EFFICIENCY AND FAIRNESS
IN AMBULANCE PLANNING

Caroline Jagtenberg
We discuss how to position ambulances across an EMS region in a ‘fair’ way. Ambulance literature often focuses on maximizing the number of people served, regardless of where they live. This is equivalent with optimizing a utilitarian Social Welfare Function (SWF). It is well known that such an approach benefits people living in cities, at the cost of people living in remote areas. An often mentioned alternative is equity: providing the same service to people at every location. However, this gives so much focus on helping people in remote locations, that it usually leads to poor overall performance. Instead, we propose to use the so-called Bernoulli-Nash SWF. This may be viewed as an appealing compromise between the two solutions above. We formulate and solve models that maximize the Bernoulli-Nash social welfare. The most straightforward model maximizes coverage, but we also use more complex measures such as survival functions. We juxtapose the Bernoulli-Nash optimal solution with the Utilitarian optimum, and show how the results differ depending on the load of the system. Calculations are done for a realistic EMS region in the Netherlands.

This chapter is based on:

5.1 Introduction

A key issue in EMS planning is the ambulance location problem: how and where to locate vehicles in order to effectively cover future demand. Much research has been focused on solving variants of this problem, and the majority has approached the problem from the same angle: their objective is to help as many patients as possible. For example, in overview papers from 2003 [22] and 2011 [71] practically all models aim to maximize the (expected) coverage.

Maximizing the number of people served seems natural in the context of ambulance planning. In fact, at first sight it seems hard to reason that we should help fewer people rather than more. However, as we will argue in this chapter, there may be reasons to consider ambulance configurations other than the one that maximizes the number of people served. For example, in order to serve as
many people as possible given a fixed number of resources, a planner inadvertently moves resources to densely populated areas - at the cost of people living in rural areas. This certainly is an efficient use of resources, but the question arises how to distribute ambulances in a *fair* way.

There are examples in literature of models that aim for equity. McLay et al. [48] review how equity can be modeled in the context of allocating public resources, and conclude that in general there is no single, best way to do so. However, when it comes to EMS problems they note that equity is almost always interpreted as ‘having equal outcomes for all patients, regardless of where they are living’.

Therefore, the aim for equity often results in egalitarian (also known as maximin) models: models that maximize the level of service to the people that are the hardest to reach. It is not surprising that this generally leads to poor performance of the overall system. This is a major drawback of those models, and it would be hard to convince anyone to actually apply such solutions in practice. Furthermore, we disagree with the statement that ‘equal outcomes for everyone’ is the correct definition of fairness in an ambulance context. In fact, we argue that ‘equal outcomes for everyone’ is far from ‘fair’, because it is obvious that it requires much more resources to serve those people who live far away. Given these two extreme solutions (maximizing the number of people served, and equal service for everyone), we believe that a truly fair solution is some sort of compromise between them.

There exist a few papers on ambulance location problems that explicitly include a form of fairness in the objective. For example, in [31] the sum of ‘envy’ among all demand zones is minimized. In [32] the authors propose three bi-objective covering models, for which fairness is a secondary objective. They consider the following three options (1) minimize the maximum distance between each uncovered demand zone and its closest opened station, (2) minimize the number of uncovered rural demand zones, and (3) minimize the number of uncovered demand zones. Although these papers are all based on ideas similar to ours - that some form of fairness should be incorporated in the objective - none of these take our approach, which we will describe next.

In this chapter we will view ambulance location problems from the perspective of social welfare. Social welfare is measured as a function of the ‘utilities’ of individuals or subgroups of a society. That way, different social welfare functions (SWFs) represent different objectives on how to balance fairness and efficiency. For example, maximizing the total number of people served is equivalent with a so-called *utilitarian* SWF: that is, maximize the sum of the individual utilities. Alternatively, aiming for equal outcomes for everyone would correspond to an egalitarian SWF. In this chapter, we propose and investigate a third option, the so-called Bernoulli-Nash SWF. The Bernoulli-Nash SWF is defined as the *product* of individual utilities. These three different SWFs can be visualized by so-called *social indifference* curves: these are solutions which are equivalent in terms of

---

1For more background regarding equity in ambulance planning, see also [81]. They discuss several ideas regarding equity, e.g., server equity (as opposed to patient equity). We consider this an interesting alternative point of view, but not the focus of our work.
social welfare. For a small problem instance with just two individuals, these curves are shown in Figure 5.1. To the best of our knowledge, the Bernoulli-Nash SWF has not previously been applied to ambulance location problems. We juxtapose the Bernoulli-Nash optimal solution with the often-used utilitarian optimum, and show how the results differ depending on the load of the system.

![Figure 5.1](image)

(a) Utilitarian SWF  (b) Bernoulli-Nash SWF  (c) Egalitarian SWF

**Figure 5.1** Social indifference curves for three different social welfare functions.

The Bernoulli-Nash SWF is related to the Nash bargaining problem [88], which was defined for two players. As Nash defines the problem, both players have to agree on the outcome, or otherwise a so-called disagreement point (denoted $\xi$) is reached. Let $f$ and $g$ be the utility functions for the two players. The optimal solution (obtained at the point $\theta$) is defined relative to this $\xi$: it is the maximum of the SWF:

$$\arg\max_{\theta} (f(\theta) - f(\xi))(g(\theta) - g(\xi)).$$

As [19] describes it: “When the Nash bargaining solution is used, it is to predict what the result would be, under certain ideal circumstances, if specimens of *homo economicus* were to bargain optimally.”

To interpret this in the context of ambulance planning, imagine these two players to be patients living in different locations - or two communities, together deciding on how to distribute their ambulances. It is not unreasonable to imagine this decision to be a bargaining process. Let us say that the disagreement point is that no ambulances will be acquired for the region, so we have $\xi = 0$. An optimal Bernoulli-Nash social welfare then comes down to the same as Nash’s bargaining solution.

