Contents lists available at ScienceDirect

Operations Research for Health Care

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

Linear formulation for the Maximum Expected Coverage Location Model with fractional coverage

P.L. van den Berg [a,b,*], G.J. Kommer [c], B. Zuzáková [c,d]

[a] Delft Institute of Applied Mathematics, Delft Institute of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
[b] Centrum Wiskunde & Informatica, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
[c] VU University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
[d] Charles University, Prague, Czech Republic

ARTICLE INFO

Article history:
Received 23 September 2014
Accepted 4 August 2015
Available online xxxx

Keywords:
Integer linear programming models
Ambulance base locations
Location analysis

ABSTRACT

Since ambulance providers are responsible for life-saving medical care at the scene in emergency situations and since response times are important in these situations, it is crucial that ambulances are located in such a way that good coverage is provided throughout the region. Most models that are developed to determine good base locations assume strict 0–1 coverage given a fixed base location and demand point. However, multiple applications require fractional coverage. Examples include stochastic, instead of fixed, response times and survival probabilities. Straightforward adaption of the well-studied MEXCLP to allow for coverage probabilities results in a non-linear formulation in integer variables, limiting the size of instances that can be solved by the model. In this paper, we present a linear integer programming formulation for the problem. We show that the computation time of the linear formulation is significantly shorter than that for the non-linear formulation. As a consequence, we are able to solve larger instances. Finally, we will apply the model, in the setting of stochastic response times, to the region of Amsterdam, the Netherlands.

1. Introduction

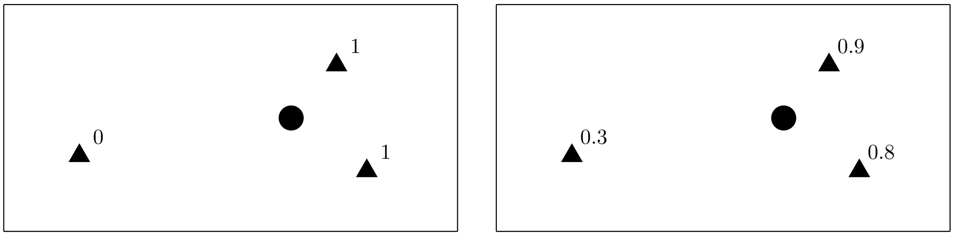
Ambulances play an important role in providing life-saving health care in case of an emergency. Since time is limited in emergency situations, it is critical that ambulance vehicles are located so as to ensure good coverage and short response times. The response time is defined as the time between the moment the call is taken in the call center and the moment the ambulance arrives at the scene of the incident. In most countries, regulations state that a minimum fraction of calls should be reached within a specified target response time. In the Netherlands, for example, an ambulance should arrive at the patient within 15 min after the call is made in 95% of the cases.

Many models have been proposed to determine good locations for ambulances. Among the first models were the Location Set Covering Model (LSCM) and the Maximal Covering Location Problem (MCLP). LSCM, introduced by Toregas [1], computes the minimum number of ambulances required to cover all demand points within the target response time. Church and ReVelle [2] developed MCLP for the case that the available capacity does not suffice to cover all demand points. The model maximizes the demand that can be covered, given the limited resources. Even though these models are useful in many applications, two strong assumptions are made.

First, the models assume that ambulances are always available for dispatch, neglecting the fact that an ambulance might be dispatched to another call. This observation resulted in numerous models that consider backup coverage. Examples that require a fixed number of ambulances to provide full coverage are Double Standard Model (DSM) [3], Backup Coverage Problem (BACOP) [4] and Maximum Availability Location Problem (MALP) [5]. The first two require two ambulances to cover a demand point, while in the last one this number depends on the average fraction of time an ambulance is unavailable, called the busy fraction. A slightly more realistic way of modeling ambulance availability uses the concept of expected coverage, introduced by Daskin [6]. Here, the busy fraction is used to estimate the probability of having at least one available ambulance within the target response time.

[*] Corresponding author at: Delft Institute of Applied Mathematics, Delft Institute of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands.
E-mail addresses: p.l.vandenberg@tudelft.nl (P.L. van den Berg), g.j.kommer@vu.nl (G.J. Kommer).

http://dx.doi.org/10.1016/j.orhc.2015.08.001
2211-6923/© 2015 Elsevier Ltd. All rights reserved.



(a) Deterministic case.

(b) Stochastic case.

Fig. 1. Representation of first example of difference between deterministic and stochastic case. A circle represents a demand point, a triangle represents a base station. The numbers next to the triangles show the probability that the demand point can be reached within the time limit from the particular base.

Repede and Bernardo [7], Schmid and Doerner [8], and Van den Berg and Aardal [9] extend DSM and MEXCLP to incorporate variation in system characteristics throughout the day. Another approach to account for ambulance availability is to use Stochastic Programming techniques, see for example [10,11].

A second assumption in LSCM and MCLP is that the obtained coverage by assigning a call to an available ambulance at a particular base is either 0 or 1. In some applications, it would be useful to relax this assumption and allow for fractional coverage. We discuss two well-studied examples: coverage probabilities and survival probabilities.

In most models, it is assumed that the response time from a particular base to a particular demand point is fixed. Typically, this is equal to the average travel time plus some fixed pre-trip delay, where the pre-trip delay is the time elapsed before the ambulance starts driving. In practice, however, these response times vary, due to traffic jams and weather conditions. At least two ways of handling this uncertainty are used in literature. Koç and Bostancioğlu [12] introduce a required reliability α, and say that a base can cover a demand point only if the response time is within the time threshold with probability at least α. This way, they have again a 0 or 1 coverage and then they apply DSM. Another, more common approach is to compute the coverage probability directly. This approach is applied to multiple models described earlier. Both Daskin [13], and Karasakal [14] included the coverage probability in MCLP. Marianov and ReVelle [15] adapted MALP so that it could handle coverage probabilities. A version for MEXCLP with coverage probabilities was introduced by Goldberg et al. [16] and Goldberg and Paz [17], for which they used heuristics to find approximate solutions. Based on these two papers, Ingolfsson et al. [18] developed a non-linear variant of MEXCLP with coverage probabilities. For small instances, typically with a fixed set of bases, the model could be solved to optimality. However, the computation time increases rapidly when the instance size increases. To determine the coverage probabilities, they assumed that both the pre-trip delay and the travel times are non-deterministic.

