POLITECNICO DI MILANO

Facoltà di Ingegneria dei Sistemi

Corso di Laurea Magistrale in Ingegneria Matematica

OPTIMIZATION MODELS FOR MULTIPERIOD PROBABILISTIC AMBULANCE LOCATION

Relatore: **Prof. Edoardo Amaldi**

Correlatore: **Prof.ssa Giuliana Carello**

Tesi di Laurea di:

Erik Cittadino 725502

Anno Accademico 2009-2010

Contents

List of Figures

List of Tables

Abstract

The aim of a Emergency Medical Service (EMS) system is to provide immediate medical care to the population.
EMS managers constantly deal with the problem of improving the system performance, in particular the response time to emergencies. A careful strategical and planning phase is a major prerequisite for the success of a system.

In this work we consider the problem of ambulance location, and we develop a multiperiod probabilistic model.
Our research is carried out in cooperation with Agenzia Regionale Emergenza Urgenza (AREU), the regional agency for medical emergency of Milan, as a follow up of the project *DECEMBRIA* (DECisioni in EMergenza sanitaRIA).
The objective of the model is to determine the minimum number of ambulances needed to meet a predetermined level of reliability for the system, together with their locations.
We propose a multiperiod model, defined on a set of consecutive time periods. Each period represents a time cluster, in which system conditions can be considered as stationary. This way, we take into account the variability of system conditions with respect to time. We also consider the probabilistic aspect of the problem. The chance of a system congestion is modeled with the introduction of the possibility that an ambulance be already busy when called for a service.
A heuristic algorithm based on Lagrangian Relaxation and neighbourhood search is proposed as an instrument to solve the problem.
We report results from the application of the model to the emergency medical system of Milan.

Riassunto della tesi

L'affidabilità del servizio di assistenza fornito da un sistema di primo soccorso dipende da molti fattori.
Il tempestivo intervento dei mezzi sul luogo dell'incidente e le cure offerte dal personale medico prima del raggiungimento di una adeguata struttura sanitaria assumono sicuramente un ruolo di primaria importanza.
Un'eccellente gestione nella dislocazione della flotta (ambulanze, automediche, eliambulanze) sul territorio è senza dubbio uno degli elementi fondamentali per garantire un buon livello di servizio. L'intervento tardivo dei paramedici sul luogo di un incidente può portare in molti casi a tragiche conseguenze.

In questa tesi di laurea si affronta il problema del posizionamento di una flotta di ambulanze.
Numerosi sono gli aspetti da considerare nella formulazione matematica del problema. Innanzitutto, un modello per la dislocazione dei mezzi sul territorio deve tenere conto dei tempi necessari per lo spostamento degli stessi dal luogo di stazionamento al luogo dell'emergenza. Come già accennato, un ritardo di pochi minuti può compromettere la buona riuscita delle operazioni di salvataggio. Pertanto, le condizioni del traffico su tutto il territorio e la loro variabilità nel tempo devono essere ragionevolmente considerate.
Un altro aspetto di cui tenere conto è la quantità di domanda proveniente dal sistema in esame: le zone con un elevato numero di richieste di intervento necessitano di una buona copertura di mezzi, in modo tale da garantire un minimo livello di affidabilità anche in casi di congestione del servizio.
In questo lavoro viene sviluppato un modello di programmazione matematica in grado di considerare questi aspetti del problema. Nello specifico, si

propone un modello multiperiodo di tipo probabilistico.
L'obiettivo è quello di creare uno strumento in grado di guidare il processo decisionale e strategico di un reale sistema di pronto soccorso.

