for coverage requirement $k \in K$
p^s	probability for scenario $s \in S$
C	constant greater than zero
Variables	
X_j	binary variable which is one when base location $j \in J$ is selected, and zero otherwise
Y_{ki}	binary variable which is one when demand location $i \in I$ is covered according to coverage requirement $k \in k$
Z_{ej}	binary variable which is one when emergency call $e \in E$ is served by ambulance base $j \in J$
W_e	waiting time of emergency call $e \in E$
A_{jt}	number of ambulances at base location $j \in J$ occupied at time $t \in T$
G_j	the number of ambulances at base location $j \in J$
B^s_{ij}	the number of ambulances at base location $j \in J$ allocated to emergencies in demand location $i \in I$ when considering scenario $s \in S$