Demand-point constrained EMS vehicle allocation problems for regions with both urban and rural areas

Martin van Buuren · Rob van der Mei · Sandjai Bhulai

Abstract Governments deal with increasing health care demand and costs, while budgets are tightened. At the same time, ambulance providers are expected to deliver high-quality service at affordable cost. Maximum reliability and minimal availability models guarantee a minimal performance level at *each* demand point, in contrast to the majority of facility location and allocation methods that guarantee a minimal performance that is *aggregated* over the entire ambulance region. As a consequence, existing models generally lead to overstaffing, particularly in 'mixed' regions with both urban and rural areas, which leads to unnecessarily high costs. This paper addresses this problem. First, we introduce the concept of *demand projection* to give fundamental insight into why this overstaffing takes place. Next, we overcome the overstaffing by the so-called *adjusted queuing* (AQ) solution that provides generalizations of the existing models. We provide mathematical proofs for the correctness of the AQ solution. Finally, to assess the performance of the AQ-solution we have performed extensive numerical experimentation, using real data from four ambulance regions in the Netherlands. The results show that in all cases the AQ-solution indeed leads to better ambulance care than the existing solutions, while reducing staffing cost.

Keywords EMS, ambulance, allocation, facility location, reliability, queuing theory

M. van Buuren
✉ *Corresponding Author*
Centrum Wiskunde & Informatica
123 Science Park, Amsterdam, The Netherlands
Tel.: +31-20-592-4365
E-mail: m.van.buuren@cwi.nl

R.D. van der Mei
Centrum Wiskunde & Informatica
123 Science Park, Amsterdam, The Netherlands
Tel.: +31-20-592-4129
E-mail: r.d.van.der.mei@cwi.nl

S. Bhulai
VU University Amsterdam
1081a De Boelelaan, Amsterdam, The Netherlands
Tel.: +31-20-598-7679
E-mail: s.bhulai@vu.nl

1 Introduction

Ambulance service providers (ASPs) need to find the right balance between good quality of service and reasonable costs. Hence, they are interested in what the best locations for ambulance bases are, and what the optimal number of ambulances is for each of these bases. Overcapacity of ambulances leads to unnecessarily high costs, while extensively reducing the costs can lead to dangerous situations. Many facility location and allocation problems address this trade-off.

Most systems try to catch the performance of an ambulance region in a single value, e.g. by calculating the fraction of late arrivals aggregated over the entire ambulance region for a duration of one year. When maximizing this covered fraction, it may lead to low coverage in rural areas in favor of densely populated cities. We call them the *regional coverage location problem* (RCLP) class of models; a name that stresses the regionally aggregated key performance indicator.

Another approach is to evaluate the performance of each district in the region individually, and satisfy at least a minimal required performance threshold that is set for each district in the best possible way. This can be achieved by giving a minimal performance constraint to every district. Either this means to find an ambulance allocation to satisfy the requirement for the best possible subset of districts when limited resources are available (maximal availability), or to determine the minimal number of ambulances and the resulting allocation such that the minimal requirement for every district is satisfied (minimal reliability). This lead tot the *maximum availability and minimal reliability* (MR-MA) class of models. Our paper has its focus on this class.

Currently RCLP is mostly used in practice, although we note an emerging balance shift in favor to MR-MA. Erkut, Ingolfsson, and Budge [**?**] wrote in what they call a critique on the MR-MA models that “the objective functions of the models in this class are not the same as the expected coverage performance measure that typically drives EMS system design”. This is still a valid argument though it loses its strength as time goes on. The trend shift from ambulance practice moving away from purely focussing on this RCLP class objective is due to pressure from local intraregional governments. Although the key performance indicator for the regions that we consider in the results section in practice still is the fraction of calls covered within a time threshold of 15 minutes measured on a yearly basis, they have to deal with multiple mayors of rural municipalities in their region who insist on a minimal coverage for their own population. As a result, every district (or: municipality) within the safety region must also receive coverage under a minimal reliability constraint. The current MR-MA models in literature are not suited for regions that have both rural and urban areas. The practical need for MR-MA models that are suitable for these so-called mixed regions provide the motivation for this research.

The existing high-end MR-MA models in literature Q-PLSCP and Q-MALP [**?**,**?**], throughout this paper referred to as the Q-models, are made for regions where the demand is fairly homogeneously spread. If they are applied to actual regions with inhomogeneous demand - against their original design - they generally lead to over-estimations for the required number of vehicles [**?**,**?**]. In that case, as a consequence, ambulance providers have an unnecessarily high cost.

In this paper we propose a new MR-MA approach that is applicable to regions with urban and rural demand. Our approach is based on many concepts that can be found in the Q-models, to such an extend that we call our approach the *adjusted queuing* solution.

The present paper can roughly be divided in three parts. First we provide an *in depth explanation* of why the over-estimation takes place when Q-models are directly applied to mixed regions (Sections 3–4). Secondly, we propose the *adjusted queuing solution* that leads to credible results (Sections 5–7). Third, we show that our proposed adjusted queuing solution leads to *credible results* for four actual mixed regions (Section 8). We use the maximum reliability model Q-PLSCP for illustration purposes throughout the current paper, although the findings are not limited to this model.

The adjusted queuing formulation AQ-PLSCP improves on Q-PLSCP as follows.

- Q-models tend to project urban demand to rural areas that may lead to major local overstaffing in rural areas. The adjusted queuing approach that we propose solves this problem, leading to better staffing for these rural areas.
- Contrary to Q-models, we allow for major differences in demand and service time.
- In AQ-models, the required reliability or availability level is demand point dependent instead of a system wide constant.

We provide mathematical proofs that the adjusted queuing approach works. Existing papers use simulation studies to illustrate that their method work [?,?]. Because the models in these papers are special cases of AQ-models, our paper also includes mathematical proofs of these models.

The remainder of this paper is as follows. In Section 2 we give an extended literature review on the line of models that leads to the Q-models. Section 3 starts with some definitions and assumptions on the ambulance practice, and it provides a detailed description of the Q-PSCLP model that is required later on in the paper. We show in Section 4 why the Q-models give over-estimations on the number of ambulances needed in an ambulance region. In Section 5 we replace the main assumption of previous models with the more general workload condition. In Section 6 this workload condition is used to give a solution to the over-estimation. Section 7 proposes the new *adjusted queuing* model formulation for PSLCP. In Section 8 we compare results of the Q-PLSCP model to those of AQ-PLSCP for four actual ambulance regions, and we show that the AQ-models lead to useful results. Section 9 contains the conclusion and gives advise for future research. Appendix A contains a convenient overview of all variables and their meaning used through the paper. The proofs of all theorem, lemmas, propositions, properties and corallaries are bundled in the Appendix B.

2 Literature review

This section starts with an overview on the RCLP class of models. Thereafter we describe the MR-MA models on which our research elaborates. We end this section by mentioning a few other related research topics in the field of emergency medical services (EMS) logistics.

The RCLP model maximum expected coverage location problem (MEXCLP) by Daskin (1983) is amongst the first that link EMS facility location to the stochastic nature of EMS logistics [?]. MEXCLP is based on the deterministic Maximal Covering Location Problem (MCLP) by Church [?]. The latter positions a given number of ambulances such that the demand that is covered at least once is maximized. Earlier models all have a deterministic nature. Many papers can be found in literature that are in some sense extension of MEXCLP. The notion of double coverage is introduced much later in the backup coverage problems (BACOP) [?] and the double standard model (DSM) [?]. A fleet with multiple vehicle types can be found in the TEAM and FLEET models that are based on MCLP. [?]. The combination of TEAM and MEXCLP can be found in the two-tiered model (TTM) [?]. The MOFLEET model combines MEXCLP and FLEET. A time-dependent version of MEXCLP [?], TIMEXCLP runs MEXCLP once for every time interval [?]. More recent, Rajagopalan et al. [?] try four multiple meta-heuristic search methods to find good solutions in the case MEXCLP becomes hard to solve.

The MEXCLP is an IP-formulation that maximizes the expected covered demand. There is a constant system wide parameter $q \in [0, 1]$ stating the probability that an ambulance is not available. Every ambulance in the system is assumed to have the same probability of being unavailable. Given the number of ambulances that can reach this demand point within the time threshold, the binomial distribution gives the probability that at least one ambulance can cover this demand point. Multiplication by the weight of the demand point, e.g. demand or the population density, yields the expected population covered at this demand point. Summation of this expected value over all demand points gives the maximized expected coverage over the entire region. MEXCLP places a fixed number of ambulances in such a way that the expected coverage is maximized. This method has two major disadvantages that other newer methods still inherit: 1) Many regions have both rural and urban areas. A constant system wide busy fraction q for each ambulance is not realistic because demand points with less demands will most likely have a lower busy fraction. 2) Rural areas have a lower population density, so the method decreases coverage in rural areas in favor of densely populated urban population. Major differences in ambulance care between a region's population can occur. From an equity perspective this may be deemed unfair. For further work on fairness in EMS logistics see Chapter 5-6 of Jagtenberg (2016) [**?**].

The MR-MA class of models has a completely different approach at which a minimal local coverage performance level is set for every demand point. Minimal reliability guarantees that every demand point has the minimal required coverage, and minimizes the total number of ambulances in the system to achieve this. Maximal availability swaps these objective function and constraints, thus allocating a fixed number of ambulances in such a way that a maximal demand is covered under the minimal reliability threshold. The Location Set Covering Model (LSCM) by Toregas (1971) is the first minimal reliability model that optimizes the number of facilities such that each demand point can be reached by at least one ambulance within a given response time threshold [**?**]. The Maximum Availability Location Problem (MALP) by Revelle and Hogan (1989) is the first maximum availability model [**?**]. A demand point is covered when the probability that it can be reached within the response time (or distance) threshold by at least one ambulance exceeds a constant α. In the first MALP model there is a system wide busy fraction for the ambulances used, and a Bernoulli approach similar to MEXCLP calculates for each demand point the minimal number of ambulances that is required to satisfy the required minimal required coverage probability constraint. MALP allocates a fixed number of ambulances in such a way that the total covered demand is maximized. In the same paper an extension is proposed where the ambulance's busy fraction depends on the demand point. The Reliability Perspective (Rep-P) by Ball and Lin [**?**] poses an upper bound on the reliability of every demand point, and it ensures that every demand point i is covered by a minimal reliability level that can be set for each demand point separately. Borrás and Pastor [**?**] adapt MALP and Rel-P in such a way that the busy fraction of each base location depends on a preference list that each demand point holds. This way the busy fraction becomes more realistic, and the authors show that it leads to a reduction in the number of vehicles. Sorenson and Church (2010) combine in the LR-MEXCLP the maximal coverage objective of MEXCLP with the reliability constraints of MALP [**?**]. They do this by not using the boolean coverage constraint by demand point but rather use a slope with the reliability by which a demand point is covered. The PLSCP model updates Rel-P in such a way that the busy fraction of a demand point depends on the number of available servers [**?**].

Q-PLSCP and Q-MALP by Marianov and Revelle are the queuing versions of PLSCP and MALP [**?**,**?**]. Instead of a binomial approach, they use an Erlang B formulation. Our AQ-models extend these models; see Subsection 3.4. Subsection 3.4.1 describes the difference in assumptions and outcomes between the binomial approach and the queuing approach, because the literature is rather limited on the subject.

Larson [**?**] describes a hypercube model that can be used to evaluate the choice of base locations and ambulance allocation. Since he uses Poisson arrivals and exponential service times, this is a Markov process. The name hypercube refers to the state of the set with ambulances, where each vehicle can be available or busy. Various extensions on this model are provided in the literature [**?**,**?**,**?**,**?**,**?**].

Good reviews on EMS facility location are available; see Marianov and Revelle (1995), Brotcorne et al. (2003) and Li et al. (2011) [**?**,**?**,**?**]. Facility location and staffing takes place in the strategic and tactical domain of EMS. Golberg [**?**] discusses the properties, advantages, and disadvantages of the various models in his recent review paper.

At the operational level dynamic ambulance management (DAM) is widely used [**?**,**?**,**?**,**?**,**?**,**?**,**?**,**?**,**?**,**?**,**?**]. Sometimes simulation tools are used to validate models. An extensive review on simulation models in EMS can be found in Aboueljinane (2014) [**?**]. DAM relocates available ambulances in such a way that the regions coverage stays optimal.

In this paper the TIFAR-framework in combination with Coin-OR is used for implementation [?,?].

3 Preliminaries: definitions, assumptions and previous model formulation

In this section we start by giving the definitions used throughout the paper. Next, we discuss the assumptions on the ambulance practice. Since the queuing approach for ambulance allocation models is not generally known, we share the differences between this approach and the binomial approach found in many papers. This section ends with a detailed description of the Q-PLSCP model that is our baseline for illustration purposes.

In the next section we show why the current implementation of the queuing approach, like Q-PLSCP, yields over-estimations when they are naively applied to mixed regions, and the remainder of this paper is dedicated to finding and proving a solution.

3.1 Definitions

Figure 1 illustrates the stages of the emergency medical service (EMS[1]) process, which holds the majority of definitions we use in this paper.

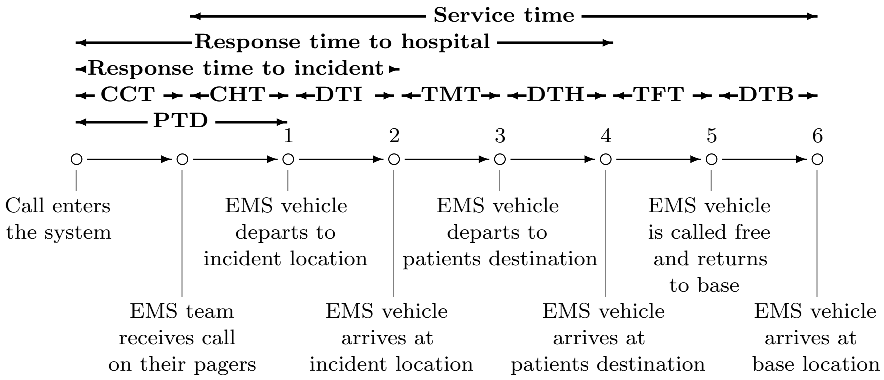
Service time
Response time to hospital
Response time to incident
CCT
CHT
DTI
TMT
DTH
TFT
DTB
PTD
1
2
3
4
5
6
Call enters the system
EMS team receives call on their pagers
EMS vehicle departs to incident location
EMS vehicle arrives at incident location
EMS vehicle departs to patients destination
EMS vehicle arrives at patients destination
EMS vehicle is called free and returns to base
EMS vehicle arrives at base location

Fig. 1 The trace of call statuses and the corresponding time intervals.