A second example of the usage of fractional coverage is the concept of survival probabilities. Erkut et al. [19] argue that even though most EMS providers are assessed on coverage related criteria, it is worth to consider performance measures related to health outcomes. They introduce a version of MEXCLP that maximizes the survival probability of a patient rather than the expected coverage. By replacing the coverage probability by the survival probabilities, we get that this model is equivalent to [18]. Again, the presented model is a non-linear integer programming problem. Knight et al. [20] extend the model to allow for different survival probabilities for different types of patients. Later, Mayorga et al. [21] used survival probabilities to develop dispatch policies.

Our main contribution is that we present an integer linear programming formulation for the version of MEXCLP with fractional coverage. Compared to the non-linear formulation [18,19], this reduces the computation time and allows solving larger instances. Erkut et al. [22] solve the non-linear model for 180 demand points and 16 bases, but note that finding optimal solution for instances with more bases would be problematic. To apply the model to determine optimal base locations rather than an optimal distribution of the ambulances given a fixed set of bases, we need to solve instances with more base locations. We will show that our linear model can be solved for larger instances. Note that the two models are equivalent and thus provide the same solutions.

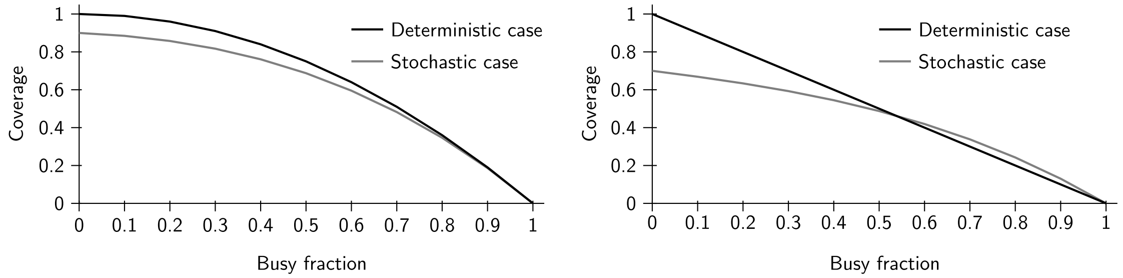
In Section 2, we will first show why a straightforward formulation will result in a non-linear model. Second, we will show how the problem can be reformulated as an integer linear programming problem. Finally, we will prove the equivalence of the two models. Section 3 provides an empirical comparison of the computation times of the linear and non-linear formulation. In Section 4, we apply the model to the region of Amsterdam to show the behavior of the model. Conclusions and possible extensions of this research are discussed in Section 5. Note that in the description of the model, we use the stochastic response times as our underlying application, but for the application to survival probabilities, the model is equivalent.

2. Model description

Even though ambulance location models typically use all-or-nothing coverage, multiple authors have noted that it might be more realistic to use fractional coverage probabilities. In this section, we present an adapted version of MEXCLP where a coverage w_{ij} is obtained when an available ambulance at base i responds to a call at demand point j. Different from the classical MEXCLP this probability does not have to be 0–1 valued. This w_{ij} can, for example, be interpreted as the probability of reaching demand point j within the time threshold from base i, or as the probability that a patient at location j survives when served by an ambulance from base i.

As in MEXCLP, we assume that each ambulance is unavailable a fraction q of the time. We call this the busy fraction. Furthermore, we assume that the availability of an ambulance is independent of the availability of the other ambulances. The probability that at least one ambulance out of k is available is then $E_k = 1 - q^k$. The expected coverage of a demand point covered by k ambulances is thus E_k. In our model, we will use this concept, introduced by Daskin [6], to determine the expected coverage.

We now give two examples to show the effect of fractional coverage probabilities on the expected coverage. In both examples, we consider a region with one demand point and three base locations, each with one ambulance located. Each base has a probability w_{ij} of covering the demand point. In the first example, these are 0.9, 0.8 and 0.3, respectively. In the second example, we have 0.7, 0.4, and 0.3. In the deterministic case, the coverage is 1 if $w_{ij} \geq 0.5$ and 0 otherwise. Fig. 1 depicts Example 1. Fig. 2 shows the expected coverage for the deterministic and stochastic case for both examples, varying the busy fraction. To show how the expected coverage is computed, we show the computation for Example 1 with a busy



(a) Example 1.

(b) Example 2.

Fig. 2. Difference between deterministic and stochastic case for different parameters. In the first example, we have three bases with coverage probabilities 0.9, 0.8, and 0.3, respectively. In the second example, we have three bases with coverage probabilities 0.7, 0.4, and 0.3. In the deterministic case, we have a coverage probability of 1 if the stochastic coverage probability is at least 0.5, and 0 otherwise. The figures show the expected coverage for different busy fractions.

fraction of 0.4. In the deterministic case, we get

$$\begin{aligned}&\mathbb{P}(\text{1st available}) * \mathbb{P}(\text{1st in time}) + \mathbb{P}(\text{1st unavailable})\\ &\quad * \mathbb{P}(\text{2nd available}) * \mathbb{P}(\text{2nd in time})\\ &\quad + \mathbb{P}(\text{1st and 2nd unavailable}) * \mathbb{P}(\text{3th available})\\ &\quad * \mathbb{P}(\text{3th in time})\\ &= 0.6 * 1 + 0.4 * 0.6 * 1 + 0.4^2 * 0.6 * 0 = 0.84.\end{aligned}$$

For the stochastic case we get

$$\begin{aligned}&\mathbb{P}(\text{1st available}) * \mathbb{P}(\text{1st in time}) + \mathbb{P}(\text{1st unavailable})\\ &\quad * \mathbb{P}(\text{2nd available}) * \mathbb{P}(\text{2nd in time})\\ &\quad + \mathbb{P}(\text{1st and 2nd unavailable}) * \mathbb{P}(\text{3th available})\\ &\quad * \mathbb{P}(\text{3th in time})\\ &= 0.6 * 0.9 + 0.4 * 0.6 * 0.8 + 0.4^2 * 0.6 * 0.3 \approx 0.76.\end{aligned}$$

Fig. 2 shows that using 0–1 coverage results in different estimations of the expected coverage than using the fractional coverage. Typically, the deterministic case overestimates the expected coverage, even though Example 2 shows that for high busy fractions, it can also be the other way around. These examples stress the importance of including fractional coverage probabilities.