The definitions of SWFs are based on *utilities*; however, it is not immediately clear how the utility of a patient should be defined. This utility should somehow represent the happiness of that patient, depending on his location with respect to the locations of ambulances. Typically, their happiness will be a function of the (expected) ambulance response time, e.g., the probability that an ambulance will reach the patient within a certain time threshold. Other - more advanced - measures are also possible; our work includes three different definitions of utility.

The rest of this chapter is structured as follows. In Section 5.2 we define the problem. In Section 5.3 we define three different measures that can be used as
utilities. Section 5.4 provides a small example that helps to build intuition for the problem. In Section 5.5 gives our optimization models, followed by a case study in Section 5.6. We finish with a discussion in Section 5.7.

5.2 Problem formulation

In this chapter we focus on the allocation of ambulances to a set of base stations with known locations. These ambulances respond to incidents that occur at demand nodes. Denote the set of demand nodes $V$. After completing service, the ambulance should return to its own home base. We address the question of which ambulance base locations to open as well as how to divide the vehicles over the bases. These questions may be addressed separately, but then obviously optimality is not guaranteed. Instead, we aim to find a base location for each vehicle, choosing from a large set of potential locations - many of which may be unused in the final solution. Note that multiple vehicles are allowed to have the same home base.

Our goal is to find a distribution of vehicles over bases that maximizes the Bernoulli-Nash SWF. The Bernoulli-Nash SWF is defined as the product of individual utilities. In the context of ambulance planning, we write the product as follows. Let $u_i$ be the utility of a person at node $i$, and let $d_i$ be the demand fraction at node $i, i \in V$. The Bernoulli-Nash SWF is then given by

$$\prod_{i \in V} u_i^{d_i}.$$

In this chapter, we compare the Bernoulli-Nash SWF to the often-used utilitarian SWF, which is denoted as

$$\sum_{i \in V} d_i u_i.$$

For completeness, we also state the egalitarian SWF:

$$\min_{i \in V} u_i.$$

Before we elaborate on definitions and interpretations of the utilities $u_i$, we find it useful to illustrate the differences between the three SWFs above. To that end, we introduce a small example of an ambulance location problem. Unlike the rest of this chapter, this particular example is not meant to be realistic. We use a simple utility measure that is good for demonstrative purposes because it helps to build intuition on (1) how the Bernoulli-Nash SWF relates to the two other SWFs, and (2) why we think the Bernoulli-Nash SWF is a somehow reasonable measure for positioning ambulances.
5.2. Problem formulation

Example. Imagine two areas or villages, together acquiring one ambulance. We model these areas as nodes, labeled 1 and 2, and the ambulance may be positioned anywhere along the line between them. Without loss of generality, assume the distance between the two nodes is normalized to 1. The two nodes both contain a proportion of the demand, \( d_1 \) and \( d_2 \), such that \( 0 < d_1 < 1 \) and \( d_1 + d_2 = 1 \). The position of the ambulance can be defined by its distance from node 1, let us call this distance \( r \). This is depicted in Figure 5.2.

![Figure 5.2](image)

**Figure 5.2** A toy example for the ambulance location problem

We define the utility \( u_i \) of an inhabitant of node \( i \), \( i \in \{1, 2\} \), to be equal to 1 minus the distance between \( i \) and the ambulance. This means, for example, that if the ambulance is placed at node 1, then \( u_1 = 1 \) and \( u_2 = 0 \). More generally, if the ambulance is located distance \( r \) from node 1, then \( u_1 = 1 - r \) and \( u_2 = r \).

Now let us compare the optimal solutions for the three SWFs. The utilitarian SWF is maximized when the ambulance is placed at the node with the most inhabitants. If \( d_1 = d_2 \), then all solutions are optimal from a utilitarian perspective. For the egalitarian SWF on the other hand, it is straightforward to see that this is maximized when \( r = 0.5 \). We finish with Bernoulli-Nash optimum, which can be found as follows. Recall that the Bernoulli-Nash SWF is given by \( u_1^{d_1} \times u_2^{d_2} \). This means that if we place the ambulance at either of the two nodes (i.e., \( r = 0 \) or \( r = 1 \)) the Bernoulli-Nash SWF is equal to zero. Therefore, for an optimal solution \( 0 < r < 1 \) will hold. Furthermore, observe that maximizing \( u_1^{d_1} \times u_2^{d_2} \) is equivalent with maximizing \( d_1 \log(u_1) + d_2 \log(u_2) \). Denote this function \( f(r) \), as the utilities can be expressed in terms of \( r \):

\[
f(r) = d_1 \log(1 - r) + d_2 \log(r).
\]

The maximum of \( f(r) \) is attained when the derivative is equal to zero:

\[
f'(r) = \frac{-d_1}{1 - r} + \frac{d_2}{r} = 0,
\]

which is equivalent with

\[d_1 r = d_2 (1 - r)\]

If we now use that \( d_2 = 1 - d_1 \) we can solve the equation:

\[r = d_2.\]

That is, the ambulance should be positioned between the two nodes, such that the ratio of the distances is inversely proportional to the ratio of the demands. In our opinion, this corresponds to a fair distribution of ambulances, that balances the distances depending on the ratio of the inhabitants, and thereby provides an attractive compromise between the utilitarian and egalitarian solution.
The simple utility function defined in the example above is not a common performance measure for ambulances. Therefore, we continue by introducing other, more realistic, utilities.

5.3 Utilities

In this section we discuss three different definitions for utility that are related to key performance indicators used by most ambulance practitioners and researchers. These three utilities will be used and compared throughout the rest of this chapter.

The utility of demand node \( i \in V \) can be interpreted as a measure of how happy an inhabitant of \( i \) is with the ambulance configuration. Although ambulance literature typically does not use the term utility, several models exist that optimize for different quantities. Such a quantity is almost always a function of the ambulance response time.