When a call enters the system, the dispatcher at the emergency medical call center (EMCC) performs a triage procedure, and, if required, dispatches an ambulance. The duration in which these two processes are done is the so-called call center time (CCT). When the pagers of the emergency medical technicians (EMTs) are activated at dispatch, some time passes before the ambulance starts moving, since they have to get to the vehicle. This time period is called the chute time (CHT). The entire duration from a call entering the system until the ambulance starts moving is called the pre-trip delay (PTD).

The next stage is the driving time to incident (DTI), which is followed by the treatment time (TMT). In some cases, the patient is treated at the incident location after which the ambulance returns to the base location. In other cases, the patient is brought to the hospital. We call this time interval the driving time to hospital (DTH), after which we have a transfer time (TFT) that bridges pre-hospital care with hospital care. Sometimes the transfer time is called the turn-around time. The last stage of the ambulance trip consists in driving back to its base location (DTB).

[1] A list with abbreviations and variables is located at the end of the document in Table 9.2.

For basic live support (BLS) transport, and occasionally for advanced life support (ALS) transport, the incident location is a hospital and the destination may be another hospital or a home address. Since our main focus is on ALS transport with random arrivals, we keep respecting these definitions, even if the patient's destination is not a hospital.

There are two response time thresholds used in the daily life of emergency medical services. First we have the response time to incident that starts from the moment the call enters the system, and it ends when the ambulance arrives at the patient. The response time to hospital is mainly used for medical outcomes and is not further used in this paper. When we write response time in this paper, we always refer to the response time to incident.

The service time is the during which an ambulance is busy and cannot respond to a newly incoming incidents. This starts when the EMS team receives a notification from the dispatch center, and it stops when the ambulance arrives back on its base location. Note that some authors define the service time as the moment that the EMS vehicle becomes available at the hospital. The choice that ambulances become available at their base location guarantees that rural demand points at the region borders directly receive coverage when an ambulance becomes available again, since hospitals are most often located in urban areas.

3.2 Assumptions

Incidents only occur at so-called demand points. The number of demand points should be large enough to give an adequate representation of the region, but small enough to perform calculations within acceptable time. Postal code areas with several thousand inhabitants are a good candidate for a demand aggregation.

We have a fixed and bounded set of demand points, potential base locations, and hospitals in each ambulance region. These locations are all assumed to be known a priori. A variant of our models gives the model free choice of base locations: it makes every demand point a potential base location. Every demand point can be reached by at least one potential base location within the response time threshold. We assume that a region contains at least one hospital in the system.

Arrivals occur according to a Poisson process. For every demand point the frequency of incident arrivals is given. There is only one urgency class, and there is one type of ambulance that can handle all calls. The service time may depend on the incident location, base location, and ambulance allocation, and it is determinable and finite for all calls.

PTD, TMT, and TFT are assumed constant in this paper. These parameters have a finite value that is the same for every call. The driving times DTI, DTH and DTB are finite and deterministically determinable through a lookup table or by navigation software. Travel times are assumed to be symmetric, i.e. swapping origin and destination of one route has no effect on the driving times.

Once assigned to a base location, the ambulance starts and ends every service on its base location. There are no relocations of ambulances between base locations. Every hospital has an unlimited capacity and any patient can be brought to any hospital. When necessary one can include ramping, i.e. the waiting duration until the hospital's emergency department (ED) has free space, by including its mean time in the TFT.

Any call that does not receive immediate service from an available ambulance within the given response time standards is lost. There is no queue for waiting calls. In practice, a lost call is either served by a neighboring ambulance service provider or by another base location that has capacity.

For the Q-models the assumption is required that ambulances from base locations that have an overlap in demand points do not have major differences in call arrival rate, mean service time and minimal reliability level. This assumption is not required for the proposed AQ-models.

3.3 Variables

Let $\mathcal{H}$ be the finite set of hospital locations, let $\mathcal{I}$ be the finite set of demand points, and let $\mathcal{J}$ be the finite set of potential base locations, such that $\mathcal{H}, \mathcal{I}$, and $\mathcal{J}$ are disjunct sets. If a demand point is located at the same location as a hospital, we add two elements with the same coordinates; one for set $\mathcal{H}$ and one for $\mathcal{I}$. The three sets are not empty. Denote $\mathcal{V} := \mathcal{H} \cup \mathcal{I} \cup \mathcal{J}$.

Denote the shortest minimal driving time from point $k \in \mathcal{V}$ to point $\ell \in \mathcal{V}$ by $t_{k\ell}$. Recall that symmetric driving times are assumed, thus $t_{k\ell} = t_{\ell k}$. We denote the *response time to incident*

$$r_{ij} := \mathrm{PTD} + t_{ji}, \qquad \forall\ i \in \mathcal{I}, j \in \mathcal{J}. \tag{1}$$

If $r_{ij} \leq R$ we say that i is covered by j, for a constant system wide *response time threshold* $R \in \mathbb{R}_{\geq 0}$.

For each demand point $i \in \mathcal{I}$ we require a reliability level of at least $\alpha_i \in [0, 1]$, i.e. the probability that an ambulance is available to reach a patient within R time units must be at least α_i. A typical value one can take is $\alpha_i = \alpha = 0.95$ for all demand points $i \in \mathcal{I}$.

The total number of ambulances in the system is denoted by Z as this is the variable we want to minimize, and the variable $x_j \in \mathbb{R}_{\geq 0}$ denotes the utilized capacity of potential base location $j \in \mathcal{J}$. The utilized capacity of a potential base location is the number of ambulances that are allocated to this base.

For each demand point $i \in \mathcal{I}$, we denote the call arrival frequency by $f_i \in \mathbb{R}_{\geq 0}$. This is the number of calls per time unit that enters the system at this demand point.

The mean service time of a demand point is the average time an ambulance is busy to a call that takes place at that demand point and therefore is not available for dispatch to a new incident. If a patient at demand point $i \in \mathcal{I}$ is served by an ambulance that departs from base location $j \in \mathcal{J}$ and brought to hospital $h \in \mathcal{H}$, the mean service time for a call, β_{hij}, is defined by

$$\beta_{hij} := r_{ij} + t_{hi} + t_{hj} + TMT + TFT - CTT. \tag{2}$$

Denote the set of demand points that can be reached, i.e. have a response time to incident within R time units from $i \in \mathcal{I}$ by $\mathcal{N}_i := \{i' \in \mathcal{I} \mid r_{ii'} \leq R\}$, throughout referred to as the *neighborhood* of demand point i. We can interpret this as the set of all demand points that would be covered if there were a base location collocated to i. For the demand points near a base location we make use of the set $\mathcal{N}_j := \{i' \in \mathcal{I} \mid r_{ji'} \leq R\}$, i.e. the set of demand points that can be reached from base location j. Similarly for the base locations near demand points or base locations, define $\mathcal{M}_i := \{j' \in \mathcal{J} \mid r_{ij'} \leq R\}$.

3.4 The queuing approach

In Subsection 3.4.1 we compare the Erlang blocking approach used in the Q-models to the binomial approach that is used in PLSCP and MALP. The authors of the Q-models show with computational results that these models require less ambulances than PLSCP and MALP for the same coverage constraints. This section considers a theoretical point of view and discusses the advantages and disadvantages of both approaches, and conclude that generally the queuing approach is the better one when all its assumptions can be satisfied. Subsection 3.4.2 describes Q-PLSCP in detail, as we use this model in our illustrations.

3.4.1 Comparison between the binomial and queuing approaches

In literature we encounter two approaches to calculate blocking probabilities in facility location and allocation problems: the binomial and queuing approaches. In this subsection we compare the two approaches

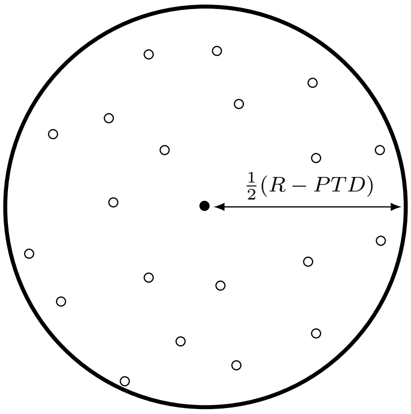

Fig. 2 Island model with only one base location j (●). Every demand point $i \in \mathcal{I}$ (○) can reach every other demand point within the response time norm.

for a toy example: the island model. Existing literature is rather limited on this subject. This comparison provides the motivation why we continue research on the queueing approach instead of the binomial.

Consider an island with one base location $j \in \mathcal{J}$, which covers the entire island, i.e. $|\mathcal{J}| = 1$. Particularly, this base location is not influenced by other base locations; see Figure 2. This means that $\mathcal{M}_i$ is the same singleton for all demand points $i \in \mathcal{I}$. Also the arrival frequency f_i for each demand point $i \in I$ is given, and thereby the total demand that must be served by ambulances that are stationed at base location j. Arrivals are independent and require service from one ambulance. The last assumption we make for this island is that every demand point can reach every other demand point within the response time to incident constraint, i.e. $\mathcal{N}_i = \mathcal{I}$ for all $i \in \mathcal{I}$. The island contains one hospital and differences between the travel times are assumed negligible compared with the mean service time of a call, hence we take $\beta_i = \beta$ constant. We assume that the reliability level requirement is constant: $\alpha_i = \alpha$ for all $i \in \mathcal{I}$.

We take a number of ambulances $Z = x_j$ at $j \in \mathcal{J}$ and calculate the blocking probability for each approach, under the assumption that any call that cannot be immediately assigned to an available ambulance is lost.

Binomial Approach The binomial approach uses a fixed busy fraction q for every ambulance, that may be based on historical recorded data. Under this assumption, the probability that an ambulance is not available at call arrival is q. Hence, the probability that none of the ambulances at this base is available is given by the following probability:

$$P_{Bin}(\text{No ambulance is available}) = q^{x_j}. \tag{3}$$

When determining the required number of ambulances such that a reliability level of at least α is met, we obtain a constraint of the form $1 - q^{x_j} \geq \alpha$. Taking $b^{Bin} := \log(1-\alpha)/\log(q)$ yields $x_j \geq b_i^{Bin}$. Thus we get constraints of the form

$$\sum_{j \in \mathcal{M}_i} x_j \geq b_i^{Bin}, \qquad \forall i \in \mathcal{I}. \tag{4}$$

The general assumption of the binomial approach is that the busy fraction of an ambulance is near-constant, and thereby that the number of available ambulances is a good approximation to handle the average ambulance busy fraction. When using the binomial approach for a realistic situation with multiple base locations an IP-formulation fits the number of ambulances to an acceptable workload.

Queueing Approach The queuing approach uses the Erlang B blocking formula for a M/G/c/c-loss system with Poisson arrivals, general service time distribution with finite mean and a server pool with a fixed number of servers, under the assumption that calls that cannot be directly served are lost. We denote the arrival rate for each demand point i as the sum of the frequencies of demand points in its neighborhood: $\lambda_i = \sum_{i' \in \mathcal{N}_i} f_{i'}$. Because of the small isolated area we consider λ_i equally valued for every demand point. We also need the mean service time β for an ambulance serving $i \in \mathcal{I}$.

Assuming Poisson arrivals with rate λ_i, a mean service time β, and $b_i^{Erl} \in \mathbb{N}$ ambulances in the vicinity of $i \in \mathcal{I}$, the Erlang B blocking function gives us the constraint for the minimal service level α_i for each $i \in \mathcal{I}$:

$$Erlang_B(\lambda_i, \beta, b_i^{Erl}) = \frac{\frac{(\lambda_i \beta)^{b_i^{Erl}}}{b_i^{Erl}!}}{\sum_{k=0}^{b_i^{Erl}} \frac{(\lambda_i \beta)^k}{k!}} \leq 1 - \alpha. \tag{5}$$

Knowing λ_i, β, and α for each $i \in \mathcal{I}$, we can determine the minimal value for b_i^{Erl} that is required to meet this constraint. In the island model b_i^{Erl} has the same value for all demand points. We should keep in mind that generally a per demand point β_i can depend on the hospital and base locations, and on the ambulance allocation $x_j, j \in \mathcal{J}$. The island model has $x_j = b_i = Z$ for all i, j. In Definition 2 of Section 5 we use another yet equivalent definition. Note the relation with the binomial approach: $q = q_i = \beta/(b_i \lambda_i)$.

This leads to the same family of constraints as the binomial approach, but with a different calculated value for b_i:

$$\sum_{j \in \mathcal{M}_i} x_j \geq b_i^{Erl}, \qquad \forall i \in \mathcal{I}. \tag{6}$$

Differences Between Approaches The main difference between the two approaches is that the queuing approach does not consider the busy fraction as an input parameter. By taking the busy fraction as an input parameter the binomial approach has an unwanted side effect. To show, we construct an example where the island has only one demand point. The busy fraction is kept fixed. In practice, when ambulances are added to the base location the busy fraction for each ambulance decreases as the workload gets shared between the ambulances. This does not happen in the binomial approach; instead we observe a strange effect. By keeping the busy fraction q fixed as the number of ambulances x increases, the binomial approach indirectly assumes that either the number of incoming calls λ or the mean service time β, or both, increase when more ambulances to a base location are added. This contradicts the fact that the number of incoming calls and mean service time are fixed input parameters. Especially at rural bases with low demand, such as one or two ambulances, adding an extra ambulance has a significant impact on the busy fraction of ambulances that are allocated at that base, which is not adequately incorporated in the binomial approach. This effect becomes smaller when more ambulances are allocated to the base locations. Therefore the binomial models provide credible results when many vehicles are required per base locations, that is, at highly populated areas.

When an ambulance region has both urban and rural areas the binomial approach may not be the best choice. In practice, in the case of a constant system wide reliability level, the busy fractions of ambulances positioned at a rural base location are much lower in contrast to ones at urban base locations. In the queuing approach this effect is incorporated through Erlang B. However, by design, the Q-models may not be applied to mixed regions because their demand may not significantly fluctuate. What happens if you apply the Q-models is discussed in detail in Section 5.

The binomial approach is a better choice in the case that the arrival process cannot be modeled by a Poisson process, or when every potential base location has many ambulances allocated such that adding or subtracting one ambulance does not have a significant impact on the busy fraction.

It is straightforward to change a binomial methods to their queuing counterpart, since they only differ in the way b_i, $i \in \mathcal{I}$, is calculated.

We conclude that Erlang blocking formulations are preferred over the binomial choice in maximum reliability and availability models when all model assumptions can be satisfied.

3.4.2 Queuing Probabilistic Location Set Coverage Problem

The Queueing Probabilistic Location Set Coverage Problem (Q-PLSCP) [?] is a maximal reliability model that minimizes the total number of ambulances in an ambulance region such that the minimal reliability level requirements are met for all demand points. This is done in two phases: the first calculates right hand side values for the constraints of a mixed integer program (MIP), and the second phase solves the resulting MIP.