2.1. *Model formulation*

In our model, we are given a set of demand points N and a set of possible base locations M. For each demand point j, we have a given demand d_j. This d_j should be a measure for the number of calls within demand point j. See, for example, Channouf et al. [23], and Setzler et al. [24] for EMS call volume forecasting methods. Each base location has a capacity b_i, which is the maximum number of ambulances that may be located at that station. In total we are allowed to use at most β base locations. The total number of available ambulances is b. The busy fraction of an ambulance is denoted by q. For each combination of a demand point j and a base location i, we have a probability w_{ij} that an ambulance departing from base i will reach demand point j within the time threshold. For fixed demand point j, given w_{ij}, we can order the base locations from the closest to the furthest for this demand point. Let a_{ij} denote the index of the base location that is in position i in this ordering for demand point j. Similarly, let $ranking(i, j)$ be the ranking of base i in the ordering of demand point j. So, by definition we have $ranking(a_{ij}, j) = i$.

The most straightforward way of modeling our problem is to introduce a decision variable x_i denoting the number of ambulances assigned to location i. The expected coverage of demand point j in terms of x_i is then

$$c_j(x) = \sum_{i \in M} q^{\sum_{k<ranking(i,j)} x_{a_{kj}}} (1 - q^{x_{a_{ij}}}) w_{a_{ij}j}. \tag{1}$$

Clearly, this formulation is not linear in the decision variables. When solving larger instances, this can result in longer computation times. To avoid this, we present a different formulation for which the objective is linear in the decision variables.

In order to formulate a linear model, we introduce a new binary decision variable z_{ijk} indicating whether the kth preferred, with respect to w_{ij}, ambulance for demand point j is located at base location i. If, for example, base location 1 is the closest one for demand point 2 and we have three ambulances located at that base location, we get $z_{121} = z_{122} = z_{123} = 1$. Additionally, we introduce a binary variable y_i, which has value 1 if and only if at least one ambulance is located at base location i. This variable is needed to limit the number of base locations that is used.

Using these decision variables we are able to formulate our model as follows:

$$\max c(z) = \sum_{j \in N} d_j c_j(z) \tag{2}$$

with

$$\sum_{k=1}^{b} z_{ijk} \le x_i \quad \forall i \in M,\ j \in N, \tag{3}$$

$$\sum_{i \in M} z_{ijk} = 1 \quad \forall j \in N,\ k \le b, \tag{4}$$

$$\sum_{i \in M} y_i \le \beta, \tag{5}$$

$$x_i \le b_i y_i \quad \forall i \in M, \tag{6}$$

$$\sum_{i \in M} x_i = b, \tag{7}$$

$$y_i, z_{ijk} \in \{0, 1\} \quad \forall i \in M,\ j \in N,\ k \le b, \tag{8}$$

$$x_i \in \mathbb{N} \quad \forall i \in M \tag{9}$$

and

$$c_j(z) = \sum_{k=1}^{b} (1 - q) q^{k-1} \sum_{i \in M} z_{ijk} w_{ij} \quad \forall j \in N.$$

The objective is to maximize the expected coverage over all demand points. This is defined as the sum of the coverages that can be provided to an individual node by the whole system, $c_j(z)$, multiplied by the total demand generated at this node, d_j. The value $c_j(z)$ is calculated by conditioning on the number of unavailable ambulances. The probability that the kth ambulance is the first available one equals $(1 - q)q^{k-1}$. If the kth preferred ambulance is located at location i, we obtain an expected coverage of w_{ij}. Constraints (3) state that no more than x_i ambulances may be assigned to base i. This makes sure that the z_{ijk}'s have the right value. Constraints (4) ensure that the kth preferred ambulance of demand point j is located at no more than one base location.

In order to design a realistic system, we add a limitation on the maximum number of base locations by constraint (5). This constraint is not included in Ingolfsson et al. [18]. They assume that the set of bases is fixed. Constraints (6) guarantee that the number of vehicles located at each station does not exceed its capacity. Finally, constraint (7) states that no more than b ambulances are used.

In the Appendix the complete description of the non-linear version is given. Now, we prove that the two formulations are equivalent.

Theorem 1. $C^{MINLP} = C^{MILP}$.

Proof. Given a solution (x', y') for MINLP, we construct the following solution (x, y, z) for MILP. Let $x := x'$, $y := y'$, and

$$z_{ijk} := \begin{cases} 1 & \text{for } k = \left\{ \sum_{l<ranking(i,j)} x_{a_{lj}} + 1, \ldots, \sum_{l\le ranking(i,j)} x_{a_{lj}} \right\}, \\ & i \in M, \text{ and } j \in N \\ 0 & \text{otherwise.} \end{cases}$$

We show that this solution is feasible and that it has the same objective value as (x', y').

Since constraints (5)–(7) are equivalent to (A.10)–(A.12), the constructed solution satisfies these constraints. In the construction of z, we set exactly $\sum_{l<ranking(i,j)} x_{a_{lj}} - \sum_{l\le ranking(i,j)} x_{a_{lj}} = x_i$ variables to 1, given i and j. Hence, constraint (3) is satisfied. Finally, constraint (4) is satisfied because the order a_{ij} fully determines at which base the kth ambulance for demand point j is located.

Now, we define
$c_{ij}(x) := q^{\delta}(1-q^{\lambda})w_{a_{ij}j}$, where $\delta = \sum_{l<ranking(i,j)} x_{a_{lj}}$ and $\lambda = x_{a_{ij}}$.
Then, we get

$$\begin{aligned} c_{ij}(x) &= q^{\delta}(1-q^{\lambda})w_{a_{ij}j} = q^{\delta}\sum_{k=1}^{\lambda} q^{k-1}(1-q)w_{a_{ij}j} \\ &= \sum_{k=1}^{\lambda} q^{\delta+k-1}(1-q)w_{a_{ij}j} = \sum_{k=\delta+1}^{\lambda+\delta} q^{k-1}(1-q)w_{a_{ij}j} \\ &= \sum_{k=1}^{b} q^{k-1}(1-q)w_{a_{ij}j}z_{a_{ij}jk}. \end{aligned}$$

For the first equality, we use the geometric sequence. This gives that $1 - q^{\lambda} = \sum_{k=1}^{\lambda} q^{k-1}(1-q)$. The last equality is true by construction of z. All terms that are added to the sum have $z_{ijk} = 0$.

