Improving fairness in ambulance planning by time sharing

Jagtenberg, C. J.; Mason, A. J.

published in
European Journal of Operational Research
2020

DOI (link to publisher)
10.1016/j.ejor.2019.08.003

document version
Publisher's PDF, also known as Version of record

European Journal of Operational Research 280 (2020) 1095–1107

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Improving fairness in ambulance planning by time sharing

C. J. Jagtenberg*, A. J. Mason

University of Auckland, Department of Engineering Science 70 Symonds Street Grafton, Auckland 1010, New Zealand

ARTICLE INFO

Article history:
Received 14 December 2018
Accepted 1 August 2019
Available online 9 August 2019

Keywords:
OR in health services
Facility location
Fairness
Nonlinear optimization
Multi-objective optimization

ABSTRACT

Most literature on the ambulance location problem aims to maximize coverage, i.e., the fraction of people that can be reached within a certain response time threshold. Such a problem often has one optimum, but several near-optimal solutions may exist. These may have a similar overall performance but provide different coverage for different regions. This raises the question: are we making 'arbitrary' choices in terms of who gets coverage and who does not? In this paper we propose to share time between several good ambulance configurations in the interest of fairness. We argue that the Bernoulli–Nash social welfare measure should be used to evaluate the fairness of the system. Therefore, we formulate a nonlinear optimization model that determines the fraction of time spent in each configuration to maximize the Bernoulli–Nash social welfare. We solve this model in a case study for an ambulance provider in the Netherlands, using a combination of simulation and optimization. Furthermore, we analyze how the Bernoulli–Nash optimal solution compares to the maximum-coverage solution by formulating and solving a multi-objective optimization model.

1. Introduction

An important aspect of Emergency Medical Services (EMS) is positioning vehicles in such a way that incidents can be served as quickly as possible. This is the objective of the ambulance location problem: where to locate bases, and how to distribute the available vehicles over those bases, in order to achieve some level of service delivered to patients.

We consider a typical ambulance operation in which, for a set of given base locations, we have to assign each vehicle to a home base. We assume that each vehicle returns to its home base whenever it is idle. Typically an ambulance organization will seek one good *ambulance configuration*, i.e., one good assignment of ambulances to bases. Such a configuration will provide good ambulance response times for people who live close to those bases, but it might provide a poor service to areas where there is no ambulance based nearby, or where the number of nearby ambulances is insufficient to meet the demand. The service level for a configuration is typically measured in terms of a response time threshold, where we say that an emergency call is served on-time (or *covered*) if an ambulance reaches the call's location within the response time threshold. The fraction of people in the region that are reached within the response time threshold is often referred to as the *coverage*. This metric is closely related to performance targets: for example, in The Netherlands the target is that for urgent patients '95% of all calls should be reached within 15 minutes' (Hoogeveen, 2010).

Performance targets such as the one above lead to the search for *efficient* configurations. That is, it raises the question how to meet the target using a minimum number of vehicles. Alternatively, another question could be how to maximize the coverage given a fixed number of vehicles. In both cases one is asking for efficiency because the objectives are stated in terms of the *total number of people* with on-time arrivals.

There is a lot of literature that focuses on maximizing efficiency, both for the more general class of facility location problems, as well as ambulance specific problems. For an overview of ambulance location models, see Brotcorne, Laporte, and Semet (2003) and Li, Zhao, Zhu, and Wyatt (2011). The majority of these models uses mixed integer linear programming to maximize the (expected) coverage. Well-known examples include the Maximum Coverage Location Problem (MCLP) (Church & Revelle, 1974) and its extensions such as the Maximum *Expected* Coverage Location Problem (MEXCLP) (Daskin, 1983) and its variant for stochastic driving times (Goldberg, Dietrich, Chen, & Mitwasi, 1990). Another well-known approach to model ambulance performance is the hypercube queuing model (Larson, 1974).

The quest for efficiency means that we are dealing with an implicit choice of a social welfare measure. It implies that we define a society's welfare as the sum of the welfare of all individuals.

* Corresponding author.
E-mail addresses: c.jagtenberg@auckland.ac.nz (C.J. Jagtenberg), a.mason@auckland.ac.nz (A.J. Mason).

https://doi.org/10.1016/j.ejor.2019.08.003
0377-2217/

This is a so-called *utilitarian* social welfare function. While this is a commonly used social welfare function, the utilitarian approach has two issues.

When we create ambulance models with a utilitarian objective, the main issues are as follows. First of all, solutions to such models tend to move vehicles to densely populated areas, at the cost of people living in more remote areas. This raises political and philosophical questions about who should get coverage and who should not. A conclusion might be that to some degree, we prefer suboptimal solutions if they offer a more equitable service. The second issue is that, regardless of which model is used, any prediction of performance is going to be approximate - and therefore so is the claim that we have found an optimal solution.

We propose to resolve the issues above by *time sharing* ambulances. That is, rather than using one configuration at all times, an ambulance provider may manage their fleet differently at different times. This is already common practice when responding to changes in demand, e.g., many ambulance providers will operate differently on business days compared to weekends; however, this is *not* what we mean. Instead, we propose to use multiple configurations even on weekdays that are very similar in terms of ambulance demand. In this situation we assume that the level of service provided to an inhabitant in a certain location is the time average of the service provided to that location. The ultimate goal is to increase fairness throughout the region, without sacrificing too much in terms of the utilitarian social welfare.

Let us illustrate our proposal by the following example. Consider an EMS region that has roughly the same demand pattern from Monday to Thursday. We first consider the utilitarian optimal configuration (call this configuration A). Suppose that this configuration provides excellent service to people living in the larger cities, and somewhat worse service to people living in small villages in the outskirts[1]. In addition to configuration A, we also consider alternative configurations B and C that are near-optimal from a utilitarian perspective. Compared to configuration A, configuration B provides better service to the people in a village in the Western part of the region, at the cost of the service people in a central city. Configuration C is similar, except it benefits a village in the Eastern part of the region. In this example, covering both villages simultaneously is inefficient, in the sense that such a system would perform poorly from a utilitarian point of view. Rather than choosing between focusing on the city, the Western or the Eastern village 100 percent of the time, we propose to spend Mondays and Tuesdays in configuration A, Wednesdays in configuration B and Thursdays in configuration C. This solution increases fairness throughout the region as compared to the utilitarian optimum (which is: spend Monday through Thursday in configuration A). Because we have only considered configurations that provide a reasonably good overall service, this ensures that our solution does not significantly reduce the utilitarian objective by too much.

In this paper we show, given a set of allowed configurations, how to compute the fraction of time to spend in each configuration in order to maximize fairness. Any set of configurations may be used, although very large sets may lead to large run times in the optimization models. The configurations can for example come from existing practice, recommendations by staff and elsewhere, with these being evaluated by simulation prior to solving the optimization problem. We argue that in order to achieve fairness, one should focus on the time average performance for each individual in the region and aim to maximize the Bernoulli–Nash social welfare measure of the society. The nonlinear optimization model that computes the corresponding time shares is presented in this paper.

[1] Such behavior is rather realistic for a utilitarian optimal solution.

We summarize the novel ideas in this paper as follows. First of all, we view ambulance planning from a social welfare perspective and place some of the existing ambulance literature in this context. Furthermore, we propose using a social welfare function that is currently unexplored in the field of ambulance planning. Lastly, we propose sharing time between several good ambulance configurations in the interest of fairness. This gives rise to our nonlinear optimization model, which is the main contribution of this paper. To the best of our knowledge, this is the first paper that proposes these ideas.

We solve our model in a case study for a realistic EMS region in The Netherlands, using a combination of simulation and optimization. Furthermore, to investigate how the Bernoulli–Nash social welfare is related to the often-used utilitarian social welfare, we solve a multi-objective optimization problem that has both the Bernoulli–Nash social welfare and the utilitarian social welfare as objectives.

The rest of this paper is structured as follows. Section 2 contains a literature review. Section 3 recaps the Bernoulli–Nash social welfare and its use in time sharing. This is done using illustrative examples that are unrelated to ambulance planning, but that are selected to provide insight into different social welfare functions and the concept of time sharing. In Section 4 we introduce the first examples of time sharing for ambulance location problems. We do this using small problem instances - small enough to compute the Bernoulli–Nash optima by hand or brute force, and small enough to develop intuition about utilitarianism versus fairness in ambulance planning. Section 5 continues this work by describing the optimization model that allows us to maximize the Bernoulli–Nash social welfare for realistically sized problem instances. We include a case study in Section 6 and extend this work with a multi-objective optimization model in Section 7. We finish with a discussion in Section 8.

2. Problem background and related work

This section summarizes a selection of papers that deal with equity in ambulance location problems. Furthermore, since equity in ambulance planning is a relatively unexplored area, we also include literature in the broader context of facility location planning.

An early example of a non-ambulance model that aims to maximize equity is the *p-center* problem: given a task to build p facilities, it seeks to minimize the maximum distance of any demand node to its nearest facility. There are two integer programming formulations of the p-center problem in the literature (Daskin, 1995; Elloumi, Labbé, & Pochet, 2004) and it is known to be NP-hard (Kariv & Hakimi, 1979). In the context of emergency services, an example of measuring equity is found in Coulter (1980), which defines a metric used to optimize the allocation of police services in Tuscaloosa, Alabama. To quantify fairness, the authors take the square root of the sum of squared deviations between some measure of service delivered to an area and that area's proportion of the total population. A review of equity in facility location models in Marsh and Schilling (1994) concludes that there has also been little agreement as to how equity should be measured. The authors summarize a range of equity measures (or rather, *in*equity measures, as they have to be minimized in order to create equity), including the maximum distance separating any demand node from its nearest facility, variance, mean absolute deviation, coefficient of variation and more. It has been argued (Mulligan, 1991) that when researchers decide which measure to use, they should consider the measure's sensitivity to extreme values (locations of users). For that reason, these authors suggests *not* to minimize the maximal distance, but instead use measures such as the mean deviation or the standard deviation.