A straightforward example of a utility is single coverage (also known as regional coverage). This quantity is defined in terms of a response time threshold (RTT). Let \( T \) denote the value of this threshold. Simply put, a demand node has (regional) coverage 1 if there is a vehicle positioned at most \( T \) minutes away. Otherwise, the node is said to have coverage 0. The early models in ambulance location literature typically used single coverage as their utility (e.g., [34, 110]). Note that this definition of coverage is somewhat shortsighted: (1) a single vehicle may not be enough to fully satisfy inhabitants of node \( i \) (because sometimes this vehicle will be busy serving other patients), and (2) one may argue that such a strict threshold is somewhat unrealistic: a vehicle that is slightly further than \( T \) away should be worth almost as much as a vehicle at distance \( T \). It may be clear that single coverage is an overly simplified measure to use for our utility; we aim for a more sophisticated measure instead. We next describe how to overcome the issues described above.

First of all, we should account for ambulance unavailability: literature describes various ways to do this. The Maximum Expected Coverage Location Problem (MEXCLP) [36] uses a so-called busy fraction: a fixed parameter that represents the probability that any given ambulance is busy at any given time. In this model, ambulances are assumed to operate independently. It should be noted that modeling unavailability this way is somewhat of a simplification: in reality vehicles are not independent, and moreover, the busy probabilities might differ between vehicles. Other ways of modeling ambulance unavailability include Erlang loss models [96], scenarios [40] and simulation, although the latter is more common when planning in a dynamic context (e.g. [118]). We decided to model ambulance availability using a busy fraction. Although there exist alternatives that may be more accurate, incorporating this in the utility makes the model far more complex. We chose not to do this, since our objective - a product of utilities - is already a hard function to optimize. This is further addressed in the discussion (Section 5.7).
Second, the model becomes more realistic if we can relax the assumption that the utility is a 0-1 function of the response time. If instead we want to use some continuous function of the response time, the model becomes harder. However, it remains possible to solve it with modern solvers and hardware. We implemented our models using three different definitions of utility, two of which relax this 0-1 assumption. The rest of this section describes those three utilities.

5.3.1 Definitions

Deterministic coverage
The most straightforward utility function that we will consider is a coverage model. Throughout the chapter we refer to this utility function as \textit{deterministic coverage}, where ‘deterministic’ refers to the underlying assumptions in the driving time model.

The coverage of demand node \(i\) can be defined in terms of how many ambulances are located within the RTT of \(i\): let \(k(i)\) denote this number of vehicles. The utility (coverage) of \(i, i \in V\) is then given by the probability that at least one of these vehicles is idle, i.e., \(c_i = 1 - q^{k(i)}\).

It is appealing to assume travel times are deterministic, for several reasons. First of all, it is quite difficult to accurately estimate a response time - let alone a whole distribution of response times. Second, stochastic travel times are harder to incorporate in optimization models. Doing this leads to less efficient solutions and scalability issues.

Stochastic coverage
As opposed to deterministic coverage (as discussed above), this section deals with coverage when travel times are stochastic. Although it is appealing to assume that travel times are deterministic, it may be argued that stochastic travel times are more realistic, see e.g. \cite{54}. In this context, our values for \(\tau\) will be interpreted as expected travel times.

The utility is then defined as the \textit{probability} that an ambulance will reach the scene of the incident within the RTT. We can compute this probability if we assume that (1) ambulance unavailability may be modeled using a busy fraction, (2) the closest idle ambulance always responds, and (3) the distribution of the travel times is known.

As described in \cite{17}, stochastic travel times for Dutch ambulances may be approximated by a normal distribution with a coefficient of variation of 0.25. Therefore, we compute the probability that an ambulance with expected travel time \(\tau\) arrives within 12 minutes (the RTT). The result is depicted in Figure 5.3.

Survival
While the majority of ambulance literature deals with response time threshold, it is sometimes argued that these do not adequately differentiate between consequences of different response times. That is, even if one uses a well-chosen
distribution of stochastic travel times, such a model still fails to accurately represent the happiness of a patient given his or her response time. Such discussions usually result in an argument for using a survival function: a monotonically decreasing function of the response time that returns the probability of survival for the patient.

We analyzed several survival functions mentioned in literature. As noted in [38], almost all of the published research relating survival rates to EMS response times focuses on cardiac arrest. Survival is typically interpreted as ‘survival until discharge from the hospital’. For the purpose of this chapter, such practical considerations are of limited importance. Our main goal is to show that our model can find a solution resulting in maximum survival, the specific survival function used is mainly illustrative of the idea.

We implemented two different survival functions. The first was introduced by De Maio et al. in [73] and is given by:

\[ f(\tau) = (1 + e^{0.679+0.262\tau})^{-1}. \]  

The second (by Valenzuela et al. [113]) uses variables that measure the time from collapse to CPR \((\tau_{CPR})\), and from collapse to defibrillation \((\tau_{defib})\). The survival probability is then given by:

\[ f(\tau_{CPR}, \tau_{defib}) = (1 + e^{-0.260+0.106\tau_{CPR}+0.139\tau_{defib}})^{-1}. \]  

As in [38], we assume that CPR is performed by the responding EMS unit immediately upon arrival, and defibrillation is performed one minute after arrival. Equation 5.2 then becomes:

\[ f(\tau) = (1 + e^{-0.260+0.106\tau+0.139(1+\tau)})^{-1}. \]  

The two survival functions (5.1) and (5.3) are depicted in Figure 5.4.

**Figure 5.3** The probability that an ambulance will arrive within the time threshold, when driving times are normally distributed with a coefficient of variation of 0.25.
Figure 5.4 The two different survival functions. The horizontal axis represents the response time in minutes. The vertical axis is the probability of survival.