By summing over all demand points j and base stations i, we get

$$\begin{aligned} \sum_{j\in N} d_j c_j(x) &= \sum_{j\in N} d_j \sum_{i\in M} c_{ij}(x) \\ &= \sum_{j\in N} d_j \sum_{i\in M}\sum_{k=1}^{b} q^{k-1}(1-q)w_{a_{ij}j}z_{a_{ij}jk} \\ &= \sum_{j\in N} d_j \sum_{i\in M}\sum_{k=1}^{b} q^{k-1}(1-q)w_{ij}z_{ijk} \\ &= \sum_{j\in N} d_j c_j(z). \end{aligned}$$

Since for every solution (x', y') for MINLP we can find a solution (x, y, z) for MILP with the same objective value, we have that $C^{MINLP} \le C^{MILP}$.

To show that $C^{MINLP} \ge C^{MILP}$, we prove that given an optimal solution (x^*, y^*, z^*) for MILP, we have that (x^*, y^*) is a feasible solution for MINLP with the same objective value. Clearly, (x^*, y^*) is feasible. Without loss of generality, we can assume that the optimal solution for MILP satisfies the relation between x and z as before. It is optimal to respect the order a_{ij} in filling z, because $q^{k-1}(1-q)$ is concave. As a result, by the same arguments as before, we have that $C^{MINLP} \ge C^{MILP}$.

Hence, $C^{MINLP} = C^{MILP}$. □

3. Comparison of computation time

To analyze the difference in computation time between our formulation (MILP) and the non-linear formulation (MINLP), used for example by Ingolfsson et al. [18], we apply both models to a set of 20 generated test instances. We implemented both models in AIMMS 3.14 [25] and used the default solvers, which are CPLEX 12.6 [26] and BARON 12 [27], respectively.

We created two sets of ten instances differing in the number of demand points and potential bases. We randomly generated demand points and base locations in the unit square. The average travel time between two points is the Euclidean distance multiplied by 1500 s. We assume that both the pre-trip delay and the travel times are stochastic. The travel times are assumed to be normally distributed with a coefficient of variation of 0.25, which corresponds to a standard deviation of 25% of the mean. The pre-trip delays are incorporated in the same way as in the case study, a lognormal distribution with mean 5.2967 and standard deviation 0.4574. The time threshold is set to 900 s, or 15 min. For each demand point, we generate a weight d_j uniformly between 10 and 30. Hence, the maximum difference in importance between two demand points is a factor of 3. We use a busy fraction of 42%, which corresponds to the observed busy fraction in the region of Amsterdam.

For the first set of instances, we take 180 demand points and 10 potential bases. We set β to 10, so that there is no limitation on the number of opened bases. Basically, the model only decides how to distribute the available ambulances over the given bases. The dimensions of these instances correspond to the test cases in Ingolfsson et al. [18].

The second set of instances is used to test how the formulations perform when not all bases can be opened. To that end, we take instances with 100 demand points and 100 base locations, while allowing to open only 10 bases, i.e. $\beta = 10$. In both sets, we set the number of ambulances to 18 and the maximum number of ambulances per base to 5.

Note that the two formulations, MILP and MINLP, are equivalent and thus have the same optimal objective value.

3.1. Results

As described above, we have a total of 20 test instances, which can be divided in two groups. To all instances, we apply both models with different time limits. For the easier set of instances, 10–180, we set the time limit to 5 min, 30 min, and 24 h. Since the linear formulation already provided all optimal solutions within 5 min, we did not run it with longer time limits. For the second set of instances, we used time limits of 30 min and 24 h. The results for all instances are summarized in Table 1.

The table shows a significant difference in performance between the two formulations for both set of instances. For the first set, MILP was able to solve all instances to optimality in less than 3 s, while for MINLP, optimality could not be guaranteed for any of the instances within 30 min. However, in seven cases the best solution found after 30 min was the optimal one. For one instance, the solver did not provide a feasible solution within 5 min. Even within a time limit of 24 h, optimality could not be guaranteed for two of the instances, although the optimal solution was found.

For the second instance set, no optimal solutions were found within 30 min. For MILP, the average gap was only 0.10%, while for MINLP this was 28.25%. Note that the gap is defined as the value

Please cite this article in press as: P.L. van den Berg, et al., Linear formulation for the Maximum Expected Coverage Location Model with fractional coverage, Operations Research for Health Care (2015), http://dx.doi.org/10.1016/j.orhc.2015.08.001

Table 1
Results for comparison of computation time. The first three columns describe the instances. Column 4 shows the number of instances that are solved to optimality. In column 5, it is stated in how many of those cases the optimality could be verified by the solver. Column 6 shows the number of instances for which no solution is returned by the solver after the time limit has exceeded. The final two columns show the average gap and the average computation time.

Size	Formulation	Time limit (s)	# opt	# verified	# no sol.	Aver. gap	Aver. time (s)
10–180	MINLP	300	2	0	1	0.50%	300
10–180	MINLP	1 800	7	0	0	0.02%	1800
10–180	MINLP	86 400	10	8	0	0.00%	39 303
10–180	MILP	300	10	10	0	0.00%	1.96
100–100	MINLP	1 800	0	0	3	28.25%	1800
100–100	MINLP	86 400	0	0	2	26.40%	86 400
100–100	MILP	1 800	0	0	0	0.06%	1800
100–100	MILP	86 400	9	9	0	0.00%	31 308

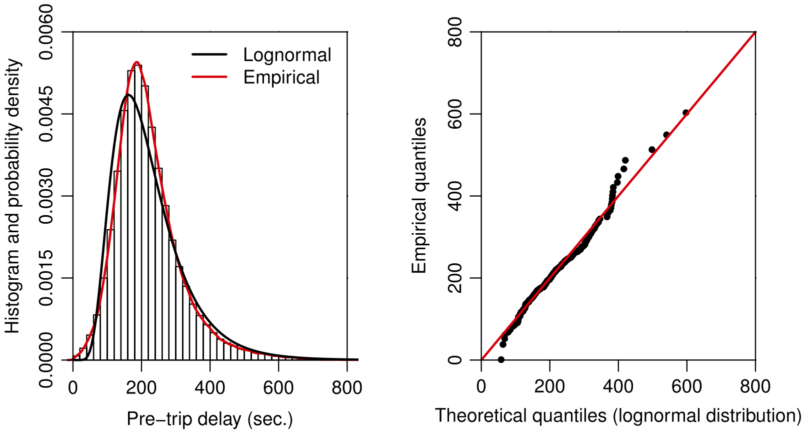


Fig. 3. Empirical distribution of pre-trip delay.

of the best found solution divided by the best found upper bound. The upper bound found in the linear formulation is also used to compute the gap for the non-linear case. When the time limit is set to 24 h, MILP was able to solve nine instances to optimality. For the remaining instance, the gap was only 0.02%. The non-linear model gave no optimal solutions and the average gap was 26.40%. In two cases, no solution was returned by the solver.