Some authors approach fairness in facility location problems from a game theoretic point of view. In Goemans and Skutella (2004) the question is where to build facilities, while also deciding how to allocate the total cost of building facilities to the customers. A central concern is that no coalition of customers should have an incentive to build their own facility or to ask a competitor to service them.

An issue that naturally comes up when dealing with equity is its trade-off with efficiency. Early work on this trade-off can be found in McAllister (1976), where it is noted that no objective means exist for determining their relative importance. That conclusion is also drawn in several other papers (Bertsimas, Farias, & Trichakis, 2011; Felder & Brinkmann, 2002; Leclerc, McLay, & Mayorga, 2012; Stone, 2001). The trade-off between efficiency and equity - and the apparent impossibility to weigh the two - has lead some researchers to focus on multi-objective optimization. This has applications in, e.g., locating undesirable facilities (Erkut & Neuman, 1992; Rakas, Teodorovic, & Kim, 2004). More specifically, in an emergency services context, there are multi-objective optimization models for disaster relief logistics (Bozorgi-Amiri, Jabalameli, & e Hashem, 2013; Zhan & Liu, 2011) (both using stochastic programming) and for minimizing the disparities in access to care between rural and urban communities Chanta, Mayorga, and McLay (2011). The paper that is perhaps most similar to our work is Enayatia, Mayorgab, Toro-Diaz, and Albert (2019), which follows a line of reasoning similar to ours, and shows that the issues we address have very recently sparked interest in the community. Their approach, however, is different from ours: they deal with equity versus efficiency by creating multi-objective optimization models for ambulance location and dispatching. Their metrics for efficiency are the mean response time and the expected coverage; their metrics for equity on the other hand are the variance, the squared coefficient of variation, and the Gini index (Ceriani & Verme, 2012; Gini, 1912) of the individual response times. (The Gini index is most well-known as a measure of inequality of wealth. It has a natural geometric interpretation as 1 minus twice the area between the Lorenz curve and the diagonal line representing perfect equality.) The objectives are computed using the hypercube model, and optimized using a genetic algorithm. Another paper that deals with balancing efficiency and equity is McLay and Mayorga (2013). We should note that in this paper, the central application is *dispatching*, i.e., deciding which vehicles goes to which patient, and not facility location; however, it is still relevant to mention their approach. The authors maximize efficiency while putting a lower bound on two equity measures: (1) the fraction of calls that are serviced from the nearest base, for each demand zone and (2) the survival probability for each demand zone. The authors classify these two measures as 'customer equity', and furthermore also deal with 'server equity': balancing the work load among vehicles (which we do not address here).

The metrics for equity mentioned above are not suitable to use in an optimization model that has a single objective function, as we will argue next. Minimizing the maximum distance from a person to its nearest facility has the downside that the measure only depends on one person (the one who is worst off), whereas we desire a lexicographic measure instead. That is, we may allow metrics that are dominated by their performance on a subset of the population, but only if they are also able to rank solutions based on differences in performance for the remainder of the population. Minimizing variance can have perverse outcomes, because we can minimize variance by getting to nobody on time. That is, this metric makes a solution that is uniformly bad look good. In contrast to the equity measures described above, we are interested in a measure that will help us identify solutions that are equitable to some degree, while at the same time also preferring efficient solutions over inefficient ones. That is, our optimization model should at the very least have the following property: improving the utility for one person, without changing the utilities for the rest of the population, should always result in an improvement in our objective value. (Note that using variance as an objective does not lead to this result.) And last but not least, we want *one* measure that prescribes exactly how to weigh equity and efficiency.

The arguments above lead us to conclude that equity by itself is not necessarily a desirable quality of a solution for ambulance planning. Maximizing equity means asking for equal outcomes for everyone, which is - although fair in a way - often very inefficient. Furthermore, it is intuitively clear that if service in one part of the region can be greatly improved, while only marginally sacrificing the service in the rest of the region, this is probably a good idea. In this paper we propose using a mathematical concept that captures this idea and prescribes how to weigh the different levels of service in different parts of the EMS region. We argue that one should view the ambulance location problem from a social welfare perspective and optimize the so-called Bernoulli–Nash social welfare. This approach is encouraged by several motivating examples where the Bernoulli–Nash social welfare measure leads to intuitively fair solutions. Throughout this paper, whenever we mention 'fairness', we are implicitly referring to the Bernoulli–Nash social welfare measure. In summary, such a 'fair' solution is more equitable than the most efficient solution, yet more efficient than the most equitable solution.

While most ambulance location problems are solved to find one permanent solution (a configuration that is to be followed at all times), there are exceptions. These exceptions typically allow some aspect of the problem to vary over time, and compute different solutions for different time periods. Examples include Repede and Bernardo (1994) and van den Berg and Aardal (2015): both are extensions of MEXCLP that incorporate temporally varying demands. Another time-dependent model is introduced in Schmid and Doerner (2010), which is an extension of the Double Standard Model (Gendreau, Laporte, & Semet, 1997). In Schmid and Doerner (2010) it is not the demand, but the travel time that varies over time. As noted before, this is different from our approach: we suggest switching between different configurations, not because the circumstances change, but in the interest of fairness.

Improving fairness through time sharing seems sensible in light of the following. Although an ambulance location problem often has one optimum, there typically exist several near-optimal solutions. These alternatives may have a similar *overall* performance but provide different coverage for different subregions. This raises the question: in choosing a theoretically optimal solution, are we in practice making 'arbitrary' choices in terms of who gets coverage and who does not? Are there alternatives that improve equity without sacrificing too much efficiency? More specifically, if we allow different configurations at different times, can we reach a long-term average performance that is almost equal to the optimum (in terms of efficiency), but with a higher Bernoulli–Nash social welfare?

3. Preliminaries

In this section we define fairness in terms of social welfare, and introduce three different social welfare functions. Furthermore, we explore how these functions can be applied to time sharing. This section contains examples that were designed to demonstrate these concepts; later, in Section 4 we will discuss what this means in an ambulance context.

3.1. Social welfare

Social welfare is measured as a function of the 'utilities' of individuals of a society. For example, a commonly used function is

the utilitarian social welfare function. Consider a society consisting of a population P, and for all people $p \in P$ denote their utility by $u_p \in \mathbb{R}^+$. The utilitarian social welfare, f_U is then given by

$$f_U = \sum_{p \in P} u_p.$$

When we are considering a large population P, it may be convenient to divide the population in smaller groups, where it is key that all people in one group have the same utility. Let I denote the set of groups that together constitute P. Then for $i \in I$, define $u_i \in \mathbb{R}$ to be the utility of a person in group i, and let $d_i \in \mathbb{N}$ be the number of people in group i. The utilitarian social welfare can then be written as

$$f_U = \sum_{i \in I} d_i u_i.$$

f_U is a common choice for a social welfare measure in ambulance planning: practically all models mentioned in the literature overview in Brotcorne et al. (2003) and Li et al. (2011) aim to maximize the total (expected) coverage; i.e. they seek a utilitarian solution where the utility of an individual is defined as the probability that an ambulance reaches them within the response time threshold. Given this definition of utility, groups may be formed based on location, since for any solution, people living in the same location will experience the same utility.

Although f_U is a common social welfare function, alternatives exist. First, let us consider an egalitarian, or 'Rawlsian', social welfare measure. This is defined as the minimum utility over all individuals:

$$f_R = \min_{p \in P} \; u_p.$$

Note that when we change our notation to groups of the population, this becomes

$$f_R = \min_{i \in I} \; u_i.$$

That is, the egalitarian social welfare measure does not depend on the size of the group (assuming that each group consists of at least one person).

Another option is the Bernoulli–Nash social welfare function. This is defined as the *product* of individual utilities. Using the notation that we introduced above, this can be written as

$$f_{BN} = \prod_{p \in P} u_p,$$

or equivalently:

$$f_{BN} = \prod_{i \in I} u_i^{d_i}.$$

Note that the Bernoulli–Nash social welfare is more egalitarian than a utilitarian measure in that it is more sensitive to the utility of the worse off individuals. To see this, it helps to realize that maximizing the Bernoulli–Nash social welfare is equivalent to maximizing the sum of logarithms of the individual utilities (as we will further demonstrate in Section 3.2). Note that $\log(0) = -\infty$, which means that individuals with a very small (near zero) utility contribute large negative values to this sum, giving them a greater influence in the objective.

We consider the Bernoulli–Nash social welfare measure's sensitivity to individuals with a utility of zero an important shortcoming of our approach. Should this issue arise in case studies, then we recommend to adapt either the data or the model in some way, for example by taking the individuals that always have a utility of zero and excluding them from the problem. An alternative option might be to increase their utilities from 0 to a very small positive number. We note that in the numerical work for this paper, we did not have to deal with this issue because for every individual there was at least one configuration that gave a nonzero utility.

The Bernoulli–Nash social welfare function is also found in other contexts under a different name: economists tend to recognize it as the 'Cobb-Douglas' function. The Cobb-Douglas function was originally introduced in 1928 in a theory of production (Cobb & Douglas, 1928). There, it has an interpretation as a *production function*, which essentially combines two or more input factors (typically labor and capital), and relates this to the amount of output that can be produced by those inputs. Throughout the rest of this paper, we do not refer to Cobb-Douglas, and only use the term Bernoulli–Nash social welfare (f_{BN}). The Bernoulli–Nash social welfare corresponds to a form of fairness, as we will demonstrate next.

3.2. A radio time sharing example

In this section we illustrate why it makes sense to use the Bernoulli–Nash social welfare for time sharing. We do this by recapping an example found in Moulin (2003), example 3.6.

Consider a group of n agents working together in a common space. They all listen to a radio which must be tuned to one of five available stations. The agents have different tastes in music, and it is up to the manager to decide how the time is shared fairly between the five stations. That is, the manager has to decide times shares λ_i, $i = 1, \ldots, 5$ such that $\lambda_i \geq 0$ and $\lambda_1 + \lambda_2 + \ldots + \lambda_5 = 1$. In this example, an agent can either like or dislike a station.

A Basic Example. In its simplest form, each agent likes exactly one station and dislikes the other four. Let d_i denote the number of fans of station i, with $d_1 + \cdots + d_5 = n$. Note that in this example, the utility of a person in subgroup i is equal to λ_i.