This model needs an extra assumption, the so-called *isolation assumption*, such that a neighborhood can be considered as an isolated problem, which is a requirement for the application of Erlang B. Every neighborhood gets a number of ambulances $b_i \in \mathbb{N}, i \in \mathcal{I}$ assigned. In the case that all b_i ambulances in $\mathcal{N}_i$ are busy, ambulances from bordering neighborhoods can respond. Similarly, ambulances from bordering neighborhoods may receive assistance from ambulances of $\mathcal{N}_i$. Because there are minor differences between bordering neighborhoods in the call arrival rate, mean service time, and reliability level, the influx and outflux cancel out. Thus each individual neighborhood may be treated as an isolated area; this is the isolation assumption. The average number of assigned ambulances that is available or serving calls in $\mathcal{N}_i$ is on average close to b_i.

Independent Poisson arrivals are assumed for each demand point, with arrival rate $f_i \in \mathbb{R}_{\geq 0}$ for all $i \in V$. The service time duration is taken constant at β time units for all calls, and the reliability level α is a fixed system wide constant. The arrival rate in neighborhood $\mathcal{N}_i$ is given by $\lambda_i = \sum_{k \in \mathcal{N}_i} f_k$. Because values for λ_i, β, and α are determined or given, the Erlang B formula provides a lower bound on the number of required ambulances $b_i \in \mathbb{N}$ to $\mathcal{N}_i$:

$$b_i = \text{argmin}_{n \in \mathbb{N}_{\geq 0}} \{1 - \text{Erlang}_{\text{B}}(\lambda_i, \beta, n) \geq \alpha\}. \tag{7}$$

Due to symmetric travel times, the number of ambulances in neighborhood $\mathcal{N}_i$ equals the number of ambulances that can reach demand point $i \in \mathcal{I}$ from their base location. Thus it poses a constraint of the form $\sum_{j \in \mathcal{M}_i} x_j \geq b_i$ to every demand point, which concludes the first phase.

In the second phase an integer program is solved to ensure that every neighborhood gets coverage by at least the minimal number of required ambulances.

$$\begin{aligned} \min\ Z = & \sum_{j \in \mathcal{J}} x_j \\ \text{s.t.}\ & \sum_{j \in \mathcal{M}_i} x_j \geq b_i,\ \forall i \in \mathcal{I} \\ & y_i \in \mathbb{N}_{\geq 0}. \end{aligned}$$

After solving the IP-formulation x_j holds the number of ambulances allocated to base location $j \in \mathcal{J}$, and Z equals the total number of ambulances in the ambulance region.

The authors chose a service time of $\beta = 45$ minutes [?]. The frequency f_i is taken proportional to the population at $i \in \mathcal{I}$. Instead of a travel time constraint, they took an action radius from the base location of $\tilde{R} = 1.5$ miles. The tilde is added to stress the unit change from time to distance. Along the lines of earlier thoughts, this leads to a set of reachable demand points $\tilde{N}_i = \{i' \in V \mid dist(i, i') \leq \tilde{R}\}$. The total demand at neighborhood $\mathcal{N}_i$ then becomes $\lambda_i = \sum_{k \in \tilde{N}_i} f_k$.

The authors give the proof that the Erlang B approach works for exponential service times, although their result is just as valid for general service time with determinable expectation; see the Queueing Approach in Subsection 3.4.1. Their method is much stronger than suggested by the paper.

4 Motivation: the cause for over-estimation of current queuing approaches in mixed regions

It is generally agreed that maximum reliability and availability methods yield over-estimation if they are applied to mixed regions. The effect is mentioned in [**?**] at multiple occasions, and computational comparisons with other facility location methods that show the over-estimation can be found in [**?**,**?**] and computations in our result section. We must state that the Q-models are originally not designed to be used in mixed regions.

This section tells us what goes wrong if we do apply Q-models to a mixed region, against their original design, and provides motivation to the search direction of the solution that we pursue in the next three sections. In Section 5 and Section 6 we address and solve the problems found in this section. Section 7 combines these insights and proposes the adjusted queuing models, e.g. AQ-PLSCP. Section 8 provides computational results that show that the AQ solution indeed solved these over-estimations.

Recall that Q-PLSCP consist of two phases. The first phase calculates the values of b_i that hold a lower bound on the required number of ambulances in the neighborhood of $i \in \mathcal{I}$, and in the second phase the methods solve an IP-formulation that provides the ambulance allocation. Recall that the first phase is equal in both methods. We show that the over-estimation is mainly caused in the first phase for regions with varying demand.

We illustrate the over-estimation using two separate situations. In Subsection 4.1 we take a realistic situation and show why Q-PLSCP yields an over-estimation. Subsection 4.2 gives a theoretical example that demonstrates that the so-called *demand projection* effect can result in an over-estimation of any extent. The solution is presented in the next two sections.

4.1 Illustration for a real situation

In this subsection we highlight the small village of Watergang (just north of the city Amsterdam), a population of 405 people. Figure 3 illustrates that this approach results in too much ambulances that are required to cover most rural neighborhoods, up to five ambulances near Watergang. After applying our proposed adjusted queuing solution, we see that only one ambulance is sufficient for this village.

In our calculation, we took the average demand at each demand point between 10:00 and 12:00 AM on working days measured over a period of 5 years, with $\alpha_i = .95$ and β_i calculated from the nearest actual base location to the nearest actual hospital with an emergency department for each demand point $i \in \mathcal{I}$.

The source of these large numbers can be found in the approximation of the arrival rates λ_i rather than in the mean service times $\beta_i, i \in \mathcal{I}$.

If the current queuing approaches are used for this region, $\mathcal{N}_{Watergang}$ contains many demand points of the nearby cities of Amsterdam and Purmerend. Consequently, the arrival rate at Watergang's neighborhood $\lambda_{Watergang} = \sum_{k \in \mathcal{N}_{Watergang}} f_k$ is not representative for the demand point Watergang itself because the arrival rate of neighborhood Watergang is dominated by urban demand, and so is the value $b_{Watergang}$. Because Purmerend is the only base location that can provide coverage to Watergang, the IP-formulation in the second phase of Q-PLSCP shows that base Purmerend gets $x_{Purmerend} \geq b_{Watergang}$ ambulances allocated.

We have shown that base Purmerend provides coverage for non-existing demand that is projected from Amsterdam onto Watergang. We call this effect *demand projection.* In general, we see that the same effect

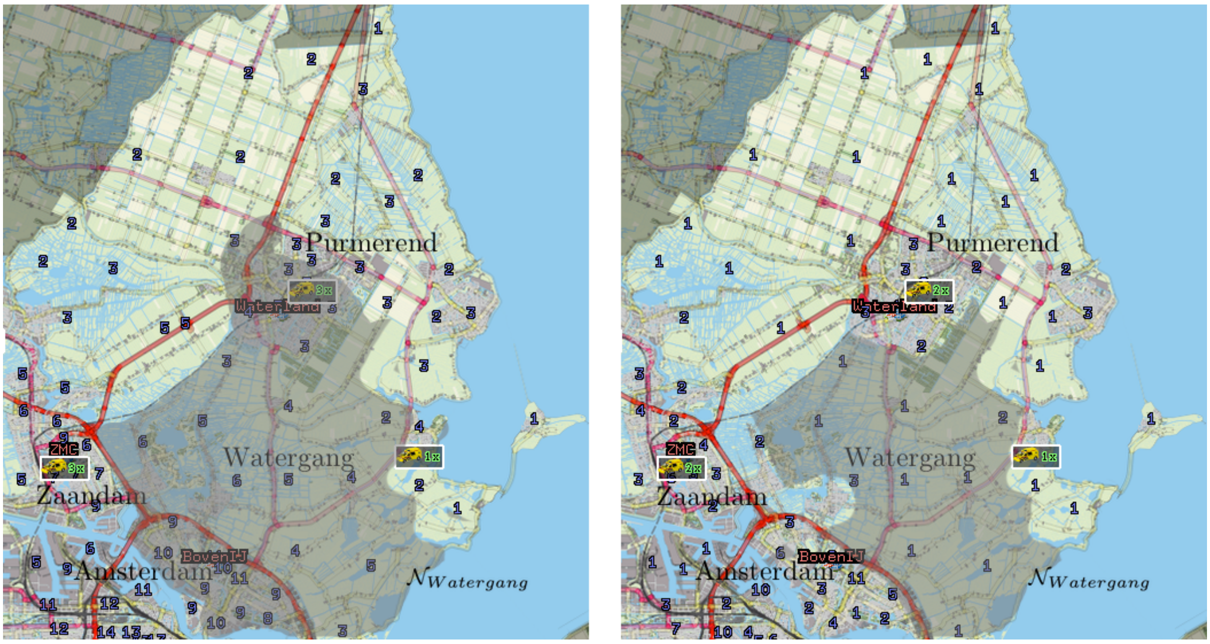

Fig. 3 Required coverage b_i for each demand point $i \in \mathcal{I}$ without (left) and with (right) frequency adjustment. The neighborhood $N_{Watergang}$ is shaded for both methods.

can occur for all rural base locations that have a driving time between R and $2R$ from a large enough city's border.

The underlying cause is that the isolation assumption says that there may no major fluctuations of demand between the region's neighborhoods when you apply Q-PLSCP. This is not realistic for ambulance regions in reality as demand fluctuations occur everywhere. The 'faulty' use of the Q-models to the mixed region in this illustration provides insight what to do next. To make the Q-models applicable to mixed regions, the isolation assumption of the Q-models needs to be replaced by a mathematical structure that allows for major fluctuations. This structure is the *workload condition* that is introduced in Section 5. Next, using this workload condition, we can redefine the way the arrival rate of in neighborhood is calculated such that demand projection cannot occur, see Section 6.

4.2 Theoretical example

This theoretical example shows that demand projection, being the over-estimation effect of rural areas, can be extended to any magnitude. We show this effect for Q-PLSCP.

To study this effect, we consider a one-dimensional scenario with five demand points, of which the first three, A, B, and C, are in an urban area and the latter two, D and E, are rural. At the locations of central urban point B and outer rural point E are potential base locations; see Figure 4.

Our focus during the remainder of this scenario lies on the arrival rates, therefor we take the service time and reliability level system wide constants: $\beta = \beta_i$ and $\alpha = \alpha_i$ for all $i \in \mathcal{I}$. Hence, the minimal required staffing b_i through Erlang B only depends on the arrival rates λ_i, i.e. $b_i = b_i(\lambda_i)$ for all $i \in \mathcal{I}$. The scenario has driving times $r_{AB} = r_{BC} = r_{DE} = R - PTD - \epsilon$ and $t_{CD} = 2\epsilon$ for a small enough constant $\epsilon \in \mathbb{R}_{>0}$, and an allowed response time threshold of $R \in \mathbb{R}_{>\epsilon}$ minutes.

The scenario is designed such that $\mathcal{M}_A = \mathcal{M}_B = \mathcal{M}_C = \{B\}$ and $\mathcal{M}_D = \mathcal{M}_E = \{E\}$.

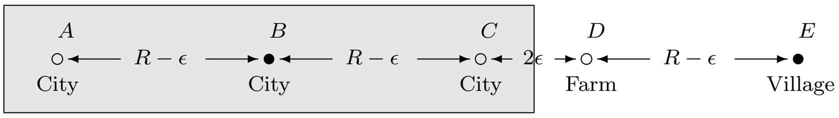

Fig. 4 Our theoretical region with urban and rural demand points. Filled points also have both demand and a base.

The substitution of $\lambda_i = \sum_{i' \in \mathcal{N}_i} f_{i'}$ yields $b_A = b_A(f_A + f_B)$, $b_B = b_B(f_A + f_B + f_C), b_C = b_C(f_B + f_C + f_D)$, and $b_D = b_D(f_C + f_D)$.

In a desirable situation Q-PLSCP allocates ambulances such that demand of A, B, and C is covered by base B, and demand of D and E is served by the ambulances at base location E. Particularly, we do not want to staff base location E for any urban demand.

This fails if $f_A = f_B = f_C = M$, $f_D = \epsilon'$, and $f_E = 1$ for small enough $\epsilon' \in \mathbb{R}_{>0}$ and a large enough $M \in \mathbb{R}_{>\epsilon'}$. After all, $x_E = \max(b_D(M+1+\epsilon'), b_E(1+\epsilon')) = b_D(M+1+\epsilon)$, which is approximately $b_D(M)$. However, we wish that our method returns a staffing of $b_D(1)$ for base location E. The *demand projection* of the urban demand point C onto demand point D results in an order of magnitude too high staffing at base location E, while demand point C is not even within reach of base location E. This is exactly the effect that causes major over-estimation in the Q-models for realistic regions.

Define the outer region $U_i \subseteq \mathcal{I}$ of demand point $i \in \mathcal{I}$ by the set of demand points that cannot be reached within time threshold R from i, but that can be reached within R time units by a demand point that can reach i within time threshold R, i.e. $\mathcal{U}_i := \{i' \in \mathcal{I} : \exists i'' \in \mathcal{I}, r_{i''i'} \leq R, r_{i''i} \leq R, r_{i'i} > R\}$. A good approximation of the outer region is the set of demand points that are contained in a torus with center i and a radius between R and $2R$. In our example $U_B = \{D\}$ and $U_E = \{C\}$. We define the outer region of a base location $\mathcal{U}_j \in \mathcal{I}$, $j \in \mathcal{J}$ similarly: $\mathcal{U}_j := \{i' \in \mathcal{I} : \exists i'' \in \mathcal{I}, r_{i''i'} \leq R, r_{i''j} \leq R, r_{i'j} > R\}$.

The number of ambulances allocated to a potential base location is only too high when the so-called density in the outer region of that base is significantly higher than any of the densities in the inner region $\mathcal{N}_i$ of that base. Various choices for the definition of the density of a demand point ψ_i can be made. For example: the density ψ_i of a demand point may be the population divided by the area that is mapped onto demand point i, or one may choose to define the density as the frequency f_i while assuming a constant area for each demand point. The example in this subsection uses the latter option, leading to $\psi_i = f_i$ for all $i \in \mathcal{I}$. In general we assume that demand points with a higher density also have a higher performance requirement: if $\psi_i \leq \psi_{i'}$ then $\alpha_i \leq \psi_{i'}$ for all $i, i' \in \mathcal{I}$.

This can also be seen as follows: Fix all variables in our example except for the frequency at demand point D. Observe the staffing at base location B: $x_B = \max(b_A(f_A + f_B), b_B(f_A + f_B + f_C), b_C(f_B + f_C + f_D)) = \max(b_A(2M), b_B(3M), b_C(2M + f_D)) = \max(b_B(3M), b_C(2M + f_D))$. Only when f_D exceeds M it can have an unwanted influence by unnecessarily increasing b_B.