The results for the two different survival functions turn out to be quite similar. Therefore, we decided to only include results for one of them in this chapter. Our choice between the ‘de Maio et al.’ and the ‘Valenzuela et al.’ survival functions was mainly based on the following observation. The probability of survival is remarkably low: even if an ambulance is present right away, the probability of survival for the most optimistic function is still less than 55%. This may be accurate for cardiac arrest, but generally speaking one might hope that the survival probability of an ambulance request would be higher. Therefore, we decided to use the highest survival function among the two (Equation (5.3)).

5.3.2 Notation

This chapter compares two different SWFs as objectives. Furthermore, numerical work is done for three different utilities (as described in Section 5.3). When combined, this leads to six different objectives.

Our goal is to show the differences between these SWFs and the solutions that correspond to their optimum. We want to emphasize the difference between objectives and models. While a model has a certain objective, we can evaluate its solution with a different objective (that is precisely what we will do in Section 5.6). We next introduce notation in order to make the distinction between the six objectives and corresponding six models.

We will denote an objective as

$$\text{SWF}_{utility}$$,

where SWF $\in \{U, BN\}$ means either the utilitarian resp. the Bernoulli-Nash social welfare. We denote utility $\in \{\text{cov}, \text{stoch}, \text{survival}\}$ to represent either deterministic coverage, stochastic coverage or survival according to the survival
function by Valenzuela et al [113].

For the optimization models, we write

\[ \text{model}_{\text{utility}}. \]

The utilitarian models are MEXCLP\(^2\), MEXSLP\(_{\text{stoch}}\) and MEXSLP\(_{\text{survival}}\). The models that optimize the Bernoulli-Nash social welfare are denoted as MaxFairness\(_{\text{utility}}\) (because the Bernoulli-Nash SWF contains a form of fairness that the utilitarian SWF is lacking). As before, utility \(\in\{\text{cov, stoch, survival}\}\).

Next, we introduce a small problem instance for which we can optimize the social welfare by brute force. This allows us to provide some insights, before we continue with our optimization models.

### 5.4 A small example

For illustrative purposes, we analyze a fictional region with two demand nodes (demand \(d_1 = 0.1\) and \(d_2 = 0.9\)). We take stochastic coverage as our utility function, and define the expected travel time between the two nodes to be 30 minutes. Both nodes are possible bases and our task is to place 2 ambulances in the region. Let the RTT to be twelve minutes, which - to the driving time distribution described in Section 5.3.1 - implies that the probability that an ambulance arrives on time while departing from the other node is \(\approx 0.0082\). Conversely, when the ambulance departs from the same node as where the incident is, the probability of being on time is \(\approx 1\). In this theoretical example we let the average busy fraction be \(q = 0.3\).

For this problem instance, there are only three different solutions. We compute the utilitarian and the Bernoulli-Nash social welfare for each of those solutions to find the following optima: the Bernoulli-Nash optimal solution has one vehicle in each zone, while the utilitarian optimum\(^3\) has two vehicles in the zone with the largest demand.\(^4\) Table 5.1 shows the obtained social welfare of these two solutions, for both the Bernoulli-Nash and utilitarian SWF.

<table>
<thead>
<tr>
<th>model</th>
<th>(BN_{\text{stoch}})</th>
<th>(U_{\text{stoch}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaxFairness(_{\text{stoch}})</td>
<td>0.70172</td>
<td>70.17%</td>
</tr>
<tr>
<td>MEXSLP(_{\text{stoch}})</td>
<td>0.56629</td>
<td>81.97%</td>
</tr>
</tbody>
</table>

Table 5.1 Max fairness vs MEXSLP, where the utility is stochastic coverage.

\(^2\)Note that the utility of MEXCLP is always (deterministic) coverage, hence we do not have to write MEXCLP\(_{\text{cov}}\) explicitly.

\(^3\)This is equivalent to the optimal solution for the Maximum Expected Survival Location Problem (MEXSLP).

\(^4\)The egalitarian optimum in this case also places one vehicle in each zone, but we will not focus on that.
The brute force approach that we used above obviously does not scale well. The next section introduces optimization models that allow us to compute optimal solutions for larger, realistic problem instances.

5.5 Methods

In this section we introduce the models used to optimize the Bernoulli-Nash SWF, which is the main contribution of this chapter. Our goal is to juxtapose this solution with the utilitarian optimum, hence, we also recap the corresponding utilitarian optimization models (MEXCLP [36] and MEXSLP [17]).

To define our models, we first introduce the notation in Table 5.2.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>The set of ambulances.</td>
</tr>
<tr>
<td>$V$</td>
<td>The set of demand locations.</td>
</tr>
<tr>
<td>$W$</td>
<td>The set of possible base locations, $W \subseteq V$.</td>
</tr>
<tr>
<td>$T$</td>
<td>The response time threshold.</td>
</tr>
<tr>
<td>$q$</td>
<td>The busy fraction.</td>
</tr>
<tr>
<td>$d_i$</td>
<td>The fraction of demand in $i$, $i \in V$.</td>
</tr>
<tr>
<td>$\tau_{ij}$</td>
<td>The driving time from $i$ to $j$ with siren turned on, $i, j \in V$.</td>
</tr>
<tr>
<td>$n_j$</td>
<td>The number of ambulances positioned at $j$, $j \in W$.</td>
</tr>
</tbody>
</table>

Table 5.2 Notation.

In the optimization models that we use, it is somewhat implicitly assumed that one always sends the closest idle ambulance.

As described in Section 5.3, we do numerical work for three different utilities. Optimizing for stochastic coverage or optimizing survival is done using the same model: they only differ in numerical input. It is also possible to use this same model for deterministic coverage; however, in this case a simpler and faster model is available. This results in four different models (two for each SWF), which we describe in the following subsections. We implemented them in Julia/JuMP [72], using Gurobi [49] as our solver.

5.5.1 Coverage optimization models

We next describe the optimization models that use coverage as utility.