A utilitarian manager would choose to play the station with the largest support all the time. (Note that if there are several such stations, mixing between them is optimal as well). An egalitarian manager, however, would do the opposite: play each station $\frac{1}{5}$th of the time.[2] Note that this ensures everyone is happy 20 percent of the time.

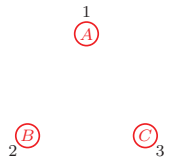
The Bernoulli–Nash social welfare can be viewed as a compromise between the two solutions above. We seek to maximize $f_{BN} = \prod_i \lambda_i^{d_i}$, which is the same as maximizing $\sum_i d_i \log(\lambda_i)$. If we maximize this under the constraint $\sum_i \lambda_i = 1$, it leads to a solution of $\lambda_i^* = d_i/n$, i.e., the time share of each station is proportional to the number of its fans.

An elaboration. In a slightly elaborated version of the example above, some agents are flexible, in the sense that they like more than one radio station. Let us assume that there are five agents with the following preferences (where a 1 indicates that an agent likes a station):

		Station				
		A	*B*	*C*	*D*	*E*
Agent	1	1	0	0	0	0
	2	0	1	0	0	0
	3	0	0	1	1	0
	4	0	0	0	1	1
	5	0	0	1	0	1

A utilitarian manager would now choose to share the time between stations C, D and E, such that there are always as many people as possible listening to a station they like (being two agents in this case). Unfortunately for Agent 1 and 2, they never get to listen to a station they like. An egalitarian manager would select $\lambda_a = \lambda_b = \frac{2}{7}$, $\lambda_c = \lambda_d = \lambda_e = \frac{1}{7}$, such that every agent likes the music $\frac{2}{7}$th of the time.

[2] Assuming each station has at least one fan.

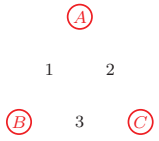


(a) Bases and nodes for example 1

	A	B	C
Base 1	1	0	0
Base 2	0	1	0
Base 3	0	0	1

(b) Coverage of nodes by bases.

Fig. 1. (a) Three bases (labeled 1, 2 and 3), and three demand nodes (*A*, *B* and *C*). (b) The associated coverages. In this example, from each base, exactly one demand location can be reached within the response time threshold.



(a) Bases and nodes for example 2

	A	B	C
Base 1	1	1	0
Base 2	1	0	1
Base 3	0	1	1

(b) Coverage of nodes by bases

Fig. 2. (a) Three bases (labeled 1, 2 and 3) and three demand nodes (*A*, *B* and *C*). (b) Their associated coverages. From each base, its two closest demand locations can be reached within the response time threshold.

The Bernoulli–Nash collective utility function again offers a compromise between the two solutions above, as it recommends playing each station $\frac{1}{5}$th of the time. To see this, observe that stations *a* and *b* are equally popular. Therefore, these time shares have the same exponent in the objective function, hence in a Bernoulli–Nash optimum they will receive the same time share *x*. Similarly, *C*, *D* and *E* will be allocated the same time share *y*. The Bernoulli–Nash maximization problem can then be written as

$$\text{maximize } f_{\text{BN}} = x^2(2y)^3 \quad \text{s.t. } x, y \geq 0, 2x + 3y = 1$$

which indeed leads to $x^* = y^* = \frac{1}{5}$.

Next, we show how the concept of Bernoulli–Nash social welfare translates to the ambulance location problem.

4. Social welfare for ambulance planning

In this section we analyze small instances of the ambulance location problem. We designed these instances specifically to explain our ambulance location problem and develop intuition about utilitarianism versus fairness in this context.

In all instances, the demand is distributed over three locations or *nodes*, labelled *A*, *B* and *C*. We claim that for any configuration of ambulances, people living in the same location can expect the same level of service (regardless of the chosen performance indicator) - and hence, have the same utility. Therefore, it is sensible to group the population by their location: this constitutes the groups *I*. In all of these examples, each node *i* contains a fraction d_i of the total demand, which means that if an ambulance is needed, this is in location *i* with probability d_i.

Throughout all these examples, there is one ambulance. We can position this ambulance in one of three *bases* (labelled 1, 2 and 3). The decision variables are λ_b: the fraction of time that the ambulance should spend at each base $b \in \{1, 2, 3\}$, where $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

For the first two examples, we define the utilities in terms of a response time threshold, where the utility of a person living in a certain demand node is 0 or 1, depending on whether or not there is an ambulance stationed at a base that is closer than the threshold. This utility is also known as *base coverage*: this is a simple measure that can be used if we assume that there are never two calls in progress simultaneously. (The average base coverage for the population as a whole is sometimes referred to as single coverage.) We will generalize this assumption and adopt a more sophisticated utility function in the third example, but the simplicity of the single coverage helps to illustrate our main idea as we build up to the more complex case.

Example 1

Consider the case where each node can be reached in time from one base, and from each base one can reach exactly one demand node in time. An example of such a setup is shown in Fig. 1. The corresponding base coverage matrix is given in Fig. 1. As discussed before, these coverage values are defined to be 0 or 1 (either a vehicle is stationed within the response time threshold, or not).

Now consider time sharing: the ambulance can be assigned to a different base at different times. More specifically, let's say the ambulance spends some fraction of time λ_i having base *i* as its home base, for $i \in \{1, 2, 3\}$. There is the matter of how to shift between the different configurations: we assume that we can ignore time spent driving between bases, perhaps because the ambulance moves between bases during periods of zero call demand, or perhaps because such moves happen infrequently. Furthermore, recall that we ignore ambulance unavailability in this example. Under these assumptions, λ_i gives the probability of the ambulance being at base *i* when any emergency call arrives. In this example, we define the utility of an individual at some node to be the fraction of time that an ambulance is stationed at 'their' base, and note that this utility can take a value between 0 and 1.

A utilitarian optimum would be to permanently assign the ambulance to the base where it serves the biggest demand. Such a utilitarian solution is what we call *efficient*; however, it is clearly far from fair. Alternatively, let us look at what it means to maximize the Bernoulli–Nash social welfare for this system. This problem is given by:

$$\begin{aligned} \max \quad & f_{\text{BN}} = \lambda_1^{d_A} \lambda_2^{d_B} \lambda_3^{d_C} \\ \text{subject to} \quad & \lambda_1 + \lambda_2 + \lambda_3 = 1 \\ & 0 \leq \lambda_1, \lambda_2, \lambda_3 \leq 1 \end{aligned}$$

Solving this gives a solution $\lambda_1 = d_A, \lambda_2 = d_B$ and $\lambda_3 = d_C$ that maximizes the Bernoulli–Nash social welfare. Just as in the basic case of Section 3, the proportion of time spent in each node is proportional to the number of inhabitants served. We argue that this can be considered a fair distribution of resources over the population.

Next, we show what happens if the situation becomes slightly more complex.

Example 2

Consider the case where each node can be reached by two bases, and each base can reach the two closest nodes. An example of this is shown in Fig. 2.

In this case, our Bernoulli–Nash measure is now maximized by solving

$$\begin{aligned} \max \quad & f_{\text{BN}} = (\lambda_1 + \lambda_2)^{d_A} (\lambda_1 + \lambda_3)^{d_B} (\lambda_2 + \lambda_3)^{d_C} \\ \text{subject to} \quad & \lambda_1 + \lambda_2 + \lambda_3 = 1 \\ & 0 \leq \lambda_1, \lambda_2, \lambda_3 \leq 1. \end{aligned}$$

The solution depends on how the demand is distributed over *A*, *B* and *C*. For the simple case $d_A = d_B = d_C = \frac{1}{3}$ the optimal solution is $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$. This corresponds to a social welfare of $f_{\text{BN}} = \frac{2}{3}^{1/3} \cdot \frac{2}{3}^{1/3} \cdot \frac{2}{3}^{1/3} = \frac{2}{3}$. For comparison, consider a case where one of the nodes has fewer inhabitants than the rest: $d_A = \frac{1}{5}, d_B = d_C = \frac{2}{5}$. In this case a utilitarian manager would choose to only serve locations *B* and *C*, i.e., the ambulance would be stationed

at base 3 at all times. The Bernoulli–Nash optimal solution on the other hand is $\lambda_1 = \lambda_2 = 0.2, \lambda_3 = 0.6$, and the corresponding social welfare is $f_{\text{BN}} = 0.4^{1/5} \cdot 0.8^{2/5} \cdot 0.8^{2/5} \approx 0.696$. (Note that this is slightly higher than in the case where the demand is distributed equally over the nodes.) This example shows that the Bernoulli–Nash social welfare function favors areas with high demand, but does not leave areas with low demand completely uncovered.

The two examples above illustrate our main idea with simple 0–1 utility functions: either an ambulance is stationed 'nearby' (where nearby is defined in terms of a response time threshold), or it is not. However, an actual EMS system is more complex than that. Therefore, we next explain how one might capture the dynamics of such a realistic system and obtain numerical values for a more complex utility function.

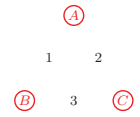
4.1. A realistic utility function

In this section, we introduce a more sophisticated utility function: utility values are no longer restricted to zeroes or ones. In fact, we may use any measure that maps a distribution of response times into a utility value. We choose to define the utility of a person to be the *probability that an ambulance will arrive within the response time threshold*. This utility is more complex than single coverage for the following three reasons. First of all, an ambulance positioned at a base is not always available; instead, it may be busy serving a patient when another call arrives. Another matter has to do with ambulances responding while they are on the road. That is, an ambulance may finish serving one patient and be on its way back to the base, when a new call arrives. The ambulance is then dispatched immediately, and no longer has to travel from the base. Lastly, in reality there will always be some randomness in response times.