We conclude this section with the following statements. Only if any demand point $i' \in \mathcal{I}$ in the neighborhood $\mathcal{N}_j$ of base location $j \in \mathcal{J}$ has a demand point $i'' \in \mathcal{N}_{i'}$ in its own neighborhood that is not in the neighborhood of base j itself, and if the density $\psi_{i''}$ is significantly higher than the density ψ_i of any demand point $i \in \mathcal{N}_j$ in the neighborhood of j, an overstaffing at base location $j \in \mathcal{J}$ is achieved through demand projection. Consequence 1: if the density of any demand point in $\mathcal{U}_j$ significantly exceeds the density of all demand points in $\mathcal{N}_j$, then base $j \in \mathcal{J}$ gets overstaffed through demand projection. Consequence 2: if all demand points in $\mathcal{N}_j$ have at least the density of all demand points in $\mathcal{U}_j$, there will be no overstaffing at base $j \in \mathcal{J}$.

4.3 Some words on differentiating the reliability level in the Q-models

It the Q-models it is not trivial to implement a differentiation of the reliability level, that is, having a parameter α_i, $i \in \mathcal{I}$, that significantly differs for bordering neighborhoods. The reason is that the isolation assumption of the Q-models loses its validity when one does.

Key element is the Erlang blocking formula for an M/G/s-loss system Erlang$(\lambda, \beta, b) \leq \alpha$ where b represents the number of servers (ambulances). Recall that in these models λ, β and α are input parameters, and the minimal value b for which the inequality is satisfied is the output parameter. For the isolation criteria to hold, the inflow and outflow over a neighborhood's border to adjacent neighborhoods must cancel out. Hence, both λ, β and α should be treated equally. If one of the three fluctuates while the other two are kept stable, it results in a non-zero flux over a neigborhood border. Thus α should also be stable across bordering neighborhoods for the isolation criteria to hold.

More in line with literature, likewise to the call arrival rate, one could say in Q-PLSCP and Q-MALP that one allows for α_i demand point dependent where α_i does not differ to a significant extent from the reliability levels of the neighborhoods that border i. Such a small difference may not be manageable from an administrative point of view. In literature we see that α does not vary over i [**?**,**?**].

4.4 Concluding: necessary changes to the Q-methods for application to mixed regions

There are two challenges that must be addressed and solved to allow MR-MA models to be applied on ambulance regions with both urban and rural areas:

Challange 1 The isolation assumption for a neighborhood should be generalized and made more explicit. In areas with both urban and rural demand the isolated neighborhood assumption is not realistic, in particular, the case when an urban located base location covers rural areas near the city that has no closer base locations, or when larger base locations are located near a neighborhood's border. A generalization of this condition is handled in Section 5.

Challange 2 We need a new neighborhood definition such that demand projection cannot happen. The solution, and the resulting adjusted Q-PLSCP formulation is discussed in Section 6.

5 Solution to Challange 1: generalization of the isolation assumption

Section 4 showed that the demand projection effect can cause over-estimation on the required numbers of ambulances to any extent when Q-PLSCP is applied to a mixed region. This section discusses an approach that can replace the isolation assumption that was introduced in Section 3.4: the *workload condition.* Another advantage of the workload condition is that it allows for a demand point dependent reliability level requirement α_i instead of a system wide constant.

From a theoretical point of view, not necessarily all ambulances at $\mathcal{M}_i$ must be able to serve demand at $i \in \mathcal{I}$. Recall that the concept of neighborhoods $\mathcal{N}_i$ is only used to get a notion of the workload if i would be the only base location within a response time radius of at most R.

Definition 1 We introduce the following notations:
a) Denote the set of ambulances in the system by $\mathcal{A}$.
b) Denote the set of ambulances that serve demand at $i \in \mathcal{I}$ by $\mathcal{A}_i \subseteq \mathcal{A}$.

c) Denote the number of ambulances that serve demand at $i \in \mathcal{I}$ by $n_i = |\mathcal{A}_i|$.
d) Denote the number of ambulances that are positioned at bases of $\mathcal{M}_i$ by $y_i = |\mathcal{A}_i|$.
e) Denote the number of ambulances that can reach demand point $i \in \mathcal{I}$ within response time R by x_i.

Note that in Q-PLSCP solving the IP-formulation yields values x_j, $\forall j \in \mathcal{J}$, and equation $b_i \leq y_i = \sum_{j \in \mathcal{N}_i} x_j$ holds for all demand points $i \in \mathcal{I}$.

It is not necessary that each ambulance for which $t_{ai} \leq R$ holds is an element of $\mathcal{A}_i$. This makes our adjustment also possible for Q-MALP.

5.1 Bounds on the busy fractions

In the remainder of the paper the (offered) workload that is generated in a neighborhood, and busy fraction of the ambulances play a central role.

Definition 2 **a)** Define for arrival rate λ and mean service time β the workload

$$\rho := \lambda\beta.$$

Similarly, the workload at demand point $i \in \mathcal{I}$ is denoted by $\rho_i = \lambda_i\beta_i$.

b) For the remainder of this paper we redefine Erlang B as a function of ρ and b by substituting $\rho = \lambda\beta$, which is equivalent to Equation 5:

$$Erlang_B(\rho, b) = \frac{\frac{\rho^b}{b!}}{\sum_{k=0}^{b} \frac{\rho^k}{k!}} \leq 1 - \alpha.$$

Similarly, we get $Erlang_B(\rho_i, b_i) = \frac{\rho^{b_i}}{b_i!} / \sum_{k=0}^{b_i} \frac{\rho^k}{k!} \leq 1 - \alpha_i$ for $i \in \mathcal{I}$.

We define the bounds on the busy fraction of an ambulance, and in Section 5.2.2 we illustrate how they are used. A starred notation $(*)$ refers to the solution of the problem that we found, and hence depends on n_i; this is not necessarily a global optimum.

Definition 3 Consider an independent system with Poisson arrivals, workload ρ, and a fixed reliability level α.

a) Define the *minimal required number of ambulances* at workload ρ by

$$b(\rho) := \text{argmin}_{b' \in \mathbb{N}}\{Erlang_B(\rho, b') \leq 1 - \alpha\}.$$

For a neighborhood $\mathcal{N}_i$, $i \in \mathcal{I}$, we have a demand point dependent α_i and define it as

$$b_i(\rho_i) := \text{argmin}_{b' \in \mathbb{N}}\{Erlang_B(\rho_i, b') \leq 1 - \alpha_i\}.$$

b) Define the *lower bound on the busy fraction per ambulance* with $n \in \mathbb{N}_{>0}$ serving ambulances by

$$\Psi^{low}(\rho, n) = \rho/n.$$

Furthermore, define $\Psi^{low}(\rho) := \rho/b(\rho)$ and $\Psi_i^{*,low}(\rho) := \rho/n_i$ for $i \in \mathcal{I}$ where we have $n_i = |\mathcal{A}_i|$ ambulances that serve demand point i.
Denote $\Psi_i^{low} := \Psi^{low}(\rho_i)$, and $\Psi_i^{*,low} := \Psi^{*,low}(\rho_i, n_i)$ for $i \in \mathcal{I}$.
c) Define the *upper bound on the busy fraction per ambulance* for a system with offered workload ρ by

$$\Psi^{upp}(\rho) := \sup_{\rho'}(Erlang_B(\rho', b(\rho)) \leq 1 - \alpha)/b(\rho).$$

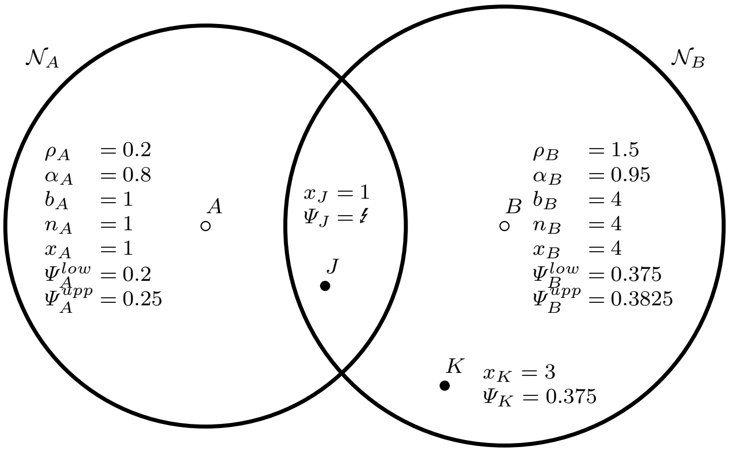

Fig. 5 A counterexample for a region where rural neighborhood $\mathcal{N}_A$ and urban $\mathcal{N}_B$ share one ambulance at J, which cannot fulfill the workload conditions of both neighborhoods simultaneously.

Consequently, the *upper bound on the busy fraction per ambulance for a neighborhood* $\mathcal{N}_i, i \in \mathcal{I}$, is the supremum on the busy fraction serving each ambulance in $\mathcal{N}_i$ such that any more workload leads to an additional ambulance:

$$\begin{aligned}\Psi_i^{upp} &:= \sup_{\rho'}(Erlang_B(\rho', b_i) \leq 1 - \alpha)/b_i,\\ \Psi_i^{*,upp} &:= \sup_{\rho'}(Erlang_B(\rho', n_i) \leq 1 - \alpha_i)/n_i.\end{aligned}$$

Denote the corresponding arguments by ρ_i^{upp} and $\rho_i^{*,upp}$, respectively.

5.2 The basic idea

The basic idea behind the workload condition consists of two insights.

1. **Handling Overcapacity** We staff for an isolated neighborhood $i \in \mathcal{I}$ with workload ρ_i and reliability level requirement α_i. This yields a minimal number of ambulances b_i required in the neighborhood $\mathcal{N}_i$. We do realize that due to b_i being an integer we have a slight overcapacity of ambulances in the neighborhood. This overcapacity can be used to serve outside neighborhood i.
2. **Workload Condition** Neighborhoods share ambulances, while an ambulance $a \in A$ may only have one busy fraction at which it operates. Ambulances can only serve neighborhoods if these neighborhoods' have no conflicting constraints on the ambulance's busy fraction. We formalize this concept later in this section.

5.2.1 Clarification by a counterexample

We clarify the need of the various bounds on the busy fraction through a counterexample where an assignment through Q-PLSCP with demand point dependent reliability level requirements fails in the case that demand between adjacent neighborhoods differs significantly; see Figure 5. Consider an ambulance region with two demand points A and B, with $\rho_A = 0.2$, $\rho_B = 1.5$, $\alpha_A = 0.8$, and $\alpha_B = 0.95$. When we assume $\mathcal{N}_A$

to be an isolated region, $Erlang_B$ yields $b_A = 1$. Because $Erlang_B(\rho = 0.25, b = 1) = 0.2$ we have $\Psi_A^{low} = 0.2/1 = 0.2$ and $\Psi_A^{upp} = 0.25/1 = 0.25$. Likewise, we have $b_B = 4$ and $Erlang_B(\rho = 1.53, b = 4) = 0.05$, hence $\Psi_B^{low} = 1.5/4 = 0.375$ and $\Psi_A^{upp} = 1.53/4 = 0.3825$. Figure 6 shows this behavior. If $y_i = n_i = b_i$ for all $i \in \mathcal{I}$, there is a slight overcapacity of workload $O_A = (0.25 - .2)1 = 0.05$ Erlang in place for $\mathcal{N}_A$, and $\mathcal{N}_B$ has an overcapacity of $O_B = (0.3825 - .375)4 = 0.03$ Erlang.

Let us now consider two base locations J and K, such that $\mathcal{M}_A = \{J\}$ and $\mathcal{M}_B = \{J, K\}$. An optimal solution to the IP is $x_J = 1$ and $x_K = 3$. If we focus on the ambulance at J we see that neighborhood $\mathcal{N}_A$ says that its workload may not exceed $\Psi_A^{upp} = 0.25$, because any extra workload yields the need of an additional ambulance at $\mathcal{M}_A$ to keep the guarantee that the reliability level is at least $\alpha_A = 0.8$. On the other hand, neighborhood $\mathcal{N}_B$ requests from each of its ambulances to handle a workload of at least $\Psi_B^{low} = 0.375$. If the busy fraction of an ambulance goes below this value we cannot guarantee the reliability level. We see that the two neighborhoods have contradicting requirements for the workload of this ambulance.

This illustrates why an extra condition on the workload for a valid ambulance allocation is required.

5.2.2 Notion behind the various busy fractions

The counterexample that an ambulance allocation *cannot* guarantee the reliability level α_i for demand point $i \in \mathcal{I}$ if there is another demand point $i' \in \mathcal{I}$, such that:

1. either $\Psi_i^{*,low} > \Psi_{i'}^{*,upp}$ or $\Psi_{i'}^{*,low} > \Psi_i^{*,upp}$, or both, and;
2. demand points i and i' share ambulances.

We adapt the queuing method in such a way that this cannot occur, which is a step forward in our quest to replace the isolation assumption by a more general structure. From the negation of this statement we draw a hypothesis. Take any ambulance $a \in \mathcal{A}$ at random. If $\Psi_{i_1}^{*,low} \leq \Psi_{i_2}^{*,upp}$ holds for all combinations

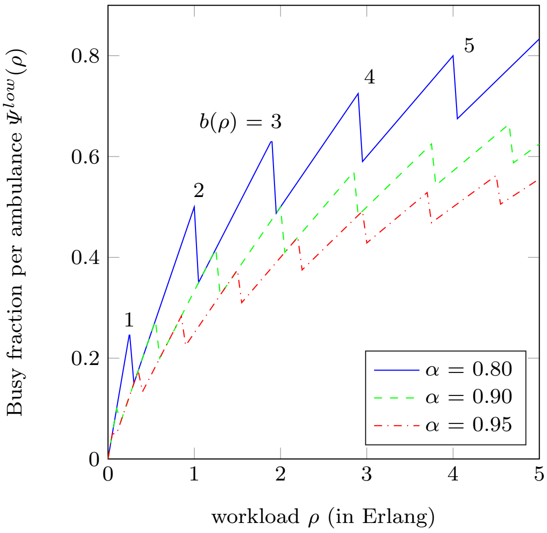

Fig. 6 Lower bound on the busy fraction per ambulance $\Psi^{low}(\rho)$ for given workload $\rho = \lambda\beta$ such that the minimal reliability constraint α is met with a minimal number of ambulances b induced by Erlang B.

of demand points $i_1, i_2 \in \mathcal{I}$ that a serves, then we may be able to guarantee the reliability level α_i for demand point $i \in \mathcal{I}$.

This is only the case if there is a variable Ψ_a^{dum} such that $\Psi_i^{*,low} \leq \Psi_a^{dum} \leq \Psi_i^{*,upp}$ holds for all demand points i that a serves. We call Ψ_a^{dum} the ambulance's dummy busy fraction. It can be easily shown that the dummy busy fraction is an upper bound to the actual busy fraction that the ambulance gets using the allocation that follows from the solution. (To proof, add *only for ambulance* a, a minimal dummy demand of the type 'keep on waiting' to all demand points that a serves until all these demand points have a similar lower bound on the busy fraction per ambulance.)