4. Case study

In this section, we apply the presented model to the region of Amsterdam, the Netherlands. We define the set N of demand points as the set of all postal codes in this region. This gives us a total of 161 points, which corresponds with an average size of 3.9 km^2 per postal code zone. We assume that each demand point is also available as a potential base location. Hence, the set M equals the set N. However, in the solution, we are allowed to use at most 9 of these bases, which corresponds to the number of bases currently in use in this region. The number of available ambulances is set to 18.

4.1. Data analysis

In order to apply the model, we have to determine the busy fraction q, the demand d_j for each $j \in N$, and the coverage probabilities w_{ij}. For the busy fraction, we take the average busy fraction during the day over the last 4 years, which is equal to 0.42. The expected demand for demand point j, d_j, is estimated by the average number of calls that have arisen from that demand point over the years 2008–2012. This data is provided by the ambulance provider in the region of Amsterdam. To compute w_{ij}, we have to estimate the pre-trip delay distribution and the travel time distribution. Below, we describe these estimations. Based on these two distributions, we compute w_{ij} by taking the convolution of the two distributions. Let R_{ij} be a random variable representing the response time for a call from demand point j served by base i. Furthermore, let $t_{ij}(x)$ be the travel time distribution for trips between i and j. Finally, let $h(x)$ be the distribution function of the pre-trip delay. Note that the pre-trip delay is independent of i and j. As in [18], we compute w_{ij} in the following way:

$$w_{ij} = \mathbb{P}(R_{ij} \leq \delta) = \int_0^{\delta} h(x) t_{ij}(\delta - x) dx. \qquad (10)$$

Here, δ is the response time target, which is 15 min in the Netherlands.

4.1.1. Pre-trip delays

The pre-trip delay is the time spent before the ambulance leaves the station. Based on 446,290 calls of high urgency, we find that a lognormal distribution gives a reasonable fit. This is the same result as obtained by Ingolfsson et al. [18]. For our data, the pre-trip delay is best approximated by a lognormal distribution with mean 5.2967 and standard deviation 0.4574. This corresponds to an average pre-trip delay of 222 s and a standard deviation of 107 s. The average is similar to the numbers reported in the annual EMS-reports in the Netherlands [28]. Fig. 3 shows the empirical and fitted distribution of the pre-trip delays.

4.1.2. Travel times

The calculation of the travel time distribution is more complicated, since we have a different travel time for each pair of base location and demand point. In order to estimate these distributions, we analyzed 10 pairs with more than 750 samples in our database. Based on these, we conclude that the travel times are well approximated with a normal distribution with a coefficient of variation of 0.25. One problem with a normal distribution is that it could generate negative values, which cannot occur in practice. However, since the coefficient of variation of 0.25, this happens only for values smaller than $\mu - 4\sigma$, which happens in only 0.3% of the cases. For the mean travel time between two points, we use

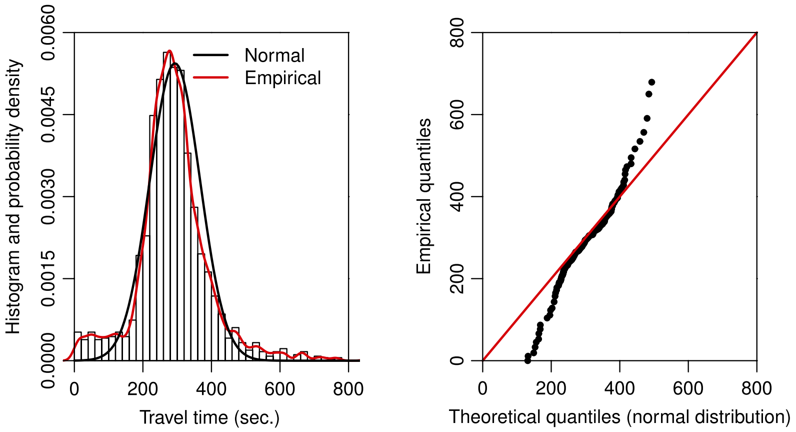


Fig. 4. Empirical distribution of travel times for one particular pair of postal codes.

the travel time model introduced by Kommer and Zwakhals [29], which is specifically developed for ambulances in the Netherlands. This model estimates the driving speed on each road type and uses that to compute the travel times. The estimated driving speeds that we use are based on rush hours, workdays from 06.30 till 09.30 and from 15.00 till 19.00. Fig. 4 shows for one pair of points that this can give a reasonable fit, although we can also see that the fit is worse than for the pre-trip delay. To account for the potential misfit, we will evaluate the sensitivity of the model with respect to the travel time distribution in Section 4.2.4.

4.2. Results

To evaluate the performance of the model, we perform multiple tests. First, we will compare the optimal solution according to the model with the current set of base locations. This shows us the potential performance increase. Second, we will compare the solution with the cases where we do not take into account randomness in either the pre-trip delay or the travel times. This provides insight into the importance of modeling the uncertainty. Furthermore, by plotting the selected bases, we get insight in the structure of the provided solutions. Third, we investigate the impact of the restriction on the number of bases. In our base case, we limit the number of bases by the current number, which is nine. These results may provide a trade-off between the number of bases to open and the coverage that can be obtained. Finally, we will evaluate the sensitivity to the chosen travel time distribution. We show how the solutions change, when different coefficients of variation are used. All computations were executed on a 2.9 GHz Intel(R) Core(TM) i7-3520M laptop with 8 GB of RAM. We used CPLEX 12.5 as our solver [26].