Consider an adaptation to Example 2. Let us take the same demand locations and bases as in Fig. 2. We denote an ambulance configuration by a vector of length 3, where the ith number indicates the number of ambulances that have location i as home base. As was the case in Example 2, we have the ability to place one ambulance, which means that the possible configurations are (1,0,0), (0,1,0) and (0,0,1). (Recall that any location can be reached in time from its two nearest bases, but not from the furthest base. Moreover, driving between two demand locations takes longer than the response time threshold.) Now, how do we determine the values in the utility matrix? The complexity of the EMS system described above makes it hard to capture the utility function analytically. Instead, we argue that the most accurate approach for constructing a realistic utility matrix is by estimating them using simulation. (A possible alternative would be to use the hypercube model (Larson, 1974) or an approximation thereof.)

Simulation is an often-used method to evaluate EMS performance; see for example Henderson and Mason (2005), Aboueljinane, Sahin, and Jemai (2013), Ridler, Mason, and Raith (2017), UoA EMS Research, Jagtenberg, Bhulai, and van der Mei (2015). We can simulate the EMS process using EMS calls emerging at demand locations, and ambulance movements between the calls and the base. The observed fraction of on time arrivals in a location can then be used as an estimate for the utility of an individual in that location. In order to simulate this system we need additional information such as travel speeds, call arrival rates, how long it takes to serve a patient, and perhaps a road network. We will explore all this in our elaborate case study in Section 6; however, for now we do not define these details. Instead, we next show some possible results that might have been obtained through simulation of the region depicted in Fig. 3, in order to illustrate our approach.

The values in Fig. 3 show the effects of the two key features described above, which we reiterate now. Consider the first configuration, in which the ambulance is stationed at Base 1. In this case, location A can often be reached on time. However, sometimes the ambulance is busy when a new call arrives in A, which typically means the patient in A ends up waiting longer than the response time threshold. That is why the probability of serving a patient in A using configuration (1,0,0) is smaller than 1. Another effect we observe is that a new call can arrive when the ambulance has just finished serving a patient in location C. If this call is in location C, it can be reached within the response time threshold: in this example the likelihood of this situation occurring is illustrated by the value 0.03 in the matrix for configuration (1,0,0) and location C.



(a) Bases and nodes for example 3

	A	B	C
(1,0,0)	0.85	0.85	0.03
(0,1,0)	0.85	0.03	0.85
(0,0,1)	0.03	0.85	0.85

(b) Utility of nodes per ambulance configuration

Fig. 3. (a): A set of three bases (labeled 1, 2 and 3) and demand nodes (A, B and C). From each base, the two closest demand locations can be reached within the response time threshold. (b): the observed fraction of on time arrivals when the ambulance is stationed at a certain base.

We next analyze the optimal time shares for the configurations defined in Fig. 3. Again, assume that the demand is distributed as follows: $d_A = 0.2$ and $d_B = d_C = 0.4$. The utilitarian solution is the same as in Example 2: the ambulance spends all its time stationed at Base 3, from where it can serve the greatest demand. The Bernoulli–Nash social welfare is maximized by solving

$$\begin{aligned} \max \quad & f_{\text{BN}} = (0.85\lambda_1 + 0.85\lambda_2 + 0.03\lambda_3)^{d_A} \\ & \cdot (0.85\lambda_1 + 0.03\lambda_2 + 0.85\lambda_3)^{d_B} \\ & \cdot (0.03\lambda_1 + 0.85\lambda_2 + 0.85\lambda_3)^{d_C} \\ \text{subject to} \quad & \lambda_1 + \lambda_2 + \lambda_3 = 1 \\ & 0 \le \lambda_1, \lambda_2, \lambda_3 \le 1. \end{aligned}$$

We solved this model using a solution technique that we will describe in Section 6.3. For now, we merely state that the solution turns out to be $\lambda_1 = \lambda_2 \approx 0.19$, $\lambda_3 \approx 0.62$, and the corresponding social welfare is $f_{\text{BN}} \approx 0.6$. The reader may compare these time shares to the time shares found in Example 2, and conclude that using the 0–1 coverage (base coverage) lead to a slightly lower time share in configuration (0,0,1), i.e., using base coverage resulted in a small underestimation of the importance of locations B and C (the ones with relatively high demand).

Throughout the rest of the paper we will continue to use as utility function the probability that an ambulance arrives within the response time threshold. We consider this a reasonable utility function. Alternatives will be discussed in Section 8.

The previous examples show that the Bernoulli–Nash social welfare corresponds to a form of fairness. Despite this anecdotal evidence, one may still ask why it would be reasonable - let alone fair - to optimize the product of the individual utilities. We now provide an interpretation that helps to further understand the differences and similarities between a utilitarian optimum and a Bernoulli–Nash optimum. Recall that we defined our utility function to be the probability that an ambulance arrives within the response time threshold. Then, maximizing the sum of all utilities (utilitarian objective) is equivalent to maximizing the *expected number of on time arrivals*. Note that a utilitarian solution will leave some areas uncovered if this helps more people than it disadvantages. In contrast, imagine the situation where every individual makes exactly one ambulance call at a random time. Maximizing the product of all utilities (f_{BN}) is then equivalent to maximizing the joint probability that *everybody* receives their ambulance on time. It is this measure of fairness that we seek to maximize.

Note that the interpretation given above, that the Bernoulli–Nash social welfare measures the joint probability that everybody receives their ambulance on time, is valid if incidents occur independently of one another. This independence assumption is often made in ambulance models, and might be realistic for incidents such as heart attacks. For some incidents however, such as incidents caused by the weather, or traffic collisions involving multiple patients (requiring more than one ambulance simultaneously), it is clear that the independence assumption is a simplification of reality.

Next, we generalize the examples from this section to a model that can handle multiple ambulances and an arbitrary set of bases and demand locations.

5. Optimization model

This section formally describes the ambulance system we wish to consider, and details the model that optimizes the Bernoulli–Nash social welfare for this system.

We assume that demand for ambulances in a region can be modeled using a set of demand nodes (or zones) V. Each node $i \in V$ has a non-negative demand fraction d_i, such that $\sum_{i \in V} d_i = 1$. There is a fixed number of ambulances available, which may be stationed at a given set of base locations. We define a *configuration* to be an allocation of ambulances to bases (and we allow a base to hold multiple ambulances). In this problem, we only wish to consider a predefined set C of possible ambulance configurations which we assume the user has constructed.

Our optimization model requires, as input, the utility u_{ic} at each demand node $i \in V$ for each configuration $c \in C$. As we pointed out earlier, in order for the Bernoulli–Nash social welfare measure to be non-zero, each zone i must have non-zero utility. That is, for each $i \in V$ there should be at least one configuration $c \in C$ such that $u_{ic} > 0$.

Assuming that these values u_{ic} have been determined, we proceed by building the optimization model. The goal is to determine the fraction of time λ_c to spend in each configuration $c \in C$. We allow $\lambda_c = 1$, which indicates that we have chosen just a single configuration for the system. We may interpret λ_c, the time share for configuration c, as giving the probability of the system being in configuration c when an ambulance is dispatched to a call. Given a solution $(\lambda_1, \lambda_2, \ldots, \lambda_{|C|})$, the long-term average utility that a person living in node i receives is given by $\sum_{c \in C} \lambda_c u_{ic}$. Therefore, the Bernoulli–Nash social welfare is given by

$$f_{\text{BN}}(\lambda_1, \lambda_2, \ldots, \lambda_{|C|}) = \prod_{i \in V} \left(\sum_{c \in C} \lambda_c u_{ic} \right)^{d_i}. \tag{1}$$

We can now form the following non-linear Bernoulli–Nash Time Sharing (BNTS) optimization model:

$$\text{BNTS:} \quad \max \quad f_{\text{BN}} = \prod_{i \in V} \left(\sum_{c \in C} \lambda_c u_{ic} \right)^{d_i} \tag{2}$$

$$\text{s.t.} \quad \sum_{c \in C} \lambda_c = 1, \tag{3}$$

$$0 \leq \lambda_c \leq 1 \quad \forall c \in C \tag{4}$$

6. Case study

This section reports the details of a case study in which we compute the Bernoulli–Nash optimal time shares, applying the BNTS optimization model from Section 5 to a realistic EMS region.

Our numerical work concerns the province of Utrecht, a region in the Netherlands with an area of approximately 1,400 square kilometers. Emergency requests are handled by a single EMS provider, that currently has nineteen bases in the region (see Fig. 4). Utrecht is one of the busiest EMS regions of the country.

Utrecht consists of 217 postal codes, the details of which can be found in Appendix. The centroids of these postal codes form our set of demand nodes V. We define the fraction of demand d_i in a single node to be proportional to the number of inhabitants in that postal code.

We consider a fleet of nineteen vehicles that we wish to distribute over the nineteen base locations that exist in this region. In this case study, we only want to consider configurations that have a good overall performance, that is, we want to consider a set of configurations that are near-optimal to the utilitarian ambulance location problem. One way to find these would be to use solutions to one or several integer programming models for the ambulance location problem. However, these solutions can only be considered optimal with respect to the model itself, which is always a simplification of the problem. Instead, we use a combination of simulation and local search to find locally optimal configurations for the utilitarian problem.

We next describe both the simulation model and our local search procedure in more detail.

6.1. Simulation

We used a discrete event simulation model that was previously described in Jagtenberg et al. (2015). This model keeps track of all incidents and vehicles. Vehicle travel occurs on a network that contains nodes for demand nodes (being postal code centroids), hospitals and ambulance bases. The simulation uses deterministic lights-and-siren driving times between nodes as estimated by the Dutch National Institute for Public Health and the Environment (RIVM) in 2009 (Kommer and Zwakhals, 2011, Chapter 3). Travel speeds without lights and sirens are assumed to be 10% lower.