The overcapacity $O_i := (\Psi_i^{*,upp} - \Psi_i^{*,low})n_i$ may be used to serve calls outside $\mathcal{N}_i$.

5.2.3 Widening a demand points acceptance gap

We show that when enough ambulances are added we can fulfil the condition from the hypothesis for any region. After applying the Q-model's IP to a possibly homogeneous region, the requirement on the various busy fractions is usually not fulfilled. We show that it is possible for any ambulance region that this 'workload condition' gets fulfilled when enough ambulances are strategically added to the base locations, on top of the allocation we already obtained from the IP. This lays the base for the proposed adjusted queuing approach.

When n_i increases, $i \in \mathcal{I}$, both the lower and upper bounds on the busy fractions for the ambulances in neighborhood $\mathcal{N}_i$ change: the lower bound on the busy fraction decreases in n_i, while the maximum workload per ambulance increases in n_i.

Proposition 1 *For $b \leq n$, the following two conditions hold for any $\rho \geq 0$ and any fixed $\alpha \in [0,1]$.*
a) $\Psi^{low}(\rho, n) \leq \Psi^{low}(\rho, b)$. *Equality holds if and only if $n = b$.*
b) $\Psi^{upp}(\rho, n) \geq \Psi^{upp}(\rho, b)$. *Equality holds if and only if $n = b$.*

Hence the gap between $\Psi^{low}(\rho, n)$ and $\Psi^{upp}(\rho, n)$ widens if n increases from two sides: both the lower bound decreases and the upper bound increases. For $\Psi_i^{*,low}$ and $\Psi_i^{*,upp}$ we call this the *acceptance gap.* As a result, once a neighborhood is covered, it stays covered when additional ambulances are included in $\mathcal{A}_i$.

As n increases, the lower bound on the busy fraction for each ambulance goes to zero, and the upper bound exceeds one. This is an important insight that is required to proof that the proposed adjusted queuing method always leads to a solution.

Theorem 1 *For any $i \in \mathcal{I}$, $\rho_i > 0$, $\alpha_i \in [0,1]$ for $n_i \to \infty$, we have*
(a) $\Psi_i^{*,low} \downarrow 0$.
(b) $\Psi_i^{*,upp} \uparrow \frac{1}{1-\alpha_i} > 1$.

Theorem 1 states that for every $i \in \mathcal{I}$ and fixed value $\Psi_a^{dum} \in (0,1)$ for all $a \in \mathcal{A}$ there exists a finite number $\Delta n_i \geq 0$ such that allocating an additional number of at least Δn_i demand points i serving ambulances at $\mathcal{M}_i$, the following holds for each $a \in \mathcal{A}_i$: $\Psi_i^{*,low} \geq \Psi_a^{dum} \geq \Psi_i^{*,upp}$. Recall that $\Psi_i^{*,low}$ and $\Psi_i^{*,upp}$ depend on Δn_i. This means for PLSCP-models that adding enough extra ambulances at base locations in a somewhat smart fashion always leads to a feasible result; Section 5.4.1 shows an example.

5.3 Workload condition

This section formalizes the hypotheses from Section 5.2. In the adjusted queuing methods we replace the isolation assumption of the Q-methods by the *workload condition.*

Theorem 2 Workload condition *The reliability level α_i is guaranteed for every demand point $i \in \mathcal{I}$ if there exists an assignment of dummy variable Ψ_a^{dum} for every $a \in \mathcal{A}$ and an assignment $\mathcal{A}_i \subseteq \mathcal{A}$ for every $i \in \mathcal{I}$ such that all these conditions hold:*

$$\Psi_i^{*,low} \leq \Psi_a^{dum} \leq \Psi_i^{*,upp}. \tag{8}$$

Hence, the workload condition provides inspiration for a new definition of coverage for a demand point. Using this definition, every demand point is covered if and only if the workload condition holds.

Definition 4 Coverage of a demand point Given a set of ambulances $\mathcal{A}$, and a coupling between a demand point and a subset of ambulances $\mathcal{A}_i$ for $i \in \mathcal{I}$, such that every $a \in \mathcal{A}_i$ is stationed at any $j \in \mathcal{M}_i$ and has a fixed effective workload Ψ_a^{dum}. We say that demand point i is *covered* if $\Psi_i^{*,low} \leq \Psi_a^{dum} \leq \Psi_i^{*,upp}$ $\forall a \in \mathcal{A}_i$ holds.

Theorem 3 *If the isolation assumption holds for an allocation through a Q-method, then the workload condition is satisfied.*

Theorem 3 shows that the workload condition is a generalization of the isolation assumption that Q-PLSCP of Section 3.4.2 uses.

Corollary 1 *The demand point coverage in Definition 4 is a generalization for the definition of coverage found in the Q-methods.*

The workload condition is a sufficient condition, not a necessary one. Using the actual average workloads of ambulances it may be possible to construct an example where the reliability level is guaranteed for all demand points, while the workload condition is not satisfied. We leave the construction of a counterexample open for future research.

5.4 A post-processor to meet the workload condition

Recall that we replace the isolation assumption of the queuing assumption by the workload condition, and drop the assumption that the demand between neighborhoods does not vary much over space. In that case the allocation through the IP-formulation does not necessarily satisfy the workload condition.

A post-processor is required to adjust the solution in the slightest possible extent, such that the workload condition gets respected. In the adjusted queuing formulation we use a basic post-processor.

In the case that a solution of a Q-method does not respect the workload condition, we demand a higher minimal coverage by stating $b_i \leftarrow b_i + 1$ for each neighborhood that cannot satisfy the condition, and solve the IP-formulation again.

There are alternative approaches possible:
1) Find an alternate optimum to the IP-formulation for which the workload condition holds. In the counterexample $x_J = 4$ and $x_K = 0$ would be such a solution with the same objective value $Z = 4$, while all constraints are still respected.
2) Put an extra ambulance on a base location in a neighborhood where the workload condition is not satisfied.
3) It is an interesting topic for further research to see if it is possible to adapt the IP-formulation such that the workload condition always holds.

Note: In this paper we do not focus on making an optimal post-processor, and we settle on a basic post-processor for PLSCP that provides a solution that satisfies the workload condition to show an improvement on the existing papers.

5.4.1 A basic post-processor for maximal reliability models

For our result section we use a basic post-processor. This post-processor assumes that every used potential ambulance within the response time threshold must be able to respond to a call. It also assumes that the dummy busy fraction of ambulances that are stationed at the same base location is always the same. It prefers to add ambulances to base locations where it has the most added value.

Algorithm 1 Basic post-processor for maximum reliability models

```
 1: repeat
 2:     for all $i \in \mathcal{I}$ do
 3:         Update $\Psi_i^{*,low}$ and $\Psi_i^{*,upp}$.
 4:     end for
 5:     for all $j \in \mathcal{J}$ do
 6:         $\Psi_j^{*,upp} \leftarrow \min\{\Psi_i^{*,upp} : j \in \mathcal{M}_i \text{ and } i \in \mathcal{I}\}$
 7:     end for
 8:     for all $j \in \mathcal{J}$ do
 9:         $m_j \leftarrow 0$                      ▷ Uncovered demand $m_j$ through workload condition
10:         $m^{*,upp} \leftarrow 0$                 ▷ And $m^{*,upp}$ holds the maximal uncovered demand
11:         $w \leftarrow$ undefined                 ▷ Base location $w$ has the highest uncovered workload
12:         for all $i \in \mathcal{I}$ do
13:             if $j \in \mathcal{M}_i$ and $\Psi_i^{*,low} > \Psi_j^{*,upp}$ and $n_j > 0$ then
14:                 $m_j \leftarrow m_j + f_i$
15:                 if $m_j > m^{*,upp}$ then
16:                     $w \leftarrow j$             ▷ $j$ has the most uncovered demand so far
17:                     $m^{*,upp} = m_j$
18:                 end if
19:             end if
20:         end for
21:     end for
22:     if $w \neq$ undefined then
23:         $x_w \leftarrow x_w + 1$                 ▷ Add an additional vehicle at $w$ and repeat this procedure.
24:     end if
25: until $w =$ undefined                        ▷ If $w \neq$ undefined, the workload condition is not satisfied.
```

It is not hard to see that the basic post-processor satisfies the workload condition; take $\Psi_a^{dum} = \Psi_j^{*,upp}$ if a is stationed at j and note that Line 13 in combination with the stop condition $w = undefined$ guarantees $\Psi_i^{*,low} \leq \Psi_j^{*,upp} (\leq \Psi_i^{*,upp})$ for all $i \in \mathcal{I}$. Note that $n_i = y_i$ holds.

A property of this post-processor is that either all ambulances of a base location cover a demand point, or none do.

6 Solution to Challenge 2: density dependent demand aggregation

Section 4 illustrated the cause of over-estimation for the required number of ambulances in each neighborhood for realistic ambulance regions. Section 5 proposed the workload condition as a generalization of the isolation assumption; see Theorem 3. This section proposes a solution that does not contain the demand projection effect.

Our solution to the over-estimations lies in the way both arrival rate λ_i and mean service time β_i are approximated for each demand point $i \in \mathcal{I}$, and we propose an alternative calculation method for these variables. The next section contains the adjusted queuing variants of the Q-PLSCP, the so-called adjusted queuing models AQ-PLSCP, that use the findings of this section as an input.

Section 8 compares results from the Q-models with the adjusted queuing models for four actual ambulance regions to show the extent of the improvements.

The adjusted queuing approach uses another calculation methodology for the three Erlang B input parameters of demand point i's neighborhood, $i \in \mathcal{I}$: arrival rate λ_i, mean service time β_i, and reliability level α_i. We discuss these input parameters one at a time.

The second change is that a post-processing phase guarantees that the reliability level requirement is met for every demand point; see Section 5.4.

The basic difference is that the reliability level α_i is a demand-point-based fixed input parameter instead of a system wide constant. The approximation of the arrival rate and mean expected service time are discussed in Subsection 6.1 and Subsection 6.2, respectively.

6.1 Arrival rate

In the adjusted queuing approach we calculate λ_i in a different way compared to the Q-models, and add a post-processing phase that guarantees a valid solution. Directly after proposing our adjustments we provide theorems that prove that they work.

Method The method consists of an initialization, solving an IP-formulation and finally a post-processing phase.

Initialization To each demand point $i \in \mathcal{I}$ we assign a fixed density value $\psi_i \in \mathbb{R}$ that reflects an approximation of the fraction of total workload we can expect in neighborhood $\mathcal{N}_i$. Various choices for the density ψ_i can be made. In this paper we use the historical arrival frequency $\psi_i = f_i$ as our primary density measure (FAQ-methods), another choice is to take the population per square meter $\psi_i = C \cdot population_{\mathcal{N}_i}/area_{\mathcal{N}_i}$ (DAQ-methods) for some system wide constant $C > 0$.

The adjusted queuing approach redefines the neighborhood $\mathcal{N}_i^{AQ}$ by only including demand points that have at most the same density value of demand point i:

$$\mathcal{N}_i^{AQ} := \{i' \in \mathcal{I} : t_{ii'} \leq R \text{ and } \psi_{i'} \leq \psi_i\}. \tag{9}$$

The arrival rate λ_i of neighborhood $\mathcal{N}_i$ is obtained by summation over its demand points of at least the same density:

$$\lambda_i^{AQ} := \sum_{k \in \mathcal{N}_i^{AQ}} f_k = \sum_{\substack{k \in \mathcal{N}_i^Q \\ \psi_k \leq \psi_i}} f_k. \tag{10}$$

The neighborhood of the Q-models is denoted by $\mathcal{N}_i^Q$. Note that $N_i^{AQ} \subseteq N_i^Q$. The general concept of neighborhood is denoted by $\mathcal{N}_i$, and it will not be used in mathematical formulations from now on because it has become ambiguous; instead we use $\mathcal{N}_i^Q$ or $\mathcal{N}_i^{AQ}$.

IP-formulation Similar to literature, Erlang B yields $b_i^{AQ} := \text{argmin}_{b' \in \mathbb{N}}(Erlang_B(\lambda^{AQ}, \beta^{AQ}, b') \leq 1 - \alpha_i)$, which takes the places of b_i in the IP-formalation of the Q-methods.

Post-processing Through local search and putting additional ambulances at base locations, we make sure that the *workload condition* of Theorem 2 holds. There are multiple possible options for a post-processor; we use Algorithm 1 for generating results in this paper.

Section 5.4 shares some thoughts about alternative post-processing methods and concludes that finding the optimal post-processor remains an interesting topic for further research. Results for four actual ambulance regions are shown in Section 8.

Correctness of adjusted queuing methods Theorem 4 states that the new adjusted neighborhood definition for demand aggregation may be used in the same manner as the ones in the Q-methods.

Theorem 4 Correctness of the AQ-methods *The ambulance allocation using the AQ-method's neighborhood definition guarantees reliability level α_i for each demand point $i \in \mathcal{I}$ that is covered.*

The next theorem states that the number ambulances with the adjusted queuing neighborhood definition will never exceed the number of ambulances using the Q-methods.

Theorem 5 *The required number of ambulances through AQ-PLSCP, being Z^{AQ}, is at most the number of required ambulances Z^{Q} through Q-PLSCP for constant $\alpha_i = \alpha$ over all demand points $i \in \mathcal{I}$ if the solution of the IP-formulation respects the workload condition.*

Some final words on the choice of density values: FAQ vs. DAQ Choosing other orders than FAQ have two implications: 1) Because demand points with higher densities are counted multiple times, we see that the mean demand over the area increases leading to a higher number of vehicles. Of course this is a valid bound to reach a reliability level of α, but our proposed ordering yields a sharper bound. In this argument we used the symmetric travel times assumption. 2) Our ordering guarantees that demand concentrates at the demand point of interest. Any other ordering leads to demand projection to some extent.

DAQ projects demand to some extent to regions where the population density is relatively high in relation to the historic incidents. This choice leads to more conservative solutions in the case of demographic changes due to aging population.

6.2 Expected service duration and iterative process

The last remaining item to discuss is a fixed-point method to calculate the mean service time β_i of an ambulance covering $i \in \mathcal{I}$. Every demand point $i \in \mathcal{I}$ may have multiple base locations by which it can be reached, i.e. $|\mathcal{M}_i| > 1$. It is reasonable to assume that ambulances that are allocated to the same base location have the same mean service time.

Differences in the mean service time between neighboring base locations are not that large due to the fact that PTD, TMT, and TFT are system wide constants, and the differences in driving time are relatively small compared to the mean service time. Moreover, when we increase the value of the service time it leads to conservative solutions because we rather allocate more ambulances than less, i.e. a slight over-estimation of β_i is allowed to honor the reliability constraint, in contrast to an underestimation that may break this constraint.