Utilitarian SWF (MEXCLP)

Maximizing the utilitarian SWF with (deterministic) coverage as a utility is equivalent to the MEXCLP [36] (see also Section 1.2.1). MEXCLP maximizes the total coverage throughout the region.

Recall that $d_i$ are parameters that represent the demand in zone $i$. We introduce variables as described in Table 5.3. The MEXCLP model uses parameters $W_i$, defined as the set of potential base locations that cover demand node $i(i \in V)$. That is, $W_i = \{ j \in W : \tau_{ji} \leq T \}$. The MILP model can then be formulated as:
variable & for & range & meaning \\ 
\hline
$c_i$ & $i \in V$ & $[0,1]$ & coverage of zone $i$ \\ 
$x_j$ & $j \in W$ & $1, \ldots, |A|$ & number of vehicles positioned at base $j$ \\ 
$y_{ik}$ & $i \in V$, $k = 1, \ldots, |A|$ & $\{0,1\}$ & there are at least $k$ ambulances near zone $i$ \\ 
\hline

Table 5.3 Interpretation of variables for the MILP formulation that maximizes coverage.

Maximize $\sum_{i \in V} d_i c_i$

subject to

\begin{align*}
  c_i & \leq \sum_{k=1}^{|A|} (1 - q)q^{k-1} y_{ik}, \\
  \sum_{j \in W_i} x_j & \geq \sum_{k=1}^{|A|} y_{ik}, \quad i \in V, \\
  \sum_{j \in W} x_j & \leq |A|, \\
  x_j & \in \mathbb{N}, \quad j \in W, \\
  y_{ik} & \in \{0,1\}, \quad i \in V, k = 1, \ldots, |A|, \\
  c_i & \in [0,1], \quad i \in V.
\end{align*}

Note that variables $c_i$ are not strictly necessary to define this model. However, we included them for ease of reading (and this allows us to make a clear comparison with the model that maximizes the Bernoulli-Nash SWF). Note that for Equation (5.4) equality actually holds.

**Bernoulli-Nash SWF**

The optimization model that maximizes the Bernoulli-Nash SWF is given by:

Maximize $\sum_{i \in V} d_i \log(c_i)$

subject to

\begin{align*}
  c_i & \leq \sum_{k=1}^{|A|} (1 - q)q^{k-1} y_{ik},
\end{align*}

\[\text{where } d_i, c_i \text{ are the same as in the MILP model.}\]
5.5. Methods

\[ \sum_{j \in W} x_j \geq \sum_{k=1}^{\mid A \mid} y_{ik}, \quad i \in V, \]
\[ \sum_{j \in W} x_j \leq \mid A \mid, \]
\[ x_j \in \mathbb{N}, \quad j \in W, \]
\[ y_{ik} \in \{0, 1\}, \quad i \in V, k = 1, \ldots, \mid A \mid, \]
\[ c_i \in [0, 1], \quad i \in V. \]

Since this is not a linear model, we approximate the logarithm in Equation (5.5) with piecewise linear functions. Thereto, we add a variable \( l_i \) for all demand nodes \( i \in V \). The objective (5.5) is then replaced by

\[
\text{Maximize } \sum_{i \in V} d_i l_i. \tag{5.6}
\]

We start by introducing a few upper bounds on the value of \( l_i \) (for each \( i \in V \)), by adding lines that are tangent to the logarithm at different points, as depicted in Figure 5.5. These lines hold as upper bounds on the value of \( l_i \). Then, we solve the MILP and analyze the result. If it turns out that \( l_i > \log(c_i) + \varepsilon \) (here, \( \varepsilon \) is our tolerance), we add another constraint that bounds the value of \( l_i \) to the line tangent to \( \log(c_i) \) at point \((c_i, \log(c_i))\). We continue until our piecewise linear approximation of the logarithm is accurate enough, i.e., all values of \( l_i \) are within tolerance of \( \log(c_i) \).\(^5\)

5.5.2 Survival optimization models

In this section we generalize the models from Section 5.5.1, using a so-called survival function. A survival function maps a response time to a survival probability. Every survival function \( f(t) \) is monotonically decreasing, i.e., \( f(t') \leq f(t) \) for all \( t' > t \) (however, note that this is not a necessary condition for our model). Note that we can pre-compute all survival probabilities. Given the driving time \( t_{ji} \) from base \( j \) to demand zone \( i \), we compute probabilities \( p_{ji} = f(t_{ji}) \) for all \( j \in W, i \in V \). Hence, these probabilities are parameters of our model, not decision variables. As before, \( d_i \) are parameters that represent the demand in zone \( i, i \in V \). We introduce variables as described in Table 5.4.

**Utilitarian SWF (MEXSLP)**

In the context of survival functions, a utilitarian’s goal is to optimize the total survival probability. This is called the Maximal Expected Survival Location Problem (MEXSLP), and was first formulated in [38]. However, this formulation is not linear, and therefore does not scale well. Later, the same problem

\(^5\)In our implementation, we used \( \varepsilon = 10^{-5} \). For our case study with 217 nodes, this ensures that the objective value is approximated within \( 217 \cdot 10^{-5} \approx 10^{-3} \) of the true value.
The value of $l_i$ is bounded by linear functions (dashed lines), such that it is approximately equal to $\log(c_i)$.

![Figure 5.5](image)

**Figure 5.5** The value of $l_i$ is bounded by linear functions (dashed lines), such that it is approximately equal to $\log(c_i)$.