4.2.1. Current versus optimal

In the current situation, there are nine base locations in the region of Amsterdam. To investigate whether these nine bases are located in an optimal way, we compare the optimal solution according to the model with the best solution given that the bases are fixed. Note that the number of ambulances remains fixed at 18. In the optimal solution, we are only allowed to open the same number of bases as in the current situation. We will refer to this case as the base case. Comparing the results, we see that without changing the base stations, we can obtain a coverage of 92.03%. By changing the bases, however, we can obtain an increase of 2.92 percentage points, to 94.95%. This corresponds with reaching 37% of the previously uncovered calls. Note that the actual coverage in 2012 was 93.3% for this region [28]. This coverage is higher than expected by the model, which can be explained by some of the simplifying assumptions of the model. The model ignores that dynamic ambulance management is used to improve the real-time performance. Additionally, in practice there is a link with non-urgent patient transportations that are partly executed with the same ambulances. We did not incorporate this link in the case study.

Table 2
Importance of taking into account randomness in pre-trip delay and travel times. Estimated coverage is the coverage with respect to the w_{ij}'s used in the optimization. Real coverage is the coverage with respect to w_{ij} where both pre-trip delay and travel times are stochastic.

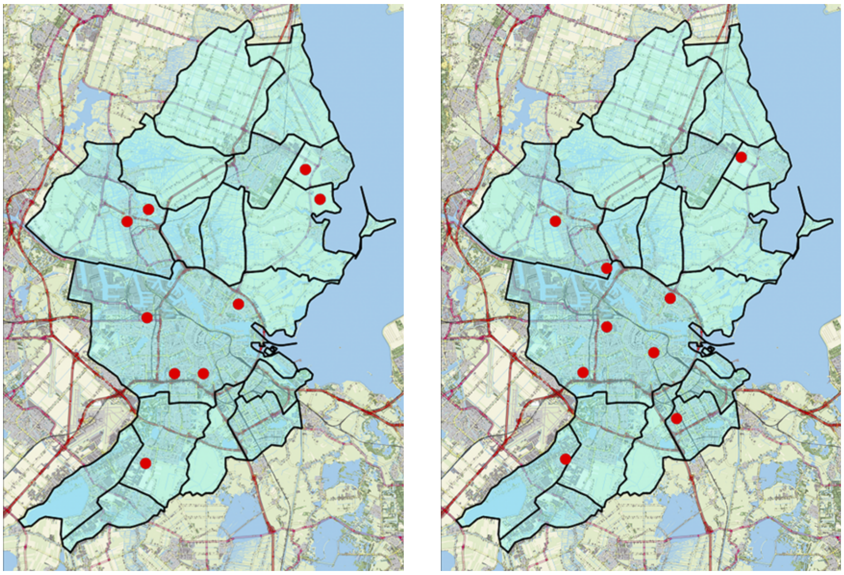
Pre-trip delay	Travel times	Estimated coverage	Real coverage
Deterministic	Deterministic	0.9852	0.9304
Deterministic	Stochastic	0.9656	0.9487
Stochastic	Deterministic	0.9623	0.9490
Stochastic	Stochastic	–	0.9495

4.2.2. Impact of randomness

To investigate the impact of the randomness in both the pre-trip delay and the travel times, we create four test instances. The first assumes stochastic pre-trip delays and travel times and is the same as the base case defined earlier. Then, we define two instances in which the randomness of one of the two response time components is ignored. The last instance has both deterministic delays and travel times and corresponds to the classical MEXCLP. In Table 2, we show the coverage according to the w_{ij}'s used in the optimization and the coverage according to the w_{ij}'s in the base case. Clearly, the coverage in the base case is the highest, since this gives the optimal solution with respect to the w_{ij}'s in the base case.

We observe that in order to get the optimal solution, it is important to take the randomness in both the components into account. In particular, when both random components are ignored, we obtain far from optimal solutions with respect to the input of the stochastic case. Note that this case corresponds to the classical MEXCLP. Furthermore, we see that the coverage is consistently overestimated when the randomness is not incorporated. In the fully deterministic case, this overestimation is almost 5.5 percentage point.

Since the non-linear formulation was not able to solve the model for many potential bases, it is interesting to see the impact of the fractional coverage probabilities on the selected set of bases. Fig. 5 shows the selected bases for the fully deterministic case, which corresponds to the classical MEXCLP, and the fully stochastic case. We see that in the stochastic case, bases are evenly spread out over the city center, so as to provide good coverage to these regions with high call volume. This is not necessary in MEXCLP, because a coverage within 15 min suffices. In three cases, two bases are located close to each other. This is necessary to avoid some demand points to be completely uncovered. This is a direct consequence of the strict 0–1 coverage.



(a) MEXCLP locations.

(b) Optimal locations.

Fig. 5. Maps of base locations selected by deterministic MEXCLP 5(a) and stochastic MEXCLP 5(b).

Table 3
Coverage for different number of bases.

# bases	Coverage	# bases	Coverage
1	0.6680	10	0.9508
6	0.9381	11	0.9517
7	0.9434	12	0.9524
8	0.9481	13	0.9533
9	0.9495	18	0.9553

4.2.3. Limited number of bases

In this part, we investigate the impact of the number of bases on the expected coverage. We run the model for different values of β and compare the coverage. Note that the total number of ambulances is fixed.

Table 3 shows that reducing the number of bases from 9 to 7 does not have a huge impact on the expected coverage. Similarly, adding one or two bases hardly increases the coverage. When no limit is set on the number of bases, which corresponds to a different base for each ambulance, the coverage increases by only 0.58 percentage point compared to the base case. For this coverage to be reached, we need twice as many bases. Ambulance providers should make the trade-off between the cost of an additional base and the increase in coverage.

4.2.4. Sensitivity to travel time distribution

Since higher or lower variation in the travel times might influence the optimal ambulance locations, we compare the outcome of the model for different levels of variation in the travel time distribution. In the base case, we used a coefficient of 0.25, corresponding to a standard deviation of 0.25 times the mean. We vary this value from 0, which corresponds to the case with deterministic travel times, to 0.5. The expected coverage and the number of changes in the ambulance distribution are given in Table 4. The third column gives the number of bases that are located differently, while the fourth column lists the number of ambulances that are assigned to a different base. Note that if a base is located differently, the ambulances assigned to that base are counted as assigned to a different base.

We see that the coverage decreases when the variability in the travel times increases. Due to the relatively high coverage percentage, the loss of coverage as a consequence of a more negative worst-case is higher than the benefit from a better best-case travel time realization. Furthermore, we can conclude that the optimal location of the ambulances does not change significantly for small changes in the variation of the travel times. Only in the two extreme cases, more than one ambulance is located differently.