We take a demand point $i \in \mathcal{I}$ at random and describe how we calculate its mean service time β_i. The values of β_i are calculated iteratively where we alternate the update of the allocation by solving the IP-formulation and values for β_i. During initialization the ambulances are allocated at random. We use a clear overcapacity in FAQ-PLSCP, e.g. we put 1000 ambulances on a single base. Recall our assumption that base locations are reasonably located and demand points are most likely to be served by the nearest ambulance base location. Hence, we assume that the response time is dominated by the travel time from the nearest base location that has at least one ambulance stationed.

Consider demand at this demand point and for now ignore that ambulances are shared with other demand points. We can approximate the mean service time of an ambulance serving exclusively this demand point i and thereby ignoring the fact that the service is influenced by all other demand points. The contribution of this exclusive mean service time $\tilde{\beta}_i$ to the mean service time of an ambulance serving demand point i can be approximated by the time from the nearest base location with at least one ambulance allocated to the nearest hospital. This value can be seen as the contribution of the demand point to the mean service time $\tilde{\beta}_j$ of an ambulance at this base location $j \in \mathcal{J}$. We have

$$\tilde{\beta}_i = \min_{h \in \mathcal{H}, j \in \mathcal{J}, x_j > 0} \beta_{hij}. \tag{11}$$

Recall that the ambulances are the servers in our Erlang B approach, and the service time depends on the base locations and the vertices they serve. For every base location, we now have to find a reasonable approximation of the mean service time by taking the weighted average mean service time over all exclusive mean service times in the base's neighborhood $\mathcal{N}_j$.

$$\tilde{\beta}_j = \frac{\sum_{i' \in \mathcal{N}_j} d_{i'} \tilde{\beta}_{i'}}{\sum_{i \in \mathcal{N}_j} d_{i'}}. \tag{12}$$

We do not use the frequency adjustment in this calculation, since the allocation by the IP-formulation assumes that this base location can service any demand that it covers. Also recall the previous discussion where we concluded that we may slightly overestimate mean service times. A corollary is that the mean service time of base locations with a large overlap in demand points is limited.

Because base locations with an ambulance stationed in the vicinity of i are more likely to cover demand points with a large response time from this base j than the demand point close to j, this $\tilde{\beta}_j$ is conservative. Here we used the assumption that base locations are reasonably spread over the region.

Staying conservative, for the mean service time of demand point i we can take the largest value $\tilde{\beta}_j$ that can be reached from the base location. Most likely this is a sharp upper bound for the actual mean service time of ambulances that serve i. We have

$$\beta_i = \max_{j \in \mathcal{M}_i} \tilde{\beta}_j. \tag{13}$$

With FAQ-PLSCP our purpose is to minimize the number of ambulances such that every demand point receives the required coverage. We stop the iterations if the total number of ambulances in successive iterations is the same.

Relation to Q-models Q-models take $\beta_i = \beta$ as a system wide constant [**?**,**?**]. Also, a statement can be found about taking β_i equal to the average amount of work on the fictional server in the neighborhood $\mathcal{N}_i$, effectively saying

$$\beta_i = \frac{\sum_{i' \in \mathcal{N}_i} \min_{h \in \mathcal{H}} d_{i'} \beta_{hi'i}}{\sum_{i' \in N_{i'}} d_{i'}}. \tag{14}$$

Our method improves on these methods by giving the actual server, i.e. the ambulance(s) on a staffed ambulance base, a central role in the calculation of β_i.

7 The solution to our problem: the adjusted queuing approach

The adjusted queuing approach consists of applying two transformations of the queuing approach. First we replace the neighborhood definition $\mathcal{N}_i^Q$ by the adjusted queuing version $\mathcal{N}_i^{AQ}$ of Section 6 through the density dependent demand aggregation for all demand point's $i \in \mathcal{I}$. Second, we make sure that the workload condition of Section 5 is satisfied by applying a post-processor.

7.1 Model formulation for AQ-PLSCP

For the adjusted queuing probability location set coverage problem (AQ-PLSCP) we provide the model formulation after applying these two transformations. We introduce an iterative process to obtain a realistic approximation of the mean service time for each demand point. Taking $\psi_i = f_i$ the arrival frequency leads to the FAQ-PLSCP, while taking $\psi_i = d_i$ the population density leads to DAQ-PLSCP (for all $i \in \mathcal{I}$).

Initialization

1. Calculate values $\lambda_i = \sum_{\substack{i' \in \mathcal{N}_i \\ \psi_{i'} \geq \psi_i}} f_{i'}$ for all $i \in \mathcal{I}$.
2. Calculate values β_{hij} for all combinations of $h \in \mathcal{H},\ i \in \mathcal{I}$, and $j \in \mathcal{J}$.
3. Set $x_j = 1000$ (or higher bound if required) for exactly one randomly chosen base location $j \in \mathcal{J}$. It also works for other choices of allocations with a huge upper bound for the number of ambulances.

Iterations

1. Calculate the values of β_i for all $i \in \mathcal{I}$,

$$\beta_i = \max_{j \in \mathcal{M}_i} \left(\frac{\sum_{i' \in \mathcal{N}_j} d_{i'} \min_{h \in \mathcal{H}, j' \in \mathcal{J}, x_j > 0} \beta_{hi'j'}}{\sum_{i \in N_{j'}} d_i} \right). \tag{15}$$

   For a stepwise procedure we refer the reader to Subsection 6.2.
2. Calculate b_i from Erlang for each $i \in \mathcal{I}$,

$$b_i = \text{argmin}_{n \in \mathbb{N}_{\geq 0}} \{1 - \text{Erlang}_{\text{B}}(\lambda_i, \beta_i, n) \geq \alpha\}. \tag{16}$$

3. Allocate x_j ambulances at base $j \in \mathcal{J}$ by solving the IP-formulation,

$$\begin{aligned} \min\ Z = \sum_{j \in \mathcal{J}} x_j \\ \text{s.t.} \sum_{j \in \mathcal{M}_i} x_j \geq b_i,\ \forall i \in \mathcal{I} \\ x_j \in \mathbb{N}_{\geq 0}. \end{aligned}$$

Stop condition Stop when the number of vehicles $\sum_{j \in \mathcal{J}} x_j$ stays the same for two successive iterations. Although a rigorous proof is missing, in practice we noticed that this algorithm converges to its accumulation point within a few iterations.

Post-processor Run the post-processor and obtain the results. See Section 5.4.1 for a basic post-processor.

7.2 Relation between Q-models and AQ-models

We conclude the section with statements about the relationship between the Q-models and AQ-models.

Theorem 6 *The AQ-methods are a generalization of the Q-methods.*

Corollary 2 Correctness of the Q-methods *The ambulance allocation using the Q-methods neighborhood definition, the isolation assumption, and constant reliability level α guarantees a reliability level of at least α for each demand point $i \in \mathcal{I}$ that is covered.*

Note that the proof of Corollary 2 does not rely on a cancelation argument like the Q-method papers.

The following property show that every demand points arrival rate and mean service time are equal in the Q-methods and AQ-methods, that is, when the isolation assumption holds. As a result, when the reliability level is a system wide constant, both models give the same number of ambulances.

Property 1 Under the assumption that demand, base locations and hospitals are evenly and well spread between adjacent neighborhoods, we see that the FAQ-methods yield Erlang B input parameters λ_i, β_i similar to the Q-models for all $i \in \mathcal{I}$.

Property 2 If a region satisfies all assumptions of Q-PLSCP, then both Q-PLSCP and FAQ-PLSCP give exactly the same facility location and allocation solutions if one chooses $\alpha_i = \alpha$ constant for all demand points $i \in \mathcal{I}$ and if one uses a post-processor that always stops if the workload condition can be met.

These properties are also valid for all other density measures that are used in minimal reliability models, and the proofs go likewise.

8 Results

In this section we show results for four actual ambulance regions, where we compare the AQ-PLSCP to other models described in the current literature.

First we determine a set of parameters that are input for all calculations, next we compare Q-PLSCP to FAQ-PLSCP and DAQ-PLSCP. These calculations give the minimal number of vehicles needed to provide coverage. We also provide a comparison with the EBNB model, where all demand is projected to the nearest base, after which a per-base minimal required number of ambulances is calculated using the Erlang B formula.

We limit ourselves to weekdays between 10 and 12 am. These intervals do not contain a change of shift for any of the ambulance regions we considered, and they have a reasonable constant but substantial arrival rate. There are also no major fluctuations in demand between these weekdays.

Table 1 Constants taken from actual call database, for weekdays 10:00–12:00 over multiple years. *Note: These durations are not necessarily equal to the regions' official performance. For the official numbers we refer to [?].*

Region	**Urgencies**	**Number**	**PTD**	**CHT**	**TMT**	**TFT**	**Fractions**	
		of calls	**of HU**				**HU**	**LU by ALS**
Utrecht	Only ALS	28109	2:51	1:35	17:58	16:52	53.74%	42.88%
	All calls	73668		2:33	16:46	18:59	25.03%	
Amsterdam	Only ALS	26259	3:32	1:22	19:05	19:37	77.32%	54.59%
	All calls	59060		2:54	17:23	20:06	30.56%	
Gooi &	Only ALS	4061	2:37	1:04	15:57	10:49	75.48%	54.49%
Vechtstreek	All calls	8930		2:08	14:14	12:13	34.32%	
Flevoland	Only ALS	6640	2:50	1:22	15:46	17:13	59.21%	40.59%
	All calls	12725		2:25	14:54	16:58	33.04%	

8.1 Parameters

Table 1 displays all input constants that are taken from actual data. They are obtained from real data to stay close to reality. We aggregate to four position postal code level; see Table 2 for the number of demand points in every region. Because maximum reliability models assume that every demand point can be reached from at least one base location, we omit a few isolated demand points to be able to perform calculations.

Note that the choice of the scaling parameter does not influence the outcome of the methods.

An ambulance partner provided us with values $\alpha_i = 0.95$ for urban and $\alpha_i = 0.80$ for rural areas. After plotting the population density per postal code area on a map we choose a good threshold value between urban and rural areas. The population density does not necessarily depend on ψ_i. Although Q-PLSCP actually is not designed for a variable reliability value, we use it for fair comparison; after all the real regions also do not meet the isolation assumption.

8.1.1 Arrival rate, density function and reliability values

The arrival frequency f_i for every demand point is calculated from the actual call detail records that were provided by each of the individual ambulance regions, over the years 2008–2012. We counted the number of dispatched calls and divided through the total duration over all these 2-hour interval blocks.

The density measure for frequency adjustment is $\psi_i = f_i$. The density measure ψ_i for density adjustment is obtained by dividing the arrival rate through the area of postal area i and multiplying the result with a system wide constant scaling parameter $C \in \mathbb{R}$:

$$\psi_i = C \frac{\lambda_i}{area_i}. \tag{17}$$

8.1.2 Mean service time

We calculate the mean service time at demand point $i \in \mathcal{I}$

$$\tilde{\beta}_i = \min_{h \in \mathcal{H}, j \in \mathcal{J}, x_j > 0} \left(CHT + t_{ji} + TMT + t_{ih} + TFT + addon \right). \tag{18}$$

This is the time from the nearest base to the nearest hospital. Let us explain how and why the value for *addon* is calculated. When a patient is transported from a hospital, and there is a base location in the same postal area, we have $t_{ji} = 0$ seconds. This is not realistic, because often the ambulance station is positioned in a separate building. We use a constant of 5 minutes driving time between these buildings. This includes the opening and closing of garage doors. When an incident happens in a hospital, the nearest hospital would be the hospital itself, leading to a DTH of 0 seconds. Because calls that originate in a hospital often concern a patient brought home or taken to another hospital, this is not realistic. From the call center database we calculate an average DTH for every demand point that contains a hospital and use that as *addon*.

We correct the travel speeds for traveling with optical and auditory signals in the case the ambulance travels to a high urgency (HU) patient. This is the highest class of ALS urgency, and the only class where an ambulance always travels with active optical and auditory signals. Medium urgency (MU) calls are also considered ALS, and all low urgency (LU) calls are considered BLS. In reality ALS ambulances can respond to both ALS and BLS calls, and BLS ambulances can only respond to BLS calls. All travel speeds in this paper are deterministic and do not change over time, i.e., they are being requested by navigation software.

We consider two cases. The first case contains only the HU and MU calls, which are all being handled by ALS ambulances. This calculation requires a correction because in practice the ambulances in our data

Table 2 Number of demand points for every ambulance region.

Region	Utrecht	Amsterdam	Gooi & Vechtstreek	Flevoland
Demand points	220	103	41	94

Table 3 Required number of ambulances for weekdays 10:00–12:00 for EBNB, Q-PLSCP, FAQ-PLSCP, and DAQ-PLSCP. C: Actual number of ALS vehicles corrected for their BLS transportations. R: Advised number of ambulances for 95% within $R = 15$ minutes subject to at least one ambulance at every base location by RIVM. Notation AQ-models: before-after applying the post-processor.

Region	**Urgencies**	**Actual**		**EBNB**	**Q-PLSCP**		**FAQ-PLSCP**		**DAQ-PLSCP**	
				Base Locs	Base Locs		Base Locs		Base Locs	
		Act.	C/R	Fixed	Fixed	Free	Fixed	Free	Fixed	Free
Utrecht	Only ALS	31	C: 21.0	25	34	28	18-27	16-30	23-25	18-27
	All calls	37	R: 35	39	65	48	25-33	25-46	35-40	28-45
Amsterdam	Only ALS	35	C: 20.7	27	41	33	18-28	17-36	20-30	19-35
	All calls	40	R: 39	44	81	59	30-42	29-54	37-49	35-66
Gooi &	Only ALS	7	C: 4.1	5	5	5	4-4	4-7	4-5	4-5
Vechtstreek	All Calls	7	R: 6	8	9	9	6-7	5-8	6-7	6-8
Flevoland	Only ALS	13	C: 9	11	17	13	12-13	11-12	13-13	10-13
	All calls	13	R: 11	13	21	18	14-17	13-17	16-16	14-18

set do also respond to BLS calls, which is discussed in Section 8.2.1. The second case has one ambulance type that responds to all ambulances.

For the Q-methods we use for all $i \in \mathcal{I}$ a mean service time

$$\beta_i = \tilde{\beta}_i. \tag{19}$$

To calculate β_i we follow the procedure of Section 6.2 for the adjusted methods, for all $i \in \mathcal{I}$.

8.2 The Q-PLSCP model compared to FAQ-PLSCP and DAQ-PLSCP

We have implemented Q-PLSCP, FAQ-PLSCP, DAQ-PLSCP, and EBNB in the TIFAR framework, and used the Coin-OR CBC solver through the CoinMP interface to solve the IP-formulation.

The model is used to solve various scenarios. We made a distinction between calls of ALS urgency only, and all calls. A second distinction we made was between a free choice of base location and fixed locations. In the free base location, also known as a greenfield scenario, a potential base location is placed at the location of every demand point. For the fixed locations we used the actual base locations that were in use during the time where our data originates from. We use a constant $R = 15$ minutes. The results are presented in Table 3.