<table>
<thead>
<tr>
<th>variable</th>
<th>for</th>
<th>range</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_i$</td>
<td>$i \in V$</td>
<td>$[0,1]$</td>
<td>utility for a patient in zone $i$</td>
</tr>
<tr>
<td>$x_j$</td>
<td>$j \in W$</td>
<td>$1, \ldots,</td>
<td>A</td>
</tr>
<tr>
<td>$z_{ijk}$</td>
<td>$i \in V, j \in W, k \in 1, \ldots,</td>
<td>A</td>
<td>$</td>
</tr>
</tbody>
</table>

**Table 5.4** Interpretation of variables for the MILP formulation that maximizes survival probabilities.

A difference with coverage models is that it is no longer sensible to pre-compute the set of bases that can reach zone $i$ within the RTT. (In the coverage models this set was denoted $W_i$.) Instead, we need to keep track of the exact preference order of vehicles for each demand zone. Thereafter, we introduce decision variables $z$ (see Table 5.4). We next formulate the MEXSLP as a MILP.

Maximize \[ \sum_{i \in V} d_i u_i \quad (5.7) \]
subject to
\[ u_i \leq \sum_{k=1}^{|A|} (1 - q)^{k-1} \cdot p_{ji} \cdot z_{ijk}, \quad (5.8) \]
5.6. Computational results

\[ \sum_{j \in W} z_{ijk} = 1, \quad i \in V, k = 1, \ldots, |A|, \quad (5.9) \]

\[ \sum_{k=1}^{|A|} z_{ijk} = x_j, \quad i \in V, j \in W, \quad (5.10) \]

\[ \sum_{j \in W} x_j \leq |A|, \]

\[ x_j \in \mathbb{N}, \quad j \in W, \]

\[ z_{ijk} \in \{0, 1\}, \quad i \in V, j \in W, k = 1, \ldots, |A|, \]

\[ u_i \in [0, 1], \quad i \in V. \]

Constraint (5.9) ensures that only one vehicle can be the \( k^{th} \) favourite for a certain demand zone. Constraint (5.10) ensures that the number of vehicles at base \( j \) that are \( k^{th} \) favorite (for any \( k \)) is equal to the number of vehicles at base \( j \) in total. Equivalently, constraint (5.8) may be formulated with an equality sign. Note that, as before, variables \( u_i \) are not strictly necessary to implement this model, but we add them for ease of reading.

Note that if one chooses survival function \( f \) to be

\[ f(t) = \mathbb{1}[t \leq T], \]

then the result is the same as the result for the models in Section 5.5.1. Therefore, the coverage optimization models may be viewed as special cases of the survival optimization models. A special case for which a more efficient formulation exists.

**Bernoulli-Nash SWF**

The model is identical to the MEXSLP, except we replace objective (5.7) with

\[ \text{Maximize} \sum_{i \in V} d_i \log(u_i) \quad (5.11) \]

As in Section 5.5.1, we deal with this nonlinear problem by creating a piecewise linear upper bound on the logarithm of \( u_i \), for all \( i \in V \). The model can then be solved with a MILP solver (Gurobi).

### 5.6 Computational results

We continue with a case study for which we computed numerical results. This section reports results based on an EMS region in the Netherlands. We first introduce the region, after which we discuss and compare the solutions of the different optimization methods.
5.6.1 Region

We apply our models to the province of Utrecht, which was described in Section 2.7.

Utrecht is divided in 217 postal codes: these will be our demand nodes $V$. We will take the fraction of demand in a single node to be proportional to the number of inhabitants in that postal code. The driving times $\tau_{ij}$ for EMS vehicles between any two nodes $i, j \in V$ were estimated by the RIVM [66, Chapter 3]. The demand nodes are depicted in Figure 5.6.

For the purpose of this chapter, we want to place sixteen vehicles in the region. Recall that we want to optimize both which base locations to open, as well as how many vehicles to put at each base. Therefore, we want to consider more than just the nineteen existing base locations. However, using all 217 demand nodes as possible base locations might be quite a lot to handle, at least for our most complex models. To limit the computation time we choose a subset of 50 of these demand nodes to be our potential base locations. We want to make sure the set of potential base locations is well spread out over the region. Thereto, we formulate and solve a MIP: when two bases are opened within distance $t$ (in minutes) from each other, we incur a penalty $e^{-0.5t}$. This implies that when two bases are opened close to one another, the corresponding penalty is very high. The objective of the MIP is to minimize the sum of these penalties. We add constraints that ensure the 19 currently existing base locations are included in...
5.6. Computational results

<table>
<thead>
<tr>
<th>model</th>
<th>run time</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEXCLP</td>
<td>1 second</td>
</tr>
<tr>
<td>MaxFair\textsubscript{cov}</td>
<td>10 seconds</td>
</tr>
<tr>
<td>MEXSLP\textsubscript{survival}</td>
<td>9 minutes</td>
</tr>
<tr>
<td>MaxFair\textsubscript{survival}</td>
<td>2-5 hours</td>
</tr>
</tbody>
</table>

Table 5.5 Run times for different optimization models. These run times are measured for the region Utrecht with 50 bases and sixteen ambulances.

the solution. This gives the 50 locations as depicted in Figure 5.7.

Figure 5.7 The demand locations (all nodes) and the 50 locations that we will use as possible bases (black nodes), as determined by our MIP.

5.6.2 Run times

We implemented and solved the six different optimization models for the region Utrecht as described above. The solve times are reported in Table 5.5. One immediately sees that the computational effort varies highly depending on the model. As expected, the coverage optimization models are more efficient than the survival models. Furthermore, the MaxFairness models are much harder than the MEXCLP/MEXSLP models. Note that the computation time of the Max-Fairness models depend on $\varepsilon$. The tolerance of the difference between the linear approximation and the true value of the logarithm, as described in Section 5.5.1. The values reported in Table 5.5 are for $\varepsilon = 10^{-5}$.

5.6.3 Results

This section reports the results of the six different optimization models described in Section 5.5, applied to the region Utrecht defined in Section 5.6.1. We ran our
models for several busy fractions \((q)\), ranging from 0 to 0.9. We next show how the optimal solution for each model performs against all objectives, and highlight some of the differences between the solutions.