I contenuti della tesi sono sviluppati in un'introduzione e quattro capitoli principali, organizzati nel modo seguente.
Nel capitolo introduttivo si descrive il problema della dislocazione delle ambulanze. Dopo aver esposto le problematiche principali che caratterizzano questo problema, si affronta la questione relativa alla sua modellizzazione.
Il problema viene discretizzato spazialmente e temporalmente. Il territorio di competenza del sistema viene partizionato in piccole zone, ognuna delle quali può essere considerata come un singolo nodo su un grafo. Ogni nodo è connesso ad un sottoinsieme di nodi circostanti grazie ad una matrice di raggiungibilità. L'orizzonte temporale del problema viene suddiviso in una serie di periodi consecutivi, in modo tale che le condizioni di traffico e di domanda all'interno di ogni periodo siano omogenee. In questo modo, ogni periodo può essere considerato come un singolo istante temporale.
Grazie ai dati forniti dal personale della Centrale Operativa del 118 di Milano, è stato possibile prendere in considerazione un reale sistema di emergenza.
Il primo capitolo è dedicato all'esposizione dei principali modelli presenti in letteratura relativi al problema della dislocazione di mezzi di soccorso. La trattazione è organizzata in modo da riflettere le considerazioni effettuate durante la prima parte del nostro lavoro, in cui sono stati valutati i vantaggi e gli svantaggi dei diversi approcci al problema.
Nel secondo capitolo viene sviluppato un modello multiperiodo di tipo deterministico. Grazie all'approccio multiperiodo è possibile considerare la dipendenza del problema rispetto alla variabile temporale. Il nostro obiettivo è di ottenere una soluzione che non presenti molte differenze nelle posizioni delle ambulanze tra un periodo e l'altro. Infatti, una configurazione che richiede la riallocazione di un alto numero di ambulanze ad ogni istante temporale non è applicabile nella realtà.
Il termine deterministico si riferisce all'ipotesi che le ambulanze siano sempre disponibili per rispondere alle chiamate di emergenza. Naturalmente tale

ipotesi non è realistica, siccome per gran parte del tempo le ambulanze sono occupate a servire richieste di intervento.
Questa ipotesi viene corretta con l'introduzione dell'aspetto probabilistico del problema. In particolare, si considera esplicitamente la probabilità che un'ambulanza non sia disponibile a rispondere ad eventuali chiamate di soccorso, in quanto già in servizio. Nel terzo capitolo, il modello deterministico multiperiodo viene riformulato nella sua versione probabilistica. Per ottenere questo scopo, adattiamo al problema in esame una metodologia generale proposta recentemente in letteratura, riguardante il problema del Set Covering robusto.
Il modello sviluppato è velocemente risolvibile in modo esatto su piccole e medie istanze del problema, grazie ad un software di ottimizzazione come CPLEX. Per quanto riguarda istanze di dimensione maggiore, il problema risulta più impegnativo da risolvere. Per questo motivo, nel quarto capitolo proponiamo un algoritmo euristico basato su rilassamento Lagrangiano e ricerca locale. I risultati preliminari ottenuti sono interessanti, e dimostrano che esiste un buon margine di miglioramento per quanto riguarda la soluzione del modello su grandi istanze.

Introduction

The aim of this thesis is to study the important problem of ambulance location.
We consider an emergency medical service system, in which ambulances are dispatched to serve requests received from the population.
An Emergency Medical Service (EMS) system consists of an operation center and a certain number of emergency vehicles, together with the staff involved in the service process.
All the strategical and tactical decisions are taken in the operation center. Emergency calls are received by the operation center staff via telephone calls. For each call, one or more idle vehicles are dispatched to provide immediate medical care.
This is one of the main features of an EMS system: ambulances wait in predetermined sites until they are assigned to a mission. Then, they immediately move to reach the patient.

The aim of an EMS is to satisfy the maximum number of emergency calls, while minimizing the required service times.
This is, actually, a surrogate for the true purpose of EMS, that is reducing as much as possible mortality, disability and suffering in persons [4, 17]. The underlying assumption is that if calls are answered and serviced quickly, then this will lead to better clinical outcomes.
A maximum service time r is usually given, within which requests should be served in order to be considered succesfull. This time threshold can be set by the EMS manager, but in general it is prescribed by the law. The United States Emergency Medical Services Act sets that 95% of requests should be

served within a time limit of 10 minutes in urban areas and 30 minutes in rural areas [2]. In Italy, the time limit is set by the law as 8 minutes in urban areas and 20 minutes in rural areas.

The arrival of an ambulance on the scene of an accident is just one step in the chain of events ending with the success of an emergency mission.
First of all, an emergency call is answered by operation center staff and the severity of the accident and its degree of urgency is quickly evaluated. Four degrees of priority can be assigned to a call: red (high urgency), yellow (medium urgency), green (no urgency), white (no priority). Red and yellow calls require an immediate intervention, to be performed as quickly as possible.
After assigning a priority to the call, the staff takes decisions about the type and number of ambulances to be dispatched. Only when these phases are completed, the intervention of paramedics on the emergency scene can occur.
In general, it is necessary that the emergency vehicles transfer the patient to a hospital. After entrusting the patient to the hospital medical staff, the vehicle returns to an idle state and is available for another call.