The naive *Erlang B on the Nearest Base* (EBNB) projects all demand to the closest base location, and solves Erlang B for each base location separately. Because EBNB takes the DTI equal to the driving time from the base location to the nearest hospital, the mean service time may be significantly lower for regions where hospitals are near base locations. This may lead to a lower required number of ambulance for EBNB.

Before we compare the two methods and the actual number, we discuss the correction for ALS ambulances that serve BLS demand.

8.2.1 Correction ALS vehicles for BLS load

Regions Utrecht and Amsterdam have both ALS and BLS vehicles. ALS vehicles are capable to handle BLS load, but this does not hold the other way around. To compare our results with the actual numbers, we have to correct the realistic number of ALS vehicles for their BLS load. We know that BLS vehicles have a higher busy fraction than ALS vehicles because they handle load that can be planned in advance. We assume that ALS vehicles will only do BLS within their own region, and that longer interregional rides are all done by the dedicated BLS vehicles. This yields similar driving times for all calls handled by ALS vehicles.

Our correction is done through an estimation. Realistic numbers for the busy fractions are 0.8 for BLS and 0.6 for ALS for dense regions like Utrecht and Amsterdam. The times used for these estimates include the driving back to base. When ALS vehicles handle BLS calls they are more efficient than handling ALS calls. This can be explained because first of all the stochastic arrival nature is eliminated and, second, because the ALS vehicles only do BLS load when there is an overcapacity in ALS vehicles. When dealing with BLS load, we assume that ALS vehicles are as effective as BLS vehicles doing the same work.

This means that when an ALS vehicle handles BLS load, it has 4/3 times the effectiveness. If we know the fraction of BLS load done by an ALS vehicle, we can use this fraction and the effectiveness measure to correct the actual number of vehicles for the BLS work they do. This yields a corrected number of 21.0 vehicles for Utrecht and 20.7 vehicles for Amsterdam.

8.2.2 Comparison of PLSCP-methods

We see that for all fixed base location scenarios the required number of ambulances in Q-PLSCP exceeds FAQ-PLSCP, and that the actual numbers for the fixed base locations are much closer to the actual (corrected) numbers than Q-PLSCP.

We give an interpretation to the results in Table 3. For *Only ALS* we see that the corrected actual number of ambulances for Utrecht and Amsterdam lies between the FAQ-PLSCP numbers with free and fixed bases. This is realistic, because DAM yields an optimization on the scenario with fixed base locations, and a free choice of base locations can be seen as optimal locations, which provides a lower bound. For *All calls* we see that FAQ-PLSCP gives a slight underestimation for these two regions. We assume the main cause is the way BLS calls differ from ALS calls, and BLS calls are a major part of the total volume.

Gooi & Vechtstreek is the smallest region of the country, and it has only ALS vehicles. We see that our results for *All calls* with fixed base locations are exactly the same. Q-PLSCP gives over-estimations.

Flevoland is a rural area with two large cities. We see that all minimal reliability models require more ambulances than what they currently have. In reality we see that the urban areas have a good score, while the rural areas underperform. The minimal reliability models allocate an additional ambulance on rural bases to meet performance which explains the slightly higher numbers. A clear difference between Q-PSCLP and AQ-PLSCP can be observed.

Let us discuss factors that influence the performance between our method and actual numbers.

Practical Implications The actual numbers are not necessarily the optimal allocation, because the regions optimize another objective function than we do. On the one hand they try to minimize the number of late arrivals, while keeping local communities satisfied. It may be that due to shift lengths they have a temporary over- or undercapacity, and they correct this in another time block to meet the constraints and objective.

Dynamic ambulance management All regions perform some kind of dynamic ambulance management (DAM). In our calculations we assumed that every ambulance returns to its home base, which lets us lose some performance.

Behavior of BLS vehicles As stated in the discussion above, BLS vehicles have a higher busy fraction because they handled planned transport that misses the stochastic nature of ALS calls. However, there are two things to keep in mind that can counter-affect this increase of performance. First of all the patient's destinations are located further away. Where ALS patients are most often transported to the nearest hospital location, BLS patients may be brought to any location in the country. The *addon* includes these travel times. Our method assumes that an ambulance becomes available again right after the transfer of the patient. When a BLS vehicle is outside of the region, it should return to its own region. This travel time back to base should be kept in mind for BLS vehicles. Second, there is only one location in a region where BLS vehicles are stationed. In the case of elective BLS transport the travel time to the patient might be higher than the travel time from the nearest base location that has at least one ambulance assigned. Third and last, available BLS vehicles are not relocated through DAM. In the further discussion we assume that BLS vehicles are all considered as effective as an ALS vehicle.

Response radius We assumed that every call should be available as if it were a high urgency call. The area in the calculation of λ_i equals that of a high urgency call. When considering other ALS calls that have a larger response radius and BLS vehicles that have no response time threshold, we see that our method may yield a slight over-estimation.

Limit on late arrivals versus mean response time There are two extremes in facility locations; one can either minimize the number of late arrivals or one can minimize the mean response time. We pose a lower bound on the late arrivals for every demand point. The result is that in some areas, the mean arrival time may approach the response time limit for various bounds. Especially when these demand points have a considerable demand, ambulance providers may choose to position more ambulances than strictly needed by our response time threshold, such that the mean response time decreases.

Spiking The demand point dependent reliability level requirement a_i, $i \in \mathcal{I}$, can result in an effect that we call *spiking*. If there is a small area with relative high demand and large α_{urban} encapsulated by a rural area with a low α_{rural}, the arrival rate $\lambda_\bullet$ of urban demand points is dominated by the summarized arrival frequency of the rural demand points. However, α_i depends only on the rural area. This combination may lead to relatively high values b_i for these urban demand points. An overcapacity of about three ambulances in ambulance region Flevoland is caused by this effect.

9 Conclusion and future research

9.1 Conclusion

In this paper we propose an adjusted queuing approach to the Erlang B queuing based facility location and allocation model that addresses, explains, and solves the over-estimation problem. We provide an application of this approach for the Q-PLSCP model. We show that these extensions lead to credible results for regions that have both urban and rural areas. AQ-models are generalization of their existing Q-model counterparts.

The theoretical framework provided in the paper proves the working of the AQ-approach and, hence, shows the correctness of the Q-models, amongst Q-PLSCP, which before relied on simulation results. The adjustment queuing model AQ-PLSCP that we propose gives sharper bounds for the required number of ambulances needed to fulfill the reliability condition than the Q-PLSCP model in practice.

9.2 Future research

There are some interesting topics left open for future research. First, in this paper we approximate a solution that respects the workload condition through a basic heuristic. It is interesting to use OR techniques to find a (proven) optimal solution that minimizes the number of ambulances that is required for the workload condition.

A key performance indicator in practice is the fraction of late arrivals, which strongly depends on the travel time distribution. To this end, it would be interesting to extend the current deterministic travel time model toward stochastic travel times. Extending the current fixed travel time model toward models with stochastic travel times is an interesting topic for further research.

Whether an ambulance is in time or late is a binary variable. An interesting extension is taking the probability that an ambulance is in time. This can be done by taking the continuous counterpart of Erlang-B, e.g. by allowing 2.5 ambulances to cover a demand point.

It is also interesting to research how we can incorporate the case when no available back-up EMS provider is available, and what effects this may have on the number of required ambulances. A good first approximation is to use Erlang C.

In the calculation of $\tilde{\beta}_i$ we assume that the ambulance comes from the nearest opened base location. An interesting enhancement would be to adapt this such that all staffed base locations that can reach the demand point are included in the calculation. We can do that by taking the probability that a base location is the one that sends the ambulance, multiplied by the driving time. One can also estimate a good distribution for the busy time at each demand point using call center records details.

Our method aggregates at demand-point level. This can be extended by giving a minimal reliability or availability on a set of demand points, so that we could say that a municipality is covered with 80% certainty instead a separate constraint for every single demand point in the municipality.

Models with a binomial approach also have a queuing approach equivalent. Hence, they have an adjusted queuing approach. Because the isolation assumption limitation has been overcome in this paper, this opens up an entire new family of models, for example AQ-REL-P and AQ-MALP. Using these insights it might be even possible to find the adjusted queuing counterpart of MEXCLP where the AQ definitions of coverage and busy fraction are being used; this loosens the hard assumption of a system wide busy fraction.

Acknowledgements Our thanks goes to RAVU, Ambulance Amsterdam, GGD Gooi & Vechtstreek, and GGD Flevoland for the call center records. We thank Connexxion NV for the area per postal code, and RIVM for the lookup table with driving times between every two postal codes in the Netherlands. This research is supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs.

Appendix A: Notation

This appendix contains all variables that are used in the current paper, in order of appearance.

Variable	**Description**
EMS	Emergency Medical Service. Short for ambulance service.
EMT	Emergency Medical Technician. Someone who works at an ambulance.
EMCC	Emerncy Medical Call Center. Also called (ambulance) dispatch center.
PTD	Pre-trip delay: the time it takes from an incoming call to when the assigned ambulance start driving.
$h \in \mathcal{H}$	A hospital in the set of hospitals.
$i \in \mathcal{I}$	A demand point in the set of demand points.
$j \in \mathcal{J}$	A potential base location in the set of potential base location.
$\mathcal{V}$	The set that contains all hospitals, demand points and potential base locations.
t_{ji}	The travel time from j to i. Similar notation for all two combinations of h, i, j.
r_{ij}	Response time from demand point i from base j.
R	Response time limit.
α_i	The reliability level: the probability that an ambulance is available to reach a patient at i within R time units must be at least α_i.
α	System wide constant version of α_i.
Z	The total number of ambulances in the system, when this number is variable.
x_j	The number of ambulances assigned to base j.
y_i	The number of ambulances that are positioned bases in neighborhood $\mathcal{M}_i$.
f_i	The arrival frequency at demand point i.
β_{hij}	The round trip time (without the driving to base time) for a trip from j to i to h.
β_i	Average service time at demand point i.
β	System wide constant version of β_i.
$\mathcal{N}_i$	The (demand point in the) neighborhood of i.
$\mathcal{N}_{\Diamond}^{\bullet}$	The demand points in the neighborhood of $\Diamond \in \{i, j\}$, using the neighborhood definition of $\bullet \in$ {Q, AQ, FAQ, DAQ}
$\mathcal{M}_i$	The base locations in the neighborhood of i.
$\mathcal{M}_{\Diamond}^{\bullet}$	The base locations in the neighborhood of $\Diamond \in \{i, j\}$, using the neighborhood definition of $\bullet \in$ {Q, AQ, FAQ, DAQ}
q	The busy fraction of an ambulance. A system wide constant variable.
$b_i^{\sharp}$	The number of ambulances that must be assigned to neighborhood $\mathcal{M}_i$, using the approach or neighborhood definition of $\sharp \in$ {Bin, Erl, Q, AQ, FAQ, DAQ}.
ρ_i	Workload that is generated at demand point i.
$b_i(\rho)$	The minimal required number of ambulances at workload ρ at demand point i.
b_i	Short notation for $b_i(\rho_i)$.
$\tilde{R}$	Response distance limit.
$\tilde{\mathcal{N}}_{\Diamond}^{\bullet}$	Distance based version of $\mathcal{N}_{\Diamond}^{\bullet}$, $\Diamond \in \{i, j\}$.
ϵ	A very short time period.
ϵ'	A very small portion of demand.
$\mathcal{A}$	The set of all ambulances.
$\mathcal{A}_i$	The set of ambulances that may serve demand at i.
n_i	The number of ambulances that serve demand at i.
$\Psi^{low}(\rho, n)$	The lower bound on the busy fraction for an ambulance at workload ρ and n serving ambulances.
$\Psi^{*,low}(\rho)$	The lower bound on the busy fraction for an ambulance in our solution.
$\Psi_i^{*,low}$	Short for $\Psi^{*,low}(\rho_i)$.
$\Psi^{upp}(\rho, n)$	The upper bound on the busy fraction for an ambulance at workload ρ and n serving ambulances.
$\Psi^{*max}(\rho)$	The upper bound on the busy fraction for an ambulance in our solution.
Ψ_i^{*max}	Short for $\Psi^{*max}(\rho_i)$.
ρ_i^{upp}	The workload at the time when b_i ambulances work that hard that exactly the allowed reliability level α_i is met.
$\rho_i^{*,upp}$	The workload at the time when n_i ambulances work that hard that exactly the allowed reliability level α_i is met.
Ψ_a^{dum}	The dummy busy fraction of an ambulance: a virtual ambulance busy fraction that after solving the program is within some bounds to the actual value.
ψ_i	A density assigned to i.
$\hat{\mathcal{I}}_i$	An ordered list of all demand points, using densities ψ_i.
$\tilde{\beta}_i$	An approximation for the mean service time at i.
$\tilde{\beta}_j$	An approximation for the mean service time of an ambulance stationed at j.

Appendix B: Proofs

In this Appendix we provide proofs to all theorems, lemmas, propositions, corollaries and properties that are mentioned in the paper.

Proposition 1 *For $b \leq n$, the following two conditions hold for any $\rho \geq 0$ and any fixed $\alpha \in [0,1]$.*
a) $\Psi^{low}(\rho, n) \leq \Psi^{low}(\rho, b)$. *Equality holds if and only if $n = b$.*
b) $\Psi^{upp}(\rho, n) \geq \Psi^{upp}(\rho, b)$. *Equality holds if and only if $n = b$.*

Proof In the proof we need the following Property: For each number of servers $s \in \mathbb{N}_{>0}$ and workload ρ, $Erlang_B(\rho, s)$ is continuous, strictly increasing in ρ, and surjective to the open interval $(0, 1)$. This is a direct result from Theorem 1 of Jagers [?], and the notion that $Erlang_B$ is a cumulative probability function.

For each $t \in \mathbb{R}$, number of servers $s \in \mathbb{N}_{>0}$ and workload $\rho > 0$, $Erlang_B(t\rho, ts)$ is strictly decreasing in t. A proof found by Burke is given in the appendix of [?]. **a)** $\Psi_i^{low}(\rho, n) = \rho/n \leq \rho/b = \Psi_i^{low}(\rho, b)$. Note that $\rho/n = \rho/b$ if and only if $n_i = b_i$.
b) That equality holds when $b = n$ is trivial. We consider the case $b < n$. The Property implies that, for given s, there is a unique value of ρ for which $Erlang_B(\rho, s) = \alpha$. This value of ρ is denoted by $\rho_s^\dagger$. Then by definition, for $s \in \mathcal{N}_{>0}$, $Erlang_B(\rho_s^\dagger, s) = \alpha$. By substituting $t = \frac{s+1}{s}$ in Burke's statement, it follows that

$$Erlang_B\left(\frac{s+1}{s}\rho_s^\dagger, s+1\right) < B(\rho_s^\dagger, s) = \alpha.$$

Since $Erlang_B(\frac{s+1}{k}\rho_s^\dagger, s+1) < \alpha$ and the Property tells Erlang B increases, we know that $\rho_{s+1}^\dagger > \frac{s+1}{s}\rho_s^\dagger$. This directly leads to $\frac{\rho_{s+1}^\dagger}{s+1} > \frac{\rho_s^\dagger}{s}$. Induction shows that $\Psi^{upp}(\rho_s^\dagger, s) = \rho_s^\dagger/s$ is increasing in s. The result follows directly.