The simulation generates events for an incident occurring, an ambulance arriving at the scene of the incident, an ambulance leaving for a hospital, an ambulance arriving at a hospital, and an ambulance becoming idle. We simulate incidents arriving according to a Poisson process with an average inter-arrival time of 6.4 minutes, and draw the incident locations based on the demand distribution, i.e., an incident occurs at node i with probability d_i, for $i \in V$. (Note that in our simulation the demand is assumed to be proportional to the number of inhabitants. When readers consider applying our work in practice, we recommend using historical EMS call data (if available) in the simulation, while using the number of inhabitants to obtain the values for the parameters d_i used in the optimization objective (1).) Each incident is served by the closest idle ambulance available at that time - including ambulances that are currently on the road returning to base. Because we do not have a full road network, we approximate the location of those ambulances on the road by using the longitude and latitude of each node, and assume that ambulances travel with constant speed in a straight line between them.[3]

After an ambulance arrives at the scene of an incident, it spends a random amount of time there. This time is drawn from an exponential distribution with an expectation of 12 minutes. After this time, it is decided whether or not the patient needs treatment at a hospital (according to a Bernoulli distribution with probability 0.8). If not, the ambulance becomes available at the scene of the incident. Otherwise, the ambulance drives to the nearest hospi-

[3] Given the estimated travel time between the ambulance's origin and destination, as well as the time that has passed since the vehicle left its origin, we then compute its longitude and latitude and round this to the nearest node in V.

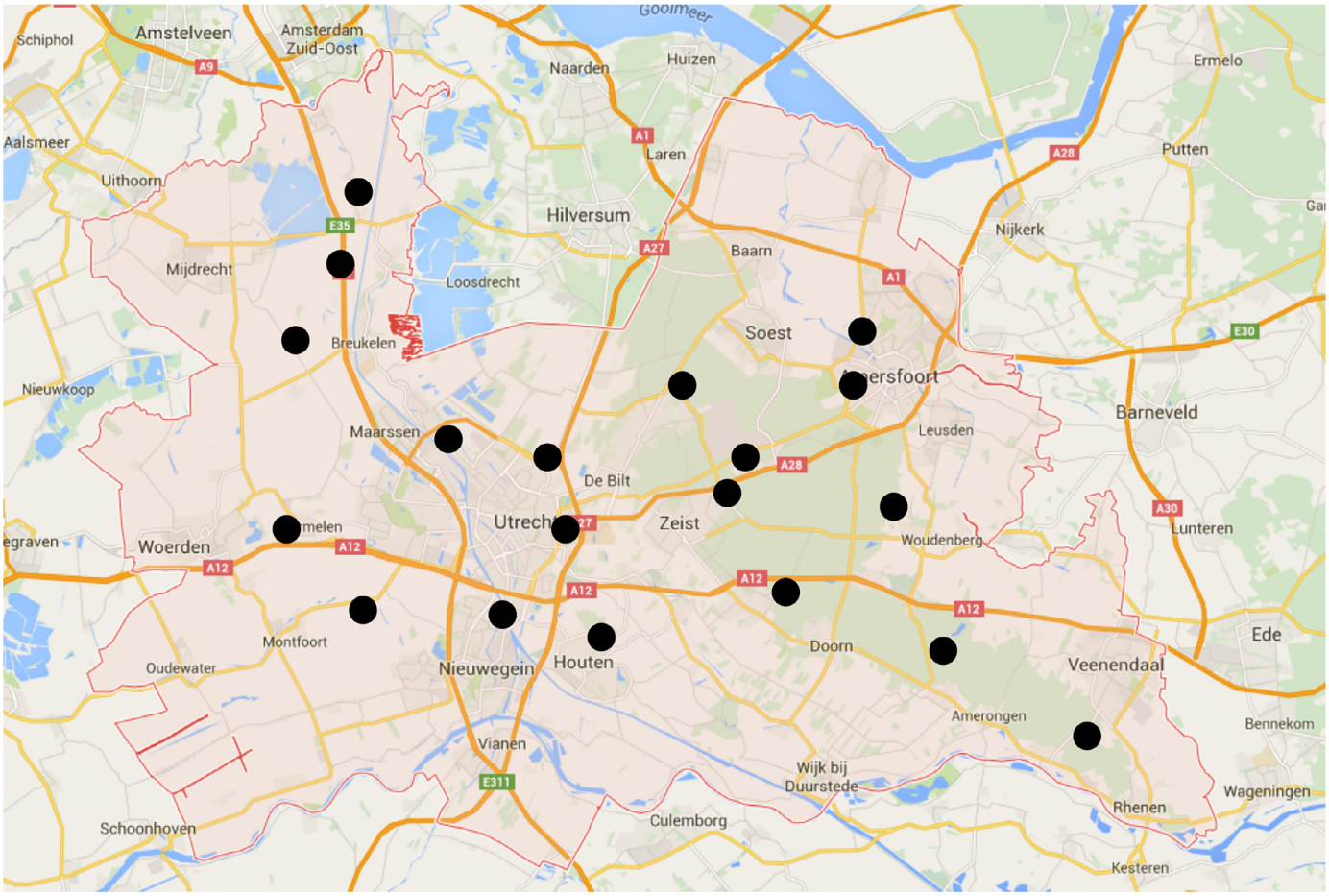


Fig. 4. The nineteen existing ambulance base locations in the region of Utrecht, The Netherlands.

tal[4], and spends an additional drop-off time there (drawn from a Weibull distribution with an expectation of approximately 15 minutes). Eventually the ambulance becomes available at the hospital location.

When an ambulance becomes available, we check if there are any unattended incidents left in the queue. If so, the ambulance is immediately dispatched to the first call in the queue. Otherwise, the ambulance stays idle, and is sent to its base location[5].

Using this model, we evaluate any given ambulance configuration by simulating 5000 hours of EMS events. In the Netherlands, ambulances should arrive within 15 minutes. Typically, 3 minutes are reserved for handling the call, therefore we use a response time threshold of 12 minutes. We estimate the utility u_{ic} of each demand zone $i \in V$ by measuring the fraction of calls there that are reached within 12 minutes. We note that longer runs can always be used to improve these estimates; however, our simulation should be sufficient to meet our goal, which is to demonstrate one approach to obtain input for our optimization model.

The purpose of this simulation model is twofold. First of all, it is used to evaluate intermediate solutions in the local search. Second, we use it to determine the utility per demand node for those solutions that we consider good enough to be in C.

[4] We use a set of 10 hospitals, excluding private clinics, that existed in the region in 2013.

[5] Note that the ambulance might not arrive at this base location, because it may be dispatched to a new call before reaching its destination.

6.2. Local optima

We want to position nineteen vehicles on the nineteen bases of region Utrecht. As it is possible to position more than one vehicle on a base, there are many different configurations to choose from. We search for solutions with high coverage, i.e., where the total fraction of calls reached within the response time threshold is high. To find these, we implemented a hill climbing algorithm which starts with a random distribution of nineteen vehicles over the nineteen base locations. In every iteration, we consider a neighboring solution in which we change the home base of one of the vehicles. This solution is evaluated by simulation, and the solution is accepted if it increases the fraction of calls reached within the time threshold.

Note that vehicles are assumed to be identical, and therefore many solutions are in fact equivalent. Hence, in order to reduce the total computation time, we check whether an equivalent solution has been simulated before, and if so, skip this. When changing the home base of one of the vehicles no longer leads to new solutions, i.e., all neighboring solutions are worse, the algorithm terminates and returns the configuration for which we observed the highest fraction of calls reached within the time threshold. This method allows a local optimum to be found in approximately 8 hours.

We repeat this procedure multiple times, using randomly generated starting solutions. This led to eleven different configurations, which together constitute our set of configurations C. For more information about these configurations, we refer the reader to Appendix.

6.3. Results

This section describes the results generated by our BNTS optimization model, using the eleven configurations generated by our local search and their corresponding utilities as input. The optimization model was solved using the Ipopt solver (Wächter & Biegler, 2006), an interior point method solver from COIN-OR. We note that all numerical results are reported within the numerical accuracy of the solver. The model was implemented in Julia/JuMP (Lubin & Dunning, 2015).

Maximizing the Bernoulli–Nash social welfare measure leads to the following time shares: $\lambda_3 = 0.1984$, $\lambda_8 = 0.1864$, $\lambda_9 = 0.2351$ and $\lambda_{11} = 0.3800$. Note that the other seven configurations are not used. This Bernoulli–Nash solution has a social welfare $f_{\text{BN}}^{\max} = 0.8525$.

To better understand this solution, it is helpful to contrast this with solutions for the more traditional utilitarian objective. We do this next.

7. Utilitarian vs Bernoulli–Nash social welfare

In this section we investigate how the Bernoulli–Nash social welfare is related to the often-used utilitarian social welfare[6]. This is done using the same set of 11 near-optimal configurations as before. However, we now consider both the Bernoulli–Nash social welfare and the utilitarian social welfare to be of interest. Since the two objectives are conflicting, we are dealing with a trade-off, and so we have a multi-objective optimization problem. We term this the Bernoulli–Nash Utilitarian Time Sharing (BNUTS) problem:

$$\text{BNUTS: maximize} \quad \begin{pmatrix} f_{\text{U}} \\ f_{\text{BN}} \end{pmatrix} = \begin{pmatrix} \sum_{i\in V} d_i \sum_{c\in C} \lambda_c u_{ic} \\ \prod_{i\in V} (\sum_{c\in C} \lambda_c u_{ic})^{d_i} \end{pmatrix} \tag{5}$$

$$\text{subject to} \quad \sum_{c\in C} \lambda_c = 1 \tag{6}$$

$$0 \le \lambda_c \le 1 \quad \forall c \in C \tag{7}$$

We can use the ϵ method Ehrgott (2005) to compute Pareto-optimal solutions for this BNUTS problem. Formally, we say that a solution $\lambda' = (\lambda'_1, \lambda'_2, \ldots, \lambda'_{11})$ *dominates* another solution $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{11})$ if either $f_{\text{U}}(\lambda') > f_{\text{U}}(\lambda)$ *and* $f_{\text{BN}}(\lambda') \ge f_{\text{BN}}(\lambda)$, or, $f_{\text{U}}(\lambda') \ge f_{\text{U}}(\lambda)$ *and* $f_{\text{BN}}(\lambda') > f_{\text{BN}}(\lambda)$. We say that a solution λ is *efficient* if there is no other solution that dominates it. We can obtain a discretized set of efficient solutions by solving a sub-problem BNUTS(ϵ_{BN}) which maximizes f_{U} subject to a lower bound $f_{\text{BN}} \ge \epsilon_{\text{BN}}$, as follows:

$$\text{BNUTS}(\epsilon_{\text{BN}})\text{: maximize} \quad \sum_{i\in V} d_i \sum_{c\in C} \lambda_c u_{ic} \tag{8}$$

$$\text{subject to} \quad \prod_{i\in V} \left(\sum_{c\in C} \lambda_c u_{ic} \right)^{d_i} \ge \epsilon_{\text{BN}} \tag{9}$$

$$\sum_{c\in C} \lambda_c = 1 \tag{10}$$

$$0 \le \lambda_c \le 1 \quad \forall c \in C \tag{11}$$

Although f_{U} in (8) is linear in λ, problem instances can be created for which f_{BN}, and hence (9), is not convex over the domain given by (10) and (11); we give examples of these in Appendix. Hence, BNUTS(ϵ_{BN}) is not, in general, a convex optimisation problem. Since the problem is not always convex, IPOPT is not guaranteed to find globally optimal solutions. However, as we will see in Fig. 5, our results generated by IPOPT form a concave Pareto frontier as expected, which suggests that the problem of local (not global) optimality has not occurred.