The Bernoulli-Nash social welfare may be computed as follows. If \(v\) represents the objective value of a MaxFairness model (e.g., the value of Equation (5.6)), then the Bernoulli-Nash social welfare is given by \(e^v\). However, note that this is not an exact answer because of errors in the approximation of the logarithm. Instead, we explicitly calculated the product of the utilities whenever we report values of Bernoulli-Nash social welfare.

![Table 5.6](image)

<table>
<thead>
<tr>
<th>model</th>
<th>(U_{cov})</th>
<th>BN(_{cov})</th>
<th>(U_{stoch})</th>
<th>BN(_{stoch})</th>
<th>(U_{surv})</th>
<th>BN(_{surv})</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEXCLP</td>
<td>0.614</td>
<td>0</td>
<td>0.610</td>
<td>0.538</td>
<td>0.119</td>
<td>0.090</td>
</tr>
<tr>
<td>MaxFairness(_{cov})</td>
<td>0.518</td>
<td>0.486</td>
<td>0.534</td>
<td>0.515</td>
<td>0.104</td>
<td>0.093</td>
</tr>
<tr>
<td>MEXSLP(_{stoch})</td>
<td>0.609</td>
<td>0</td>
<td>0.616</td>
<td>0.519</td>
<td>0.126</td>
<td>0.092</td>
</tr>
<tr>
<td>MaxFairness(_{stoch})</td>
<td>0.589</td>
<td>0</td>
<td>0.601</td>
<td>0.565</td>
<td>0.124</td>
<td>0.106</td>
</tr>
<tr>
<td>MEXSLP(_{surv})</td>
<td>0.5923</td>
<td>0</td>
<td>0.6028</td>
<td>0.4842</td>
<td>0.1305</td>
<td>0.0899</td>
</tr>
<tr>
<td>MaxFairness(_{surv})</td>
<td>0.5766</td>
<td>0</td>
<td>0.5925</td>
<td>0.5551</td>
<td>0.1224</td>
<td>0.1071</td>
</tr>
</tbody>
</table>

Table 5.6  Performance of the optimal solution for each model against all objectives. These values correspond to \(q = 0.75\). Note that the numbers on the diagonal correspond to the objective for that model, hence they are the maximum in each column.

The first thing to notice is that BN\(_{cov}\) is often zero. This means that in those solutions there is at least one postal code, however small, that cannot be reached within twelve minutes by any ambulance. Therefore the product of utilities is also zero.

![Figure 5.8](image)

Figure 5.8  The optimal solutions for the utilitarian solution and the MaxFairness solution, for \(q = 0.75\). The grey nodes are demand points, the black nodes are possible base locations. The numbers represent the number of ambulances placed at each base. The utility in both cases is deterministic coverage.

Let us compare the two solutions that use deterministic coverage as their utility (i.e., MEXCLP and MaxFairness\(_{cov}\)). We selected the solutions for \(q = 0.75\),
5.6. Computational results

because differences become more clear for higher values of $q$. As Figure 5.8 shows, the two solutions are quite different. What stands out is the fact that MEXCLP places multiple vehicles at the same base (up to five), which is due to the high busy fraction: a single vehicle does not provide a lot of coverage. Hence, giving densely populated areas additional vehicles is preferred over giving areas of less demand a first vehicle. In contrast, MaxFairness$_{cov}$ shows more of a tendency to spread out over the region. Note that this same effect could already be seen on a smaller scale in the illustrative example from Section 5.4.

Figure 5.9 The coverage versus demand of each node, using deterministic driving times.

Let us further analyze the two different solutions depicted in Figure 5.8. Thereto, we compare the utilities (coverage) of individual demand nodes: in Figure 5.9 we plot the coverage versus the demand of each node. This shows that the MEXCLP solution gives some of the highest coverages, but also some of the lowest (even zero). The values for the MaxFairness$_{cov}$ solution are closer to one another. This is consistent with what might intuitively be considered ‘fair’. Furthermore, note that the coverage only takes a few different values: this is due to the relatively simple definition of deterministic coverage.

Next, let us look at the stochastic counterpart of the previously described case. That is, instead of deterministic coverage, we take stochastic coverage as utility. The busy fraction remains 0.75. We again see that the Bernoulli-Nash optimum spreads the ambulances more than the utilitarian optimum (Figure 5.10). If we plot the stochastic coverage against the demand (Figure 5.11) we no longer see the clear discretization of values that we observed in the deterministic case.

For our next argument, imagine a utilitarian EMS manager positioning ambulances: that is, he would place them according to the MEXSLP solution. We investigate how much the fairness$^6$ of his solution can be improved by positioning the ambulances in a different way. This improvement is depicted in Figure 5.12 for different busy fractions. We show both the results for stochastic coverage and survival as utilities. Deterministic coverage is omitted, because the Bernoulli-Nash SWF of the utilitarian solution is often 0, and hence the ratio would become infinitely large. Figure 5.12 shows that for small busy fractions, the ratio is very close to 1, hence a utilitarian manager’s choices are actually quite fair. However,

---

$^6$ the value of the Bernoulli-Nash SWF
Chapter 5. Fairness in the ambulance location problem

Figure 5.10 The optimal solutions for the utilitarian solution and the MaxFairness solution, for $q = 0.75$. The grey nodes are demand points, the black nodes are possible base locations. The numbers represent the number of ambulances placed at each base. The utility is stochastic coverage.

Figure 5.11 The coverage versus demand of each node, using stochastic driving times.

for higher values of $q$, the improvement factor increases up to 1.4.

As Figure 5.12 shows, the improvement factor is larger for the survival function than for the stochastic coverage. This can be explained by the fact that the survival is more rapidly declining with distance (compare Figure 5.4 to Figure 5.3). This increases the gap between a solution that places many vehicles on one base (as we have seen in Figure 5.8), and a solution that tends to spread vehicles.