It is clear that the good performance of EMS does not only rely on practical operations. Strategical and tactical decisions as the planning of ambulance locations are critical issues. An efficient devising phase is a major prerequisite for EMS success.
EMS administrators and managers often face the difficult task of locating a limited number of ambulances in order to guarantee the best service possible to a constituent population.

The problem of locating emergency vehicles on the territory while optimally managing the limited available resources can be set up in the Operational Research framework.
As mentioned, the aim of this Master of Science thesis is to study the problem of ambulance location and to develop tools to solve it. The objective of

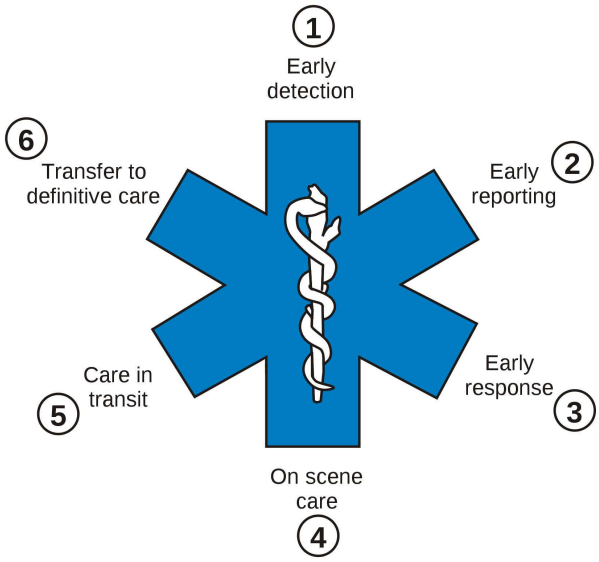


Figure 1: The Star of Life

the problem is to determine the minimum number of ambulances needed to meet a predetermined level of reliability for the system, together with their locations. This is what we refer to when we use the the expression *ambulance configuration*.

We propose a multiperiod model defined on a set of consecutive time periods. Each period represents a time cluster, in which system conditions can be considered as stationary. This way, we take into account the variability of system conditions with respect to time.

We also face the probabilistic aspect of the problem. The chance of a system congestion is modeled considering the possibility that an ambulance is already busy when called for a service.

We solve the proposed model to optimality, with the optimization software CPLEX. A heuristic algorithm based on Lagrangian Relaxation and neighbourhood search is also proposed.

We report results from the application of the model to the emergency medical system of Milan.

Our research is carried out in cooperation with Agenzia Regionale Emergenza

Urgenza (AREU) and the Milan Operation Center C.O.118 at Ospedale Niguarda, as a follow up of the project *DECEMBRIA* (DECisioni in EMergenza sanitaRIA).

Figure 2: AREU: Azienda Regionale Emergenza Urgenza

Discrete model

Every EMS system is responsible for providing medical care on a territory which is typically wide.
Since most urban and suburban EMS have tens of thousands of calls per year, it is impossible to model down to the call level. All the calls in a small area must be aggregated to a single zone, creating a partition of the system. These zones may have any shape, but the assumption is that all calls from a zone are considered as coming from the zone center point. Thus, every zone can be considered as collapsing into a single point.
In order to guarantee that such hypothesis is reasonable, the zone size has to be carefully chosen. If a zone is too big, different travel times may be necessary to reach two points lying in the same zone. This obviously leads to inaccuracies in travel times handling. On the contrary, choosing a small size leads to large instances, which may be very hard to solve.

We partition the territory in a set of square zones, and we represent each zone as a point.

Two sets of points are usually considered in EMS models: the set I of demand points i and the set J of location points j. The first set contains the points from which it is possible to receive a call. In most situations, all the zones are considered as potential demand points. Location sites set J contains all the sites in which an ambulance can be located when idle and waiting for calls. Typically, ambulances can be placed almost everywhere when they are not busy, even at very rudimental locations such as parking lots or big crossroads. We define a graph, by introducing a set of edges connecting the location and demand points. The rule is that two points i and j are connected by an edge if the travel time t_{ij} needed to move from i to j is less than a fixed value r. Typically, the parameter r is fixed to a value lower than or equal to the maximum service time set by the law.
To take into account the time threshold r, we introduce a *reachability matrix*, whose binary elements a_{ij} state if two points are reciprocally connected:

$$a_{ij} = \begin{cases} 1 & \text{if } t_{ij} \leq r \\ 0 & \text{otherwise} \end{cases} \tag{0.1}$$