Theorem 1 *For any $i \in \mathcal{I}$, $\rho_i > 0$, $\alpha_i \in [0,1]$ for $n_i \to \infty$, we have*
(a) $\Psi_i^{*,low} \downarrow 0$.
(b) $\Psi_i^{*,upp} \uparrow \frac{1}{1-\alpha_i} > 1$.

Proof Part (a) follows directly from the fact that $\lim_{n_i \to \infty} \Psi_i^{*,low} = \rho_i/n_i \downarrow 0$. To prove part (b), note that as $n_i \to \infty$, while there is precisely enough work to keep all agents fully busy, through economy of scale the busy fraction of an agent approaches 1. Because the agent is allowed to miss a fraction α_i of its total offered load, we conclude that the maximum allowed workload that the agent receives equals $\frac{1}{1-\alpha_i}$. The inequality then follows directly from $0 \leq \alpha_i < 1$.

Theorem 2 Workload condition *The reliability level α_i is guaranteed for every demand point $i \in \mathcal{I}$ if there exists an assignment Ψ_a^{dum} for every $a \in \mathcal{A}$ and an assignment $\mathcal{A}_i \subseteq \mathcal{A}$ for every $i \in \mathcal{I}$ such that all these conditions hold:*

$$\Psi_i^{*,low} \leq \Psi_a^{dum} \leq \Psi_i^{*,upp}. \tag{20}$$

Proof Assume that there exists an assignment Ψ_a^{dum} for every $a \in \mathcal{A}$ and an assignment $\mathcal{A}_i \subseteq \mathcal{A}$ for every $i \in \mathcal{I}$ such that the condition holds. We need to show that the reliability level α_i is guaranteed for every demand point $i \in \mathcal{I}$ by two steps: i) showing that the ambulances work hard enough that they can handle the workload at each demand point they serve and ii) other demand points outside the neighborhood $\mathcal{N}_i$ do not use an overdose of capacity from $\mathcal{M}_i$.

i) Choose $i \in \mathcal{I}$ at random and keep it fixed. Take $\Psi^{dum}_{a,min} = \min_{a' \in \mathcal{A}_i} \Psi^{dum}_{a'}$ the minimum value of the assigned ambulances' dummy busy fractions. The workload at i can be served by its $n_i = |\mathcal{A}_i|$ serving ambulances since $\Psi^{*,low}_i \leq \Psi^{dum}_{a,min}$ $\forall a \in \mathcal{A}_i$ yields $\rho_i = \Psi^{*,low}_i n_i \leq \Psi^{dum}_{a,min} n_i \leq \sum_{a' \in \mathcal{A}_i} \Psi^{dum}_{a'} =: \rho^{dum}$. Effectively, the workload that the n_i ambulances can handle within the performance level α_i exceeds the workload the neighborhood $\mathcal{N}_i$ generates.

ii) By assumption, the capacity of ambulances in $\mathcal{A}_i$ to serve demand points outside $\mathcal{N}_i$ is limited to $\Psi^{*,upp}_i$ $\forall a \in \mathcal{A}_i$. By Definition 3c, $\rho^{*,upp}_i := \text{argsup}_{\rho'}(Erlang_B(\rho', n_i) \leq 1 - \alpha_i)/n_i$ is the maximum workload that n_i ambulances can handle with a reliability level of α_i. It corresponds with a maximum workload of $\Psi^{*,upp}_i$. Because $\Psi^{dum}_a \leq \Psi^{*,upp}_i$ $\forall a \in \mathcal{A}_i$ we know that $\rho_i \leq \rho^{*,upp}_i$. Combine this with the knowledge that $Erlang_B(\rho_i, n_i)$ is strictly increasing in the workload (see Property from proof Proposition 1) and $\Psi^{dum}_a n_i = \rho_i$ to conclude that if $Erlang_B(\Psi^{*,upp}_i n_i, n_i) \leq 1 - \alpha_i)$, then also $Erlang_B(\rho_i, n_i) \leq 1 - \alpha_i$ holds. Hence, other demand points outside the neighborhood $\mathcal{N}_i$ do not use an overdose of capacity from ambulances in $\mathcal{A}_i$.

Theorem 3 *If the isolation assumption holds for an allocation through a Q-method, then the workload condition is satisfied.*

Proof The isolation assumption states that the demand over the neighborhood borders cancel out and that demand between neighborhood borders does not vary much over space. This means that the effective busy factions in neighborhood $\mathcal{N}_i$ may equal or slightly exceed that of the minimal busy fraction. However, the slight excess in the assumption says that the busy fraction does not exceed the maximal busy fraction. Combining these two statements yields $\Psi^{*,low}_i \leq \Psi^{dum}_a \leq \Psi^{*,upp}_i$ $\forall a \in \mathcal{A}_i$. The IP-formulation through a Q-method guarantees $b_i \leq n_i$. Since i is chosen at random it holds for all demand points.

Corollary 1 *The demand point coverage in Definition 4 is a generalization for the definition of coverage found in the Q-methods.*

Proof The Q-methods say that an ambulance is covered if $y_i = n_i \geq b_i$. In Q-PSCLP the isolation assumption holds, Theorem 3 states that the workload condition is satisfied. Hence, every demand point is covered in Definition 4. Remark: for Q-MALP a prior step is required. If a demand point is not covered in Q-MALP, $n_i \leq b_i$, and thus it cannot be covered through our definition. Remove all uncovered demand points from $\mathcal{I}$ so that we are left with the covered demand points. For these, we can use the same proof as in Q-PLSCP.

Theorem 4 Correctness of the AQ-methods *The ambulance allocation using the AQ-method's neighborhood definition guarantees reliability level α_i for each demand point $i \in \mathcal{I}$ that is covered.*

Proof Assume that the mean service time β_i is known for each demand point $i \in \mathcal{I}$. Consider a density measure $\psi = \{\psi_i : i \in \mathcal{I}\}$ on $\mathcal{I}$. Remark 1: Any density measure works, though some will perform significantly better than others. Remark 2: Value λ_i depends on the set of density values $\{\psi_i, i \in \mathcal{I}\}$. Another choice for the density measure ψ leads to other neighborhood definitions $\mathcal{N}_i$, and hence it can change arrival rate λ_i in neighborhood $\mathcal{N}_i$.

Because there are a finite number of demand points, there is always a demand point with the greatest density. An ordering over the set of demand points by their density in any descending order provides an index $v_i \in \{0, 1, \ldots, |\mathcal{I}| - 1\}$ to each demand point starting by the one with the greatest density value, i.e. $\psi_i \geq \psi_I$ if $v_i \leq v_I$, $i, I \in \mathcal{I}$. Take a fixed density order in the remainder of this paper in the case there that are a multiple that are equivalent.

Take a so-called *density ordering* v such that $v_i \in \{0, 1, \ldots, |\mathcal{I}| - 1\}$ such that $\psi_i \geq \psi_j$ if $v_i \leq v_j$.

Define the set of demand points that has at least density ψ_i by

$$\hat{\mathcal{I}}_i := \{i' \in \mathcal{I} \mid v_{i'} \leq v_i\} \text{ for all } i \in \mathcal{I}.$$

The proof follows from induction over the density ordering v and application of Theorem 2 on the adjusted queuing formulation of the neighborhood definition $\mathcal{N}_{\bullet}^{AQ}$. Note that the workload condition already guarantees the correctness if we are allowed to omit the arrival frequency of demand points with higher density value than the demand point of which we calculate the arrival rate.

Consider v_0 and the corresponding i_0. If i_0 is covered, we can apply the workload condition of Theorem 2 on the singleton $\hat{\mathcal{I}}_{i_0}$ to get the *guarantee* that the reliability level α_i holds. Using the interchangeability between the demand point i and neighborhood $\mathcal{N}_i^{AQ}$ and this cannot be altered by further induction steps, we can omit the demand point from our further calculations. Note that we are doing the same for demand points from other regions that are reachable within the response time thresholds; because their coverage is already taken care of, we omit them in the calculations.

Now take any other random v_i with corresponding neighborhood $\mathcal{N}_i^{AQ}$. We assume that all other demand points in $\hat{\mathcal{I}}_i$ are covered. The purpose of the neighborhood definition is to determine how many ambulances should be at the virtual base location at i if it should serve all uncovered demand around i. Because all demand points in $\hat{\mathcal{I}}_i$ are covered, we can omit those demand points; it is not necessary that ambulance capacity at virtual demand point i provides aid to these demand points. It does not matter that the ambulances that we assign in i also serve $\hat{\mathcal{I}}_i$, because the workload condition ensures that the effective workload of each ambulance is between the minimum and maximum allowed busy fractions for neighborhood i. Consider the set $I_{v_i} = I \backslash \hat{\mathcal{I}}_i$, and use the same argument as the first demand point v_0. Now, we can apply the workload condition of Theorem 2 on the singleton $\hat{\mathcal{I}}_i$ to get the *guarantee* that the reliability level α_i holds. Hence, we can omit i in the next iteration step.

Induction over v gives the guarantee for all demand points.

Proposition 2 is not part of the main text, but will be used in the theorem that follows.

Proposition 2 *Increased workload yields at least the same minimal required number of ambulances:*

$$b(\rho) \leq b(\varrho) \text{ if } \rho \leq \varrho.$$

Proof Because $Erlang_B(\rho, b) \geq Erlang_B(\varrho, b)$ for $\rho \leq \varrho$ and $Erlang_B(\rho, b) \geq Erlang_B(\rho, B)$ for $b \leq B$ there is a balancing effect between the workload and the required number of vehicles for a stable blocking probability. Hence

$$\begin{aligned} b(\rho) &= \operatorname{argmin}_{b' \in \mathbb{N}}\{Erlang_B(\rho, b') \leq 1-\alpha\} \\ &\leq \operatorname{argmin}_{B' \in \mathbb{N}}\{Erlang_B(\varrho, B') \leq 1-\alpha\} = b(\varrho). \end{aligned}$$

Theorem 5 *The required number of ambulances through AQ-PLSCP, being Z^{AQ}, is at most the number of required ambulances Z^Q through Q-PLSCP for constant $\alpha_i = \alpha$ over all demand points $i \in \mathcal{I}$ if the solution of the IP-formulation respects the workload condition.*

Proof The arrival rate for each demand point in AQ-PLSCP is bounded by the arrival rate for the same demand point in Q-PLSCP:

$$\lambda_i^{AQ} = \sum_{k \in \mathcal{N}_i^{AQ}} f_k = \sum_{\substack{k \in \mathcal{N}_i^Q \\ \psi_k \leq \psi_i}} f_k \leq \sum_{k \in \mathcal{N}_i^{AQ}} f_k = \lambda_i^Q.$$

For not-significantly varying service times this means $\rho_i^{AQ} \leq \rho_i^Q$. Proposition 2 then gives $b_i^{AQ} \leq b_i^Q$. Because the required number of vehicles on each demand point in AQ-PLSCP is at most the corresponding number in Q-PLSCP, it means that the IP-formulation yields $Z^Q \leq Z^{AQ}$. Because the solution of the IP-formulation respects the workload condition, the post-processor does not have any effect on the outcomes.

Theorem 6 *The AQ-methods are a generalization of the Q-methods.*

Proof Take $\psi_i = \psi$ and $\alpha_i = \alpha$ constant for all $i \in \mathcal{I}$ and assume the isolation assumption. Theorem 3 states that now the workload condition is satisfied. We have equal neighborhood definitions for every demand point $\mathcal{N}_i^{AQ} = \mathcal{N}_i^{Q}$, because the density values are the same for every demand point $i \in \mathcal{I}$.

Corollary 2 Correctness of the Q-methods *The ambulance allocation using the Q-methods neighborhood definition, the isolation assumption, and constant reliability level α guarantees a reliability level of at least α for each demand point $i \in \mathcal{I}$ that is covered.*

Proof Theorem 4 states that all AQ-methods are correct. Theorem 6 states that the Q-models are a special case of the AQ-methods. Hence, also the Q-methods are correct.

Property 1 Under the assumption that demand, base locations and hospitals are evenly and well spread between adjacent neighborhoods, we see that the FAQ-methods yield Erlang B input parameters λ_i, β_i similar to the Q-models for all $i \in \mathcal{I}$.

Proof For parameter λ_i: Because demand is evenly spread between adjacent neighborhoods, we see that $\varphi_i = f_i$ takes equal values, hence $\mathcal{N}_i^{FAQ} = \mathcal{N}_i^{Q}$. Thus λ_i is similar for all demand points i$\in \mathcal{I}$. For parameter β_i: When base locations and hospitals are evenly spread between adjacent neighborhoods, Equation 11 yields similar values $\tilde{\beta}_i$ for all $i \in \mathcal{I}$. Hence, $\tilde{\beta}_j$ and β_i are similar for all demand points i$\in \mathcal{I}$.

Property 2 If a region respects all assumptions of Q-PLSCP, then both Q-PLSCP and FAQ-PLSCP give exactly the same facility location and allocation solutions if one chooses $\alpha_i = \alpha$ constant for all demand points $i \in \mathcal{I}$ and uses a post-processor that always stops if the workload condition can be met.

Proof Take any region that respects all assumptions of the Q-models. Q-models use the isolation assumption that states that demand is evenly and well spread between adjacent neighborhoods. Using Equation 11 we can say that a constant β_i corresponds to base locations and hospitals being evenly and well spread between adjacent neighborhoods. Use Property 1 to see that λ_i and β_i are similar in both the FAQ-models and the Q-models for all $i \in \mathcal{I}$. By assumption $\alpha_i = \alpha$ is constant, hence through identical computations b_i gets the same value for Q-PLSCP and FAQ-PLSCP for all $i \in \mathcal{I}$. Now, note that b_i is similar valued for all neighborhoods. In other words, all neighborhoods get the same number of ambulances. Take $\mathcal{A}_i$ equal to the set of all ambulances on $\mathcal{M}_i$, that is, all ambulances may serve all demand points in reach. Now $\Psi_i^{*,low}$ is the same for all demand points, and $\Psi_i^{*,upp} =: \Psi^{*,upp}$ is the same for all demand points. Taking $\Psi_a^{dum} = \Psi_i^{*,upp}$ for all ambulances shows that the workload condition can be met. Because the workload condition can be met, the post-processor, by assumption, will not add any additional ambulances. Therefore, Q-PLSCP and FAQ-PLSCP give exactly the same facility location and allocation solutions.