We further compare the differences between the utilitarian and Bernoulli-Nash solutions. For now, let us focus on the stochastic coverage, i.e., we compare the MaxFairness$_{stoch}$ solution to the MEXSLP$_{stoch}$ solution. We investigate how each model performs under its own objective, as well as under the objective of the other model. Figure 5.13a shows these values for several values of $q$. First of all, note how surprisingly similar the values for BN$_{stoch}$ and U$_{stoch}$ are for low values of $q$. For higher values of $q$, we see that gap in the BN$_{stoch}$ is slightly bigger than the gap in U$_{stoch}$. This means that if one optimizes for fairness (maximize BN$_{stoch}$), the loss in coverage (U$_{stoch}$) is less than it would be vice versa. Furthermore, as before, we see that the gap between solutions widens as
Figure 5.12 The relative improvement in the Bernoulli-Nash SWF, comparing the optimum to the utilitarian solution.
Chapter 5. Fairness in the ambulance location problem

Figure 5.13 Comparing the objectives $\text{BN}_{stoch}$ and $\text{U}_{stoch}$, for both optimization models that use stochastic coverage as utility. Dashed lines represent the MaxFairness$_{stoch}$ solution. Solid lines represent the MEXSLP$_{stoch}$ solution.
5.7. Discussion

the load of the system increases.

A quick conclusion might be that a high system load (large \( q \)) directly causes the gap between solutions that maximize utilitarian and Bernoulli-Nash SWFs. However, recall that throughout these computations we varied \( q \) and kept the number of ambulances equal to sixteen. This implies that when we increase \( q \), fewer people can be served. This is not necessarily the case for all problem instances with a large \( q \): one can imagine a very busy EMS region, where vehicles are almost always busy, yet there are so many ambulances that the overall performance is still very high.

To analyze what truly causes the differences between the MaxFairness\(_{\text{stoch}}\) and the MEXSLP\(_{\text{stoch}}\) solution - the value of \( q \), or the total number of people that can be served - we performed additional computations. We fixed \( q \) at a value for which Figure 5.13a showed differences between the MaxFairness\(_{\text{stoch}}\) and the MEXSLP\(_{\text{stoch}}\) solution (we choose \( q = 0.75 \)) and increased the number of ambulances. In Figure 5.13a we observe the range for which the MaxFairness\(_{\text{stoch}}\) and the MEXSLP\(_{\text{stoch}}\) solution appear very similar: roughly where \( U_{\text{stoch}} \approx 0.8 \). Therefore, we start with 16 ambulances and increase this number until \( U_{\text{stoch}} \approx 0.8 \). The results are depicted in Figure 5.13b. As before, this shows that the social welfare for both solutions are remarkably similar when \( U_{\text{stoch}} \approx 0.8 \). This indicates that not the load of the system \((q)\), but the utilitarian social welfare \((U_{\text{stoch}})\) determines in which cases the MaxFairness and the MEXSLP differ.

5.7 Discussion

This chapter approaches ambulance location problems from the perspective of social welfare. Our main contribution is that we introduced and implemented two models that maximize the Bernoulli-Nash social welfare. These solutions are juxtaposed against well-known utilitarian solutions. We showed numerical results for a realistic EMS region in the Netherlands. The differences between the utilitarian and Bernoulli-Nash optima turned out to depend on the total number of people that can be served. When at least 80% of the population can be served within the RTT, we found that both solutions are remarkably similar. This is a somewhat surprising, but reassuring result. For a coverage smaller than 80%, the utilitarian and Bernoulli-Nash optima start to show their differences. Generally speaking the Bernoulli-Nash optimum tends to spread vehicles throughout the region, whereas the utilitarian optimum clusters vehicles in areas with high demand. We conclude that as the total coverage of the system decreases, it becomes more important to explicitly think about fairness.

When we translate our result to implications for ambulance providers in practice, the outcome depends on the EMS region. For example, an ambulance provider in the Netherlands typically serves 95% of the most urgent requests within 15 minutes [89]. Our results indicate that, even if they aimed for a utilitarian optimum without thinking about fairness, their solution will be rather fair. For other EMS providers - e.g. in the UK where ambulances typically reach
75% of their most urgent requests within eight minutes [90] - there will be more differences between the most efficient and the most fair solution. In such cases, the political debate about fairness in ambulance care deserves more attention.

The trade-off between efficiency and fairness remains unavoidable. While classical ambulance equity models - that use an egalitarian approach - are not suitable to use in practice because of their poor overall performance, we believe the Bernoulli-Nash optimum provides a reasonable alternative to the utilitarian optimum. This means that, besides the theoretical importance of our work, our solutions could truly be worth considering for ambulance providers who believe fairness should play a role in their decisions. Additionally, the Bernoulli-Nash social welfare of current practice could be evaluated, and used as a measure to detect how far current solutions are from a maximally fair solution.

The models in Section 5.5.2 implicitly define the situation in which there are no ambulances available to have survival probability zero. Alternatively, one might extend these models with a nonzero survival probability in case no ambulance is available. This probability may depend on the demand node \( i \), therefore denote it \( \delta_i \). This affects the constraints as follows: one should add \( + q_i |A| \delta_i \) to the right-hand side of (5.8).

In [17] the authors add a decision variable and constraint to limit the total number of base locations used. We chose to leave such a constraint out of this chapter, because it might distract the reader from the main topic. However, it could be added without further complicating the models.

As already described in Section 5.3, using a busy fraction to model ambulance availability is an approximation of the true system dynamics. One might suggest to relax the assumptions of independent vehicles all having the same busy probability, by using a more advanced model. This is done in Hypercube Queuing Models (HQM) [69] (which are compared to MEXCLP in [9]). However, considering that our most complex model already takes five hours to solve, we chose not to make the problem harder. Other researchers that optimized a form of fairness while using hypercube correction factors faced computational difficulties - even for small cases - and needed to resort to tabu search [31], hence losing a guarantee of finding a globally optimal solution.