4.2.5. Sensitivity to busy fraction

The model, as presented, takes the busy fraction of an ambulance as an input. Typically, this busy fraction is hard to estimate and might depend on the selected bases and ambulance distribution. To overcome this, one could use an iterative method where, based on the outcomes of the model, the busy fraction is estimated. With the updated value, the model is solved until some convergence criterion is met [18]. To gain insight in the sensitivity of the model to the busy fraction, we run the model for different values of q. Furthermore, the solution obtained with $q = 0.42$ is evaluated for different busy fractions. The results are shown in Fig. 6. For values of q between 0.3 and 0.5 the solution does not change. Only when very high or very low busy fractions are used in the optimization, we obtain suboptimal solution with respect to a busy fraction of 0.42. Similarly, if we use 0.42 in the optimization, the obtained solution is also optimal for some cases with different busy fractions. Even if the busy fraction is significantly different, the coverage loss as a result of the incorrect estimation is limited. This shows that the model is rather robust against busy fraction estimation errors.

5. Conclusions and future work

In this paper, we presented an ambulance location model based on the maximum expected coverage model, introduced by Daskin [6]. In contrast to the classical MEXCLP, we allow the coverage provided by base *i* to demand point *j* to be fractional. This allows to include stochastic travel times and survival probabilities. These applications were already studied by Ingolfsson et al. [18] and Erkut et al. [19]. They used a non-linear formulation to model fractional coverage probabilities. We presented a linear formulation for this problem, which is proved to be equivalent to their formulation. We compared the computation time of our linear formulation with the non-linear formulation and observed that significant improvement can be obtained. Instances of the

Table 4
Solution for different coefficients of variation (Var). Column two gives the coverage of the optimal solution with respect to a particular Var. Column three gives the coverage of the solution provided by the base case with Var of 0.25 with respect to different coefficients of variation. Column four evaluates the different solutions with respect to the base case of 0.25. Column five and six give the number of bases and ambulances that are located differently.

Var	Coverage	Coverage of solution for Var = 0.25 with respect to different values of Var	Coverage of different solutions with respect to Var = 0.25	Changed bases	Changed assignment
0	0.9623	0.9608	0.9490	2	3
0.1	0.9600	0.9590	0.9494	1	1
0.2	0.9539	0.9535	0.9494	1	1
0.25	0.9495	0.9495	0.9495	0	0
0.3	0.9448	0.9448	0.9495	0	0
0.4	0.9342	0.9342	0.9495	0	0
0.5	0.9231	0.9226	0.9462	1	3

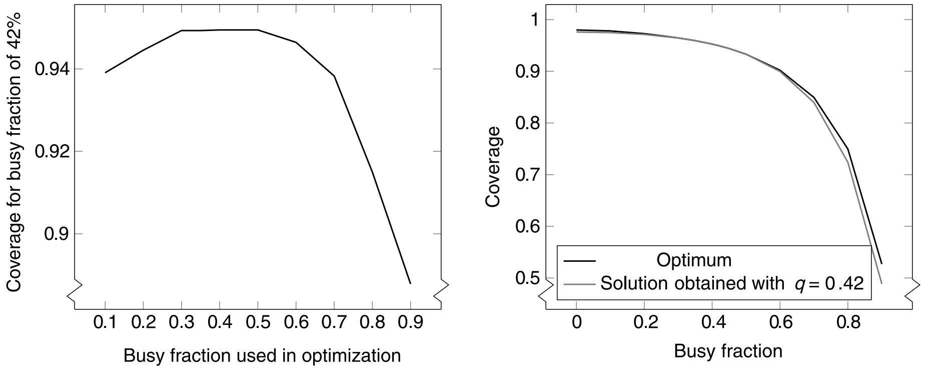


Fig. 6. Impact of busy fraction on obtained solution. On the left, the coverage, with respect to busy fraction of 0.42, of solution obtained with different busy fractions is given. On the right, the solution obtained with busy fraction of 0.42 is compared with optimum for different busy fractions.

non-linear model that take more than 30 min to solve can now be solved within a few seconds. We further applied the model to the region of Amsterdam and observed that higher coverage can be obtained according to our model. Furthermore, we saw that including the randomness in pre-trip delay and travel times has an important impact on the obtained solution. Since travel time distributions are hard to estimate, we evaluated the impact of different levels of variation in the travel time distribution. The results show that small changes in the standard deviation do not have a high impact on the optimal solution. Nevertheless, it would be useful for future research to investigate potential improvements in the estimation of the travel time distributions.

An interesting extension of this research would be to incorporate busy fractions that depend on the base station. This would allow to incorporate workload variations within a region. In the current formulation, this would result in a non-linear model. The results of Section 3 show that tractability benefits significantly from a linear formulation. Hence, investigating potential linear formulations might be worthwhile for future research.

Finally, we highlight that most proposed extensions of the Maximum Expected Coverage Location Model can be included in this model as well. For example, although the model is formulated to maximize the coverage given fixed resources, it can also be used to determine the required number of ambulances to reach a fixed coverage level. This can be done by applying the model iteratively for different values of b.

Acknowledgments

This research was financed in part by Technology Foundation STW under contract 11986, which we gratefully acknowledge. The authors thank Karen Aardal for useful comments. We further thank the reviewers and participants of the 2014 EURO Summer Institute on OR applied to health for fruitful discussions. Finally, we thank Ambulance Amsterdam for providing the data for the case study.

Appendix. Model formulation

In this Appendix, we state both the MILP and MINLP formulation. Both models use the variables x_i and y_i. Here, x_i is the number of ambulances located at base i and y_i takes value 1 if base i is opened and 0 otherwise. Additionally, MILP uses the variables z_{ijk} indicating whether the kth preferred, with respect to w_{ij}, ambulances for demand point j is located at base location i. The two formulations are then as follows.