Description of data

In our work we consider a real Emergency Medical Service system, in order to validate the introduced models and to present results relative to a real system.
In particular, we study the EMS system of Milan, one of the main Italian cities. All the necessary data was provided by our partners at AREU.
The territory of Milan is very interesting for the emergency vehicles location problem.
In fact, Milan comprises urban areas as well as rural and residential ones. In addition, the critical and highly variable traffic conditions of the city offer an excellent scenario for the analysis to be performed.
In Fig. 3, the city of Milan and its main districts are exposed.

The territory of Milan has been partitioned into a set of 492 square zones, as it can be observed in Fig. 4.

Figure 3: The city of Milan.

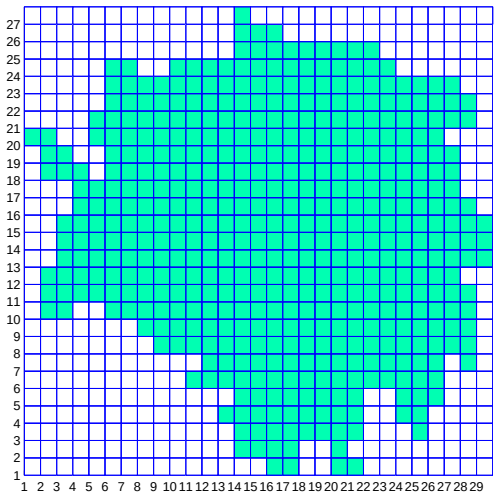


Figure 4: Partition of the territory of Milan.

The area of each zone is $0.36Km^2$, corresponding to a zone edge of 600 meters. This choice is guided by an analysis of travel times performed on historical data. The aim is to guarantee that every point belonging to a square is reachable within the maximum time standard r from the same subset of

potential ambulance sites.
As indicated by AREU staff, all the zones have been considered as potential location points. This is due to what previously observed: within an area of $0.36Km^2$, it is always possible to find a good location for ambulances.

The reachability matrix was obtained considering the average traffic conditions in every zone of the city, during different time intervals. A statistical analysis was performed in order to identify a set of time periods, such that the traffic conditions are homogeneous within each of them and for each zone of the city. Thus, it was possible to define a different reachability matrix for each homogeneous time period. A detailed description of the performed analysis and the relative results can be found in [31].
Six homogeneous time periods within a workday were identified after the analysis:

$$\begin{array}{cc} 7-9 & 9-14 \\ 14-16 & 16-19 \\ 19-21 & 21-23 \end{array}$$

The traffic conditions can be considered as homogeneous during every time interval. The average travel times do not vary significantly within the same period.
The obtained periods reflect the characteristics of a typical working day in a big city.
In the early morning $(7-9)$ many people living in the suburbs travel to reach their offices in the center of the city. In that period traffic conditions are particularly critical. During the day, three main periods are distinguishable: one in the morning and two in the afternoon. Traffic conditions vary also during the last part of the day, when people return to their houses. Two different evening periods are detected.
The effects of traffic conditions in the city can be observed in Fig. 5. In the picture, the reachability zones of three different locations are highlighted by red squares. The results are relative to the first time period $(7-9)$.
The three reachability zones do not have a regular shape. In addition, they

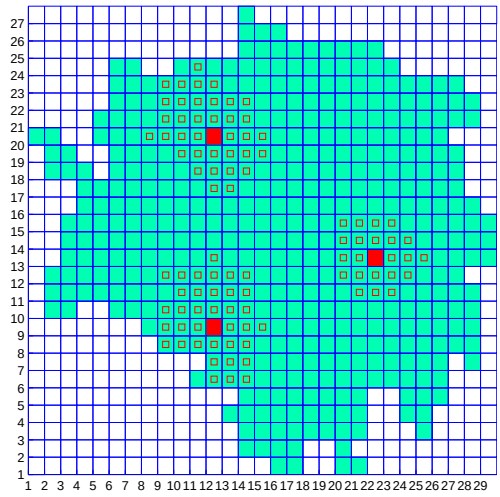


Figure 5: Reachability zones

have different extensions.