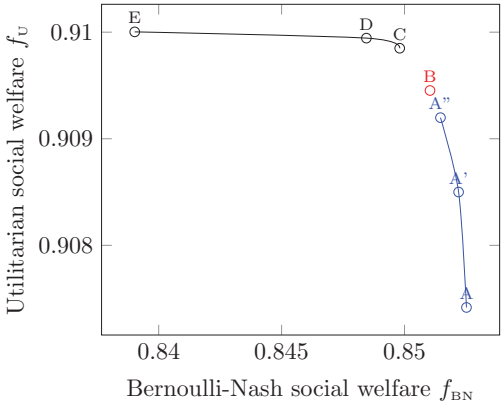


Fig. 5. Pareto efficient solutions for the region Utrecht, with respect to the utilitarian social welfare f_{U} and Bernoulli–Nash social welfare f_{BN}.

Note that one can obtain an equivalent model to construct points on the Pareto frontier by swapping the roles of f_{BN} and f_{U} between the objective (8) and constraint (9). We did numerical experiments to find out if one model is easier to solve than the other, and concluded that the computation times are of similar magnitude. (For the problem instance of our case study, it is just under 10 seconds.) We note that besides the data and problem size, computation times will depend on the particular implementation, as well as the language and solver used.

We solved BNUTS(ϵ_{BN}) for the case study described in Section 6.[7] In this experiment, we used $\epsilon_{\text{BN}} \in \{0.8525, 0.8510, 0.8498, 0.8484, 0.8390\}$, where the first $\epsilon_{\text{BN}} = 0.8525$ value is the single-objective maximum $f_{\text{BN}}^{\max}$ for the Bernoulli–Nash solution found in Section 6.3, and the other values were chosen experimentally to give a good characterization of the Pareto frontier. Those solutions are summarized in columns A, B, C, D and E in Table 1. We observe that solving for $\epsilon_{\text{BN}} = f_{\text{BN}}^{\max}$ gives the lexicographically optimal efficient solution in which f_{BN} achieves its best possible value. We found the other lexicographic solution, being the efficient solution where f_{U} achieves its maximum $f_{\text{U}}^{\max} = 0.9100$, by simply choosing the configuration $c \in C$ giving the largest f_{U} (configuration 8). We then observed that only one such configuration existed, hence the solution is Pareto efficient, and included as column F in Table 1.

Table 1 also includes the *egalitarian* social welfare, denoted f_{E}, for each point. Recall that the egalitarian social welfare is defined as the minimum utility over all nodes, regardless of the demand d_i in that node (as long as $d_i > 0$). That is, it represents the welfare of the individual who is worst off, typically making it a very low number compared to other welfare measures. For this reason, we do not think the egalitarian social welfare is very suitable for comparison; we have however included it for completeness.

Table 1 shows that as ϵ_{BN} increases (i.e., as we put more emphasis on fairness), one has to share time between more configurations. Whereas the utilitarian solution comes down to using one configuration at all times, increasing the Bernoulli–Nash social welfare f_{BN} requires two, three, or four different configurations to be combined. We were interested in better characterizing the

[6] Note that since we defined utilities in terms of a response time threshold, the utilitarian social welfare is equal to the total coverage of the region.

[7] We actually solved an equivalent problem in which (9) was modified by taking the logarithms of both sides.

Table 1
Eight solutions that are Pareto efficient. Only non-zero λ_i values are shown. Where an objective function value f_U or f_{BN} is marked *, it indicates that this is the maximum possible value for this objective, and thus this column denotes a lexicographic solution. Note that some values differ in decimal point values that are not shown; however, f_U is strictly increasing from left to right, whereas f_{BN} is decreasing.

	A	A'	A"	B	C	D	E	F
f_U	0.9074	0.9085	0.9092	0.9095	0.9098	0.9099	0.9100	0.9100*
f_{BN}	0.8525*	0.8522	0.8515	0.8510	0.8498	0.8484	0.8390	0
f_E	0.0072	0.0063	0.0060	0.0062	0.0060	0.0023	0.00001	0
λ_3	0.1984	0.2715	0.3303	0.3482	0.3354	0.1286	0.0008	
λ_8	0.1865	0.3093	0.3853	0.4716	0.6646	0.8714	0.9992	1.000
λ_9	0.2351	0.0945	0.0073					
λ_{11}	0.3800	0.3247	0.2748	0.1802				

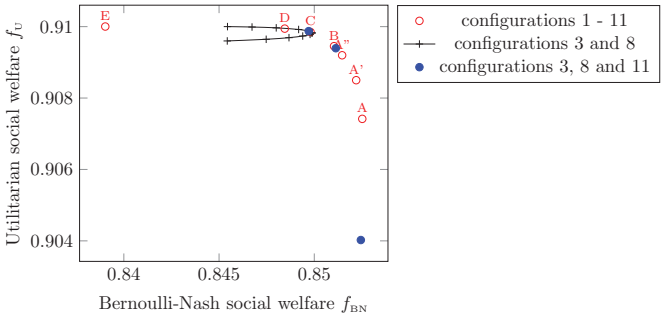


Fig. 6. The utilitarian versus Bernoulli–Nash social welfare that can be obtained using only a subset of the 11 configurations. Note that some of these points are dominated by other points.

Pareto frontier, so we searched for the points where a Pareto efficient solution shifts from three configurations to four. To do this, we performed a binary search between points A and B. We added two extra points marked A' and A" to Table 1, which show how λ_9 decreases to 0 as we move closer to point B.

The ϵ method gives us discrete points on the Pareto frontier, but we expect the frontier to be continuous. We observe that we can easily characterize all solutions formed by time sharing two configurations, let's say 3 and 8, as a parametric function of λ_3. Characterizing frontiers defined by more than two configurations is more difficult, however we expect that these will also result in a smooth and continuous frontier, which we have approximately represented by the lines in Fig. 5.

We further examine the system performance when only configurations 3 and 8 are combined. Such solutions may appeal to an ambulance organization as it may be easier to operate a system with fewer configurations. We investigate all combinations of configurations 3 and 8, including ones that are not on the Pareto frontier. To that end, we construct different solutions by taking time shares defined by $\lambda_3 \in \{0, 0.025, 0.05, 0.1, 0.2, \ldots, 0.9\}$ and $\lambda_8 = 1 - \lambda_3$. We compute the utilitarian and Bernoulli–Nash social welfare of each solution. These values are depicted as the black points in Fig. 6. Furthermore, we also calculate the λ_3 for which the Bernoulli–Nash social welfare measure is maximized (given that only configuration 3 and 8 are combined), to make sure that this point is also included in the graph. Note that all intermediate solutions also exist, therefore we added a curve through those points. We note that solutions for which $\lambda_3 > 0.422079$ are dominated (in a multi-objective sense) by solutions for which $\lambda_3 \leq 0.422079$, and so there would be no value in operating such solutions.

Fig. 6 also shows the results for a similar analysis in which we combine three configurations (configurations 3, 8 and 11). We considered all combinations for which $\lambda_3, \lambda_8, \lambda_{11} \in \{0, 0.1, 0.2, \ldots, 1\}$ and $\lambda_3 + \lambda_8 + \lambda_{11} = 1$. For each combination, we computed the utilitarian and the Bernoulli–Nash social welfare, and depicted a few selected points in Fig. 6.

Figs. 5 and 6 show that for this particular case study there is a trade-off between the Bernoulli–Nash social welfare and the utilitarian social welfare, which raises the question whether this is true in generality. To better understand how these objectives relate, we did some additional experiments. We constructed small test instances and randomly generated utility matrices for all nodes and policies. We found that it is possible to construct small examples where these two objectives are not conflicting; however, in any realistic problem instance we believe that the trade-off will occur. Our reasoning is as follows: in a realistic region, there will be configurations that achieve a high utilitarian welfare, by leaving rural areas uncovered. (This is typically true unless the ambulance provider is completely overstaffed, which is not realistic.) Leaving a rural area permanently uncovered will drive the Bernoulli–Nash social welfare value to zero, while typically increasing the utilitarian social welfare.

8. Discussion

In this paper we viewed the ambulance location problem from a time sharing perspective. We proposed alternating between different ambulance configurations to increase the fairness of the system. The optimal mix is found by computing the maximum of the Bernoulli–Nash social welfare, which requires solving a nonlinear optimization problem. This section discusses the choices made in our approach, as well as its practical applicability and possibilities for extending this work.

- The examples in this paper demonstrate several different utility functions. Section 3 started with a simple a 0–1 function (base coverage). Later, we defined the performance in terms of a probability of being reached within a response time threshold. We believe the latter is a realistic choice for the case study in Section 6, but do not claim that this is the single best way to define utilities. In fact, any measure that maps a distribution of response times to a utility value might be considered. For example, rather than a response time threshold, one might prefer to use *average* response times. Another option would be to use *survival probabilities* (see, e.g., Erkut, Ingolfsson, & Erdoğan, 2008; van den Berg, Kommer, & Zuzáková, 2015). Whichever utility function one chooses to use, the main ideas in this paper remain valid (in the sense that time sharing can be used to improve fairness, and that the Bernoulli–Nash social welfare function can be used to evaluate the fairness of the system). When applying our models to a region in practice, the utility definition might depend on the region and the local rules applicable, and should most definitely include discussions with the ambulance provider.
- When applying our model, it is important to realize that the Bernoulli–Nash social welfare measure is very sensitive to utilities with value zero. This sensitivity makes it important to consider what it actually means when the utility function is zero.