MILP

$$C^{MILP} = \max \sum_{j \in N} d_j c_j(z) \tag{A.1}$$

with

$$\sum_{k=1}^{b} z_{ijk} \le x_i \quad \forall i \in M,\ j \in N, \tag{A.2}$$

$$\sum_{i \in M} z_{ijk} = 1 \quad \forall j \in N,\ k \le b, \tag{A.3}$$

$$\sum_{i \in M} y_i \le \beta, \tag{A.4}$$

$$x_i \le b_i y_i \quad \forall i \in M, \tag{A.5}$$

$$\sum_{i \in M} x_i = b, \tag{A.6}$$

$$y_i, z_{ijk} \in \{0, 1\} \quad \forall i \in M,\ j \in N,\ k \le b, \tag{A.7}$$

$$x_i \in \mathbb{N} \quad \forall i \in M \tag{A.8}$$

and

$$c_j(z) = \sum_{k=1}^{b} (1-q) q^{k-1} \sum_{i \in M} z_{ijk} w_{ij} \quad \forall j \in N.$$

MINLP

$$C^{MINLP} = \max \sum_{j \in N} d_j c_j(x) \tag{A.9}$$

with

$$\sum_{i \in M} y_i \le \beta, \tag{A.10}$$

$$x_i \le b_i y_i \quad \forall i \in M, \tag{A.11}$$

$$\sum_{i \in M} x_i = b, \tag{A.12}$$

$$y_i \in \{0, 1\} \quad \forall i \in M, \tag{A.13}$$

$$x_i \in \mathbb{N} \quad \forall i \in M \tag{A.14}$$

and

$$c_j(x) = \sum_{i \in M} q^{\sum_{k<ranking(i,j)} x_{a_{kj}}} (1 - q^{x_{a_{ij}}}) w_{a_{ij}j} \quad \forall j \in N.$$

References

[1] C. Toregas, R. Swain, C. Revelle, L. Bergman, The location of emergency service facilities, Oper. Res. 19 (6) (1971) 1363–1373.

[2] R. Church, C. ReVelle, The maximal covering location problem, Pap. Reg. Sci. 32 (1) (1974) 101–118.

[3] M. Gendreau, G. Laporte, F. Semet, Solving an ambulance location model by tabu search, Locat. Sci. 5 (2) (1997) 75–88.

[4] K. Hogan, C. ReVelle, Concepts and applications of backup coverage, Manage. Sci. (1986) 1434–1444.

[5] C. ReVelle, K. Hogan, The maximum availability location problem, Transp. Sci. 23 (3) (1989) 192.

[6] M. Daskin, A maximum expected covering location model: formulation, properties and heuristic solution, Transp. Sci. 17 (1) (1983) 48–70.

[7] J. Repede, J. Bernardo, Developing and validating a decision support system for locating emergency medical vehicles in louisville, kentucky, European J. Oper. Res. 75 (3) (1994) 567–581.

[8] V. Schmid, K. Doerner, Ambulance location and relocation problems with time-dependent travel times, European J. Oper. Res. 207 (3) (2010) 1293–1303.

[9] P.L. van den Berg, K. Aardal, Time-dependent MEXCLP with start-up and relocation cost, European J. Oper. Res. 242 (2) (2015) 383–389.

[10] P. Beraldi, M. Bruni, D. Conforti, Designing robust emergency medical service via stochastic programming, European J. Oper. Res. 158 (1) (2004) 183–193.

[11] P. Beraldi, M. Bruni, A probabilistic model applied to emergency service vehicle location, European J. Oper. Res. 196 (1) (2009) 323–331.

[12] M. Koç, M. Bostancioğlu, A reliability based solution to an ambulance location problem using fuzzy set theory, Int. J. Nat. Eng. Sci. 5 (1) (2011) 13–17.

[13] M. Daskin, Location, dispatching, and routing model for emergency services with stochastic travel times, in: A. Ghosh, G. Rushton (Eds.), Spatial Analysis and Location Allocation Models, Van Nostrand Reinhold, New York, 1987, pp. 224–265.

[14] O. Karasakal, E. Karasakal, A maximal covering location model in the presence of partial coverage, Comput. Oper. Res. 31 (2004) 1515–1526.

[15] V. Marianov, C. ReVelle, The queueing maximal availability location problem: a model for the siting of emergency vehicles, European J. Oper. Res. 93 (1) (1996) 110–120.

[16] J. Goldberg, R. Dietrich, J. Ming Chen, M. Mitwasi, T. Valenzuela, E. Criss, Validating and applying a model for locating emergency medical vehicles in tuczon, AZ, European J. Oper. Res. 49 (3) (1990) 308–324.

[17] J. Goldberg, L. Paz, Locating emergency vehicle bases when service time depends on call location, Transp. Sci. 25 (1991) 264–280.

[18] A. Ingolfsson, S. Budge, E. Erkut, Optimal ambulance location with random delays and travel times, Health Care Manage. Sci. 11 (3) (2008) 262–274.

[19] E. Erkut, A. Ingolfsson, G. Erdoğan, Ambulance location for maximum survival, Nav. Res. Logist. (NRL) 55 (1) (2008) 42–58.

[20] V.A. Knight, P.R. Harper, L. Smith, Ambulance allocation for maximal survival with heterogeneous outcome measures, Omega 40 (2012) 918–926.

[21] M.E. Mayorga, D. Bandara, L.A. McLay, Districting and dispatching policies for emergency medical service systems to improve patient survival, IIE Trans. Healthcare Syst. Eng. 3 (1) (2013) 39–56.

[22] E. Erkut, A. Ingolfsson, T. Sim, G. Erdoğan, Computational comparison of five maximal covering models for locating ambulances, Geogr. Anal. 41 (2009) 43–65.

[23] N. Channouf, P. L'Ecuyer, A. Ingolfsson, A.N. Avramidis, The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta, Health Care Manage. Sci. 10 (1) (2006) 25–45.

[24] H. Setzler, C. Saydam, S. Park, EMS call volume predictions: A comparative study, Comput. Oper. Res. 36 (6) (2009) 1843–1851.

[25] AIMMS BV, AIMMS, The User's Guide, 2013. http://download.aimms.com/aimms/download/manuals/.

[26] ILOG, ILOG Cplex 12.5 reference manual, 2009. http://www-03.ibm.com/software/products/us/en/ibmilogcpleoptistud.

[27] The Optimization Firm, BARON user manual v. 13.0.0, 2014. http://www.theoptimizationfirm.com/downloads/docs/baron.

[28] I. Boers, P. Duijf, G. Leerkes, J. van Rhijn, H. Van der Werff, Ambulances in-zicht 2012, Vereniging Ambulancezorg Nederland, 2013.

[29] G. Kommer, S. Zwakhals, Referentiekader spreiding en beschikbaarheid ambulancezorg 2008, RIVM Briefrapport 270192001/2008, 2008.

Please cite this article in press as: P.L. van den Berg, et al., Linear formulation for the Maximum Expected Coverage Location Model with fractional coverage, Operations Research for Health Care (2015), http://dx.doi.org/10.1016/j.orhc.2015.08.001