This is due to the variable travel times detected in each zone of the city. For example, the western reachability zone is smaller than the others, since it is closer to the historical center of the city.

The definition of a set of time periods is of fundamental importance for our work. In fact, our objective is to formulate a new model which is able to explicitely consider different time periods, in order to introduce a time-dependence into the ambulance location problem.

Chapter 1

Literature Review

In this chapter we propose and comment some of the models present in literature regarding the emergency vehicles location problem. This review aims at introducing the reader to the important problem of vehicles location and at outlining the state of the art in this research field.
This review, which does not mean to be fully exhaustive, is organized so as to revise the path followed during the early stage of this work. All the observations and comments that led to the development of our multiperiod probabilistic model are illustrated and clarified.

The ambulance location models are usually divided into two main categories: deterministic models and probabilistic models.
Deterministic models aim at finding an optimal configuration for the location of ambulances without considering any stochastic aspect in the problem formulation.
In particular, they assume that ambulances are always available when called for service. Actually, this assumption is not true. In fact, ambulances are idle very rarely during the day, because of the large quantity of requests received by the operational center. In big cities, it is almost impossible that all ambulances are not working and ready for answering calls at the same moment, as it is assumed in deterministic models.
The fact that an ambulance might be busy when it is called for a mission can cause delay in answering the call. Clearly, fatal consequences could arise

in these cases.
The pobjective of probabilistic models is to overcome this weakness.
In such models the probability that an ambulance is not available is explicitely considered. This probability is usually referred to as *busy probability*.
As a consequence, solutions of probabilistic models are more stable with respect to stochastic events.

Emergency vehicle location models can be also divided into static models and dynamic models.
Static models represent the first born models in the field of emergency vehicles location. They are developed without taking into account the time dependence of the problem. Averaged data is exploited to obtain solutions which are completely time-independent.
Since system conditions (i.e. traffic intensity, service demand, ambulance availability,...) usually vary during the time, averaged data cannot fully describe their behaviour during long periods.
Dynamic models introduce a dynamic feature into the location problem. They are formulated so as to provide solutions which depend at every instant on the current system conditions.

1.1 Deterministic models

The first models for the emergency vehicle location were proposed in the early 70s. Many authors focused their work on this very important matter.

One of the earliest models presented in literature is the Location Set Covering Model (LSCM). This model, proposed by Toregas et. al in 1971 [33], approaches the location problem as a Set Covering Problem (SCP).
The aim of the model is to determine the minimum number of necessary ambulances to cover all the demand points. A demand point is said to be covered if it can be reached by at least one ambulance within the maximum service time r.
Thanks to the reachability matrix introduced in Chapter , it is possible to

define the sets J_i containing the location points j which can be reached from demand point i within the time limit:

$$J_i = \{j \in J : t_{ij} \leq r\} = \{j \in J : a_{ij} = 1\}.$$

In LCSM model, binary variables x_j indicate if an ambulance is located in zone j:

$$x_j = \begin{cases} 1 & \text{an ambulance is located in } j \\ 0 & \text{no ambulances are located in } j \end{cases}$$

The model is written as follows:

$$\text{(LSCM)} \quad \text{minimize} \quad \sum_{j \in J} x_j \tag{1.1}$$

$$\text{subject to} \quad \sum_{j \in J_i} x_j \geq 1 \quad \forall i \in I \tag{1.2}$$

$$x_j \in \{0, 1\} \quad \forall j \in J \tag{1.3}$$

The objective function (1.1) minimizes the total number of deployed ambulances. The covering constraints (1.2) state that every possible demand point must be covered by at least one ambulance.
This model presents the problem mentioned in the previous section: it gives an ambulance configuration which is able to cover all possible demand points only provided that every single vehicle is available for requests. When an ambulance is dispatched to a service, it leaves its area completely uncovered. Clearly, if a new request is received from that area, a farther ambulance has to be assigned to the emergency. Longer travelling times are then required.
Although very simple, LSCM model is very useful since it rapidly produces a lower-bound for the number of ambulances needed to cover the entire territory.