In fact, it can even be a reason to reconsider the chosen utility function: perhaps it makes sense to replace it with one that recognizes there is a non-zero benefit in an ambulance arriving, no matter how long the response time was. When dealing with a utility function that can be zero, one must ensure that all individuals have a nonzero utility under at least one configuration, or otherwise exclude those individuals who do not. Finally, we recommend to take extra caution when estimating those utilities close to zero.

- A positive aspect of our solution method is that determining the utilities for each demand node and each configuration can be done in a preparatory phase: they are input to the optimization model. This allows for the use of complex performance indicators: if they are too difficult to compute, they may be estimated by simulation (as showed in Section 6). Simulation has a few more advantages over analytical models: for example, it allows the user to incorporate practical limitations such as labour legislation or dispatch rules. Additionally, a trace-driven simulation (using historic call records) would further benefit accuracy because using a trace avoids modeling of the incident arrival process - and the potential corresponding modeling errors.
- An interesting extension would be to incorporate the fact that individuals move throughout the region during the day. This mainly affects how the population can be grouped, i.e., basing the demand nodes V on inhabitants is no longer sufficient to model this situation.
- This paper only considers *static* ambulance configurations. That is, in a certain configuration each ambulance has its own home base, from which it always responds. Alternatively, one may consider using dynamic ambulance redeployment policies, e.g., (Alanis, Ingolfsson, & Kolfal, 2013; Gendreau, Laporte, & Semet, 2001; Maxwell, Henderson, & Topaloglu, 2013; Maxwell, Restrepo, Henderson, & Topaloglu, 2010; Schmid, 2012). Dynamic policies are different from time sharing, in the sense that they use real-time information about ambulance unavailability in order to make redeployment decisions. We demonstrate our approach on a system that does not use dynamic redeployment, but the principles we are advocating can still be applied in ambulance systems that do use a dynamic redeployment system. For example, one might compute the Bernoulli–Nash optimum combining several dynamic policies, or a mix between dynamic and static configurations. As long as the utilities for each node and each configuration can be estimated, this is a straightforward extension of our model and does not increase the complexity.
- We discussed the practical applicability of our work with several ambulance providers world wide (Bremer, 2019; Clough, 2019; Hatenboer, 2019; Lindner, 2019; Richards, 2019; Uleberg, 2019; van Breukelen, 2019). In our discussions, we learned that fairness is important to all of them. When we introduced them to the concept of time sharing, we learned that - although they do not use the terminology 'time sharing' - some EMS providers already time share their ambulances. That is, they use different configurations at different times. In this paper, we illustrated the concept with an example that depended on the day of the week; in practice, we see that it is more common for ambulance providers to use a different configuration depending on the time of the day. Note that this is just a different implementation of the same idea. What we can conclude from this is that ambulance providers are using time sharing; however, the question remains if they are using time sharing in order to improve fairness. We argue that this does happen, as illustrated by the following example. The ambulance provider for the region Utrecht in the Netherlands has a base in a rural area in the South East of their region. There is an ambulance stationed at this base during the day; at night time the base is not in use (Bremer, 2019). This ambulance's performance, as measured by the total number of on-time calls, would improve if it were moved into an urban area. This leads us to conclude that the decision to place this ambulance in the rural area is motivated by fairness, not efficiency.
- The example above is complicated because in practice the time of day typically also affects the demand. We chose not to model time-dependent demand in our work; however, it is straightforward to see how one might extend our approach to incorporate this. For an ambulance organization that uses different configurations at different times of the day to deal with changes in demand, one could use our approach to generate sets of different configurations for each of those times.
- When we suggested a time sharing approach, some ambulance providers pointed out the similarities with dynamic redeployment, as they tend to be familiar with that concept (Clough, 2019; Richards, 2019; Uleberg, 2019; van Breukelen, 2019). We argue that time sharing is *at least as practical to use* as a dynamic redeployment policy: in a way it is easier because it does not depend on the current system status. At least two of the ambulance providers we contacted use a software tool that gives real-time advice on the next redeployment move (Clough, 2019; van Breukelen, 2019), and another has expressed interest in using such a tool in the future (Richards, 2019). Current redeployment algorithms tend to focus on maximizing the number of calls reached on time. Where such systems are already implemented, a natural extension of these approaches would be to include fairness in the objective. We conclude that as dispatching systems become more computerized it becomes easier to implement our ideas, although the concept of time sharing itself does not depend on having such a system in place.
- In practice we see examples of ad hoc approaches to address fairness, for example as a result of lobbying by politicians in areas that feel like they are not being treated fairly. This often does not lead to outcomes that are inherently more fair for the population as a whole. Instead, we advocate a more systematic approach where fairness is treated rigorously as part of the decision making.

Acknowledgments

The first author thanks the Netherlands National Institute for Public Health and the Environment (RIVM) for giving access to the travel times for EMS vehicles in the Netherlands. Funding: This work was supported in part by Technology Foundation STW (contract 11986); and the Netherlands Organization for Scientific Research (NWO) in the form of a Rubicon grant (number 5208).

Appendix A

The numerical study in this paper used a set of 11 ambulance configurations. A summary of properties of each of those configurations can be found in Table A1. Moreover, we illustrate one of these configurations, configuration 8, in more detail in Fig. A1. This figure also shows the coverage measured per postal code in our simulation.

In Table A1 we observe that the configurations with the highest utilitarian social welfare (configurations 8 and 3) have a very poor Bernoulli–Nash social welfare and so neither of these is fair when used in isolation. But, we note that both configurations 8 and 3 received positive time shares in our solutions, showing that they can produce fairer outcomes when used in a time-sharing policy.

Table A1
Characteristic properties of each of the 11 configurations used in our case study. The column f_E represents the egalitarian social welfare of the configuration. All decimal point values are rounded.

Configuration	f_U	f_{BN}	f_E	Fraction inhabitants having utility	max # vehicles per base
1	0.903	0	0	0.0023	3
2	0.908	0	0	0.0092	2
3	0.910	0	0	0.0005	2
4	0.903	0	0	0.0085	2
5	0.907	0	0	0.0022	2
6	0.905	0.84	0.003	0	2
7	0.906	0	0	0.0013	2
8	0.910	0	0	0.0022	2
9	0.903	0	0	0.0003	3
10	0.906	0	0	0.0016	3
11	0.908	0	0	0.0032	3

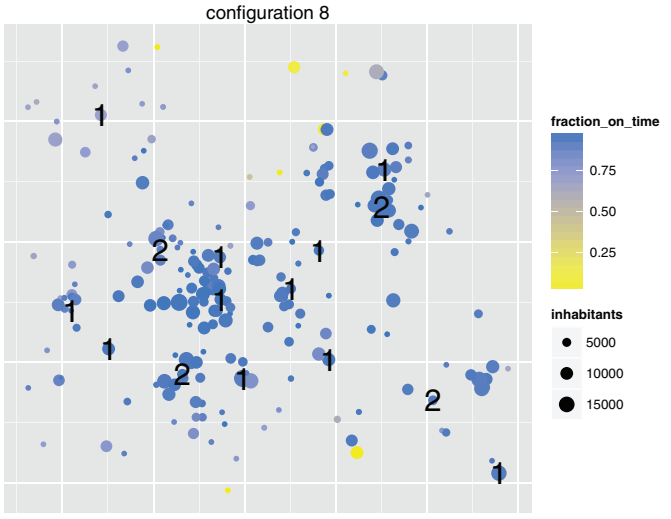
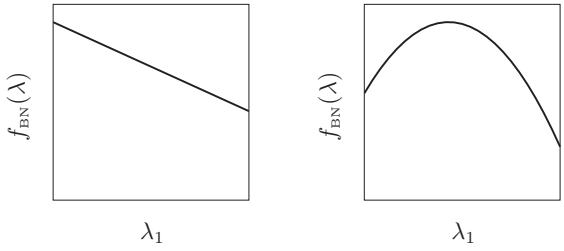


Fig. A1. The numbers show the number of ambulances per location in configuration 8. The nodes show the centers of the postal codes, where the shading shows the coverage observed in simulation, while the size of the nodes depicts the number of inhabitants.



(a) Instance 1 (b) Instance 2

Fig. B2. The Bernoulli–Nash social welfare function behaves differently for different problem instances. Note that this affects the location of the maximum.

Table B3
Utility matrix for a problem instance where $f_{BN}(\lambda)$ is neither convex nor concave.

	Instance 3	
	Node 1	Node 2
Configuration 1	0.2	1
Configuration 2	0.9	0.8
Configuration 3	0	0.1

Appendix B

We next analyze the convexity of (9) over the domain given by (10) and (11). As mentioned before, this depends on the problem instance. Recall that a problem instance consists of demands d_i and utilities u_{ic}. In this appendix we introduce three different problem instances, which show that (9) can be either convex, concave or neither.

Note that in this paper, we defined d_i to be the fraction of demand that is located in node i, so $0 \le d_i \le 1$. However, one might use an alternative definition in which d_i denotes the demand, i.e., the number of inhabitants in node i, in which case it follows that $d_i \ge 1$. In both cases, the optimum is found in the same place, however, the convexity (or concavity) properties might differ between the two.

Instance 1 and 2 both consist of two demand nodes, each with one inhabitant (i.e., $d_1 = d_2 = 1$), and two configurations. Let the utilities be as shown in Table B2. Fig. B2 shows how the Bernoulli–Nash social welfare depends on λ_1, the proportion of time spent in configuration 1. We conclude that f_{BN} can be convex (see Fig. B2) or concave (see Fig. B2), depending on the problem instance.

Table B2
Utility matrices for two problem instances with different convexity properties.

Instance 1			Instance 2		
	Node 1	Node 2		Node 1	Node 2
Configuration 1	0.5	0.2	Configuration 1	0.1	0.5
Configuration 2	1	0.6	Configuration 2	1	0.1

In higher dimensions, one can even construct instances such that $f_{BN}(\lambda)$ is *neither* convex nor concave. The following example shows this.

Instance 3 is a region with two demand nodes, each with one inhabitant (i.e., $d_1 = d_2 = 1$), and *three* configurations. Let the utilities be as shown in Table B3.

Consider two vectors of time shares, $\lambda = [0.1, 0.7, 0.2]$ and $\lambda' = [0.5, 0.3, 0.2]$. Then, the following holds:

$$f_{BN}(0.5\lambda + 0.5\lambda') = 0.3672 > 0.3616 = 0.5 f_{BN}(\lambda) + 0.5 f_{BN}(\lambda')$$

Therefore, $f_{BN}(\lambda)$ is not convex. Similarly, by choosing $\lambda = [0.6, 0.3, 0.1]$ and $\lambda' = [0.1, 0.4, 0.5]$ we obtain that

$$f_{BN}(0.5\lambda + 0.5\lambda') = 0.2541 < 0.2551 = 0.5 f_{BN}(\lambda) + 0.5 f_{BN}(\lambda')$$

and conclude that $f_{BN}(\lambda)$ is not concave.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2019.08.003.

References

Abouelinane, L., Sahin, E., & Jemai, Z. (2013). A review on simulation models applied to emergency medical service operations. *Computers & Industrial Engineering, 66*(4), 734–750.

Alanis, R., Ingolfsson, A., & Kolfal, B. (2013). A Markov chain model for an EMS system with repositioning. *Production and Operations Management, 22*(1), 216–231.

Bremer, P. (2019). *Head of planning at Regionale Ambulance Voorziening Utrecht (RAVU)*. Personal Communication.

Clough, A. (2019). *Manager operational planning & improvement at Ambulance Victoria*. Personal Communication.

Bertsimas, D., Farias, V., & Trichakis, N. (2011). The price of fairness. *Operations Research, 59*(1), 17–31.

Bozorgi-Amiri, A., Jabalameli, M., & e Hashem, S. M. A. (2013). A multi-objective robust stochastic programming model for disaster relief logistics under uncertainty. *OR Spectrum, 35*(4), 905–933.

Brotcorne, L., Laporte, G., & Semet, F. (2003). Ambulance location and relocation models. *European Journal of Operational Research, 147*(3), 451–463.

Ceriani, L., & Verme, P. (2012). The origins of the Gini index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *The Journal of Economic Inequality, 10*, 421–443.

Chanta, S., Mayorga, M. E., & McLay, L. A. (2011). Improving emergency service in rural areas: A bi-objective covering location model for EMS systems. *Annals of Operations Research, 221*(1), 133–159. doi:10.1007/s10479-011-0972-6.

Church, R., & Revelle, C. (1974). The maximal covering location problem. *Papers of the Regional Science Association, 32*, 101–118.

Cobb, C., & Douglas, P. (1928). A theory of production. *The American Economic Review, 18*(1), 139–165.

Coulter, P. (1980). Measuring the inequity of urban public services. *Policy Studies Journal, 8*(5), 683–698.

Daskin, M. (1983). A maximum expected location model: Formulation, properties and heuristic solution. *Transportation Science, 7*, 48–70.

Daskin, M. (1995). *Network and discrete location: models, algorithms, and applications*. New York: Wiley.

Ehrgott, M. (2005). *Multicriteria optimization*. Springer.

Elloumi, S., Labbé, M., & Pochet, Y. (2004). A new formulation and resolution method for the p-center problem. *INFORMS Journal on Computing, 16*(1), 84–94.

Enayatia, S., Mayorgab, M., Toro-Diaz, H., & Albert, L. (2019). Identifying trade-offs in equity and efficiency for simultaneously optimizing location and multipriority dispatch of ambulances. *International Transactions in Operational Research, 26*, 415–438.

Erkut, E., Ingolfsson, A., & Erdoğan, G. (2008). Ambulance location for maximum survival. *Naval Research Logistics (NRL), 55*(1), 42–58. doi:10.1002/nav.20267.

Erkut, E., & Neuman, S. (1992). A multiobjective model for locating undesirable facilities. *Annals of Operations Research, 40*(1), 209–227.

Felder, S., & Brinkmann, H. (2002). Spatial allocation of emergency medical services: minimising the death rate or providing equal access? *Regional Science and Urban Economics, 32*(1), 27–45.

Gendreau, M., Laporte, G., & Semet, F. (1997). Solving an ambulance location model by tabu search. *Location Science, 5*(2), 75–88.

Gendreau, M., Laporte, G., & Semet, F. (2001). A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing, 27*, 1641–1653.

Gini, C. (1912). *Variabilità e Mutuabilità. contributo allo studio delle distribuzioni e delle relazioni statistiche*. Bologna: C. Cuppini.

Goemans, M., & Skutella, M. (2004). Cooperative facility location games. *Journal of Algorithms, 50*(2), 194– 214 doi:https://doi.org/10.1016/S0196-6774(03)00098-1.

Goldberg, J., Dietrich, R., Chen, J., & Mitwasi, M. (1990). Validating and applying a model for locating emergency medical services in Tucson, AZ. *Euro, 34*, 308–324.

Henderson, S., & Mason, A. (2005). Ambulance service planning: Simulation and data visualisation. In *Operations research and health care: A handbook of methods and applications* (pp. 77–102). Boston: Springer.

Hoogeveen, M. (2010). Report ambulance care in Europe. *Technical Report*.

Hatenboer, J. (2019). *Policy and innovation manager at UMCG ambulancezorg*. Personal Communication.

Jagtenberg, C., Bhulai, S., & van der Mei, R. (2015). An efficient heuristic for real-time ambulance redeployment. *Operations Research for Health Care, 4*, 27–35.

Kariv, O., & Hakimi, S. (1979). An algorithmic approach to network location problems. part 1: The p-centers. *SIAM Journal on Applied Mathematics, 37*, 513–538.

Kommer, G., & Zwakhals, S. (2011). Modellen referentiekader ambulancezorg 2008 Documentatie rijtijden- en capaciteitsmodel. *Rapport 270412001/2011*.

Lindner, T. W. (2019). *Director at regional centre for emergency medical research and development in Western Norway (RAKOS)*. Personal Communication.

Larson, R. (1974). A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers & Operations Research, 1*(1), 67–95.

Leclerc, P., McLay, L., & Mayorga, M. (2012). Modeling equity for allocating public resources. In M. Johnson (Ed.), *Community-based operations research* (pp. 97–118). New York: Springer.

Li, X., Zhao, Z., Zhu, X., & Wyatt, T. (2011). Covering models and optimization techniques for emergency response facility location and planning: A review. *Mathematical Methods of Operations Research, 74*(3), 281–310. doi:10.1007/s00186-011-0363-4.

Lubin, M., & Dunning, I. (2015). JuMP: A modeling language for mathematical optimization. *INFORMS Journal on Computing, 27*(2), 238–248.

Marsh, M., & Schilling, D. (1994). Equity measurement in facility location analysis: A review and framework. *European Journal of Operational Research, 74*(1), 1–17 doi:https://doi.org/10.1016/0377-2217(94)90200-3.

Maxwell, M., Henderson, S., & Topaloglu, H. (2013). Tuning approximate dynamic programming policies for ambulance redeployment via direct search. *Stochastic Systems, 3*(2), 322–361.

Maxwell, M., Restrepo, M., Henderson, S., & Topaloglu, H. (2010). Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing, 22*, 226–281.

McAllister, D. (1976). Equity and efficiency in public facility location. *Geographical Analysis, 8*, 47–63.

McLay, L., & Mayorga, M. (2013). A dispatching model for server-to-customer systems that balances efficiency and equity. *Manufacturing and Service Operations Management, 15*(2), 205–220. doi:10.1287/msom.1120.0411.

Moulin, H. J. (2003). *Fair division and collective welfare*. The MIT Press. doi:10.1007/s10888-004-7784-8.

Mulligan, G. (1991). Equality measures and facility location. *Papers in Regional Science, 70*(4), 345–365. doi:10.1007/BF01434593.

Richards, D. (2019). *Business intelligence manager at St John New Zealand*. Personal Communication.

Rakas, J., Teodorovic, D., & Kim, T. (2004). Multi-objective modeling for determining location of undesirable facilities. *Transportation Research Part D: Transport and Environment, 9*(2), 125–138.

Repede, J., & Bernardo, J. (1994). Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *European Journal of Operational Research, 75*, 567–581.

Ridler, S., Mason, A., & Raith, A. (2017). A simulation package for emergency medical services. In *Proceedings of the 51st annual conference of the ORSNZ*.

Schmid, V. (2012). Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research, 219*, 611–621.

Schmid, V., & Doerner, K. (2010). Ambulance location and relocation problems with time-dependent travel times. *European Journal of Operational Research, 207*(3), 1293–1303.

Stone, D. (2001). *Policy paradox: the art of political decision making*. New York: Norton.

Uleberg, O. (2019). *Flight physisican at Olav's University Hospital*. Personal Communication.

UoA EMS Research Julia package for emergency medical services simulation. https://github.com/uoa-ems-research/JEMSS.jl.

van Breukelen, A. (2019). *Policy officer at regionale ambulancevoorziening flevoland / Gooi & Vechtstreek*. Personal Communication.

van den Berg, P., & Aardal, K. (2015). Time-dependent MEXCLP with start-up and relocation cost. *European Journal of Operations Research, 242*(4), 358–368.

van den Berg, P., Kommer, G., & Zuzáková, B. (2015). Linear formulation for the maximum expected coverage location model with fractional coverage. *Operations Research for Health Care*. https://doi.org/10.1016/j.orhc.2015.08.001.

Wächter, A., & Biegler, L. T. (2006). On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Mathematical Programming, 106*(1), 25–57.

Zhan, S., & Liu, N. (2011). A multi-objective stochastic programming model for emergency logistics based on goal programming. In *Proceedings of the IEEE Computational sciences and optimization, fourth international joint conference on.* (pp. 640–644).