An alternative approach to that of LCSM is given by MCLP (Maximal Coverage Location Problem), a new model proposed by Church and ReVelle in 1974 [9].
The authors stated that also the number of requests in each demand zone should be considered when solving the problem. A zone with a very high rate

of call requests should be covered by a larger number of ambulances.
In the formulation of MCLP, the number of ambulances to be allocated is an input of the model and not a result as in LSCM: it is represented by the parameter p. Given this limit on the available resources, MCLP maximizes the quantity of potential covered calls.
Demand data is easy to obtain on the basis of historical statistics, or to be simply inferred from the population density of each zone. The parameter d_i represents the number of calls coming from demand point i.
The binary variable y_i is equal to 1 if and only if zone i is covered by at least one ambulance. The binary variable x_j is equal to 1 if an ambulance is located in site j.
The model is formulated as follows:

$$\text{(MCLP)} \quad \text{maximize} \quad \sum_{i \in I} d_i y_i \tag{1.4}$$

$$\text{subject to} \quad \sum_{j \in J_i} x_j \geq y_i \quad \forall i \in I \tag{1.5}$$

$$\sum_{j \in J} x_j = p \tag{1.6}$$

$$x_j \in \{0,1\} \quad \forall j \in J \tag{1.7}$$

$$y_i \in \{0,1\} \quad \forall i \in I \tag{1.8}$$

The objective function (1.4) maximizes the number of covered calls. Constraints (1.5) state the relationship between the variables y_i and x_j: zone i is covered if and only if there is at least one ambulance within its neighbourhood J_i. Contraint (1.6) sets the maximum number of deployed ambulances. MCLP model aims at making the best possible use of limited resources, given a deterministic demand pattern.

Both LSCM and MCLP are quite simple, but it must be underlined that they can provide relevant information in a very small amount of time. Furthermore, such models represent the basis in emergency vehicles location problems.

The formulation of new models was suggested by the instability of LSCM and MCLP solutions with respect to little changes in data, such as travel times or availability of ambulances. The main idea is that more stable configurations can be obtained if more than one ambulance is deployed to cover the same demand point. This way, a backup coverage is guaranteed in case of many emergencies occurring at the same time.
Requiring a multiple coverage for demand zones represent an answer to partially fix the problems of the deterministic approach.

In 1986 Hogan and ReVelle [20] proposed two extensions of MCLP model in which backup coverage is considered.
In the first one, denominated BACOP1, the fraction of demand covered twice is maximized. The authors considered a deterministic demand pattern for each zone i of the system, which is represented by the parameter d_i.
As in MCLP model, binary variables x_j indicate if an ambulance is located in point j. Binary variables u_i are introduced: they assume the value 1 if zone i is covered by at least two vehicles within the standard time r.
The model is formulated as follows:

$$\text{(BACOP1)} \quad \text{maximize} \quad \sum_{i \in I} d_i u_i \tag{1.9}$$

$$\text{subject to} \quad \sum_{j \in J_i} x_j - u_i \geq 1 \quad \forall i \in I \tag{1.10}$$

$$\sum_{j \in J} x_j = p \tag{1.11}$$

$$u_i \in \{0,1\} \qquad \forall i \in I \tag{1.12}$$

$$x_j \in \{0,1\} \qquad \forall i \in I \tag{1.13}$$

Objective function (1.9) maximizes the number of calls which are covered twice. At least single covering is required for every demand point, as stated by constraints (1.10). Constraint (1.11) sets the maximum number of ambulances to be deployed.

The second model proposed in [20] is BACOP2. In combination to the previous one, it includes an objective function considering a balance between

demand covered once and twice, weighted by the parameter θ. This weight can be set by the EMS manager so as to reflect a predetermined coverage policy. Parameter d_i stands for the quantity of demand received from zone i, and p is the maximum number of available vehicles.
Binary variable x_j is equal to 1 if a vehicle has to be located in zone j. Binary variables y_i and u_i indicate if zone i is covered once and twice, respectively. The model is written as follows:

$$\text{(BACOP2)} \quad \text{maximize} \quad \theta \sum_{i \in I} d_i y_i + (1-\theta) \sum_{i \in I} d_i u_i \tag{1.14}$$

$$\text{subject to} \quad \sum_{j \in J_i} x_j - y_i - u_i \geq 0 \qquad \forall i \in I \tag{1.15}$$

$$u_i - y_i \leq 0 \qquad \forall i \in I \tag{1.16}$$

$$\sum_{j \in J} x_j = p \tag{1.17}$$

$$x_j \in \{0,1\} \qquad \forall i \in I \tag{1.18}$$

$$y_i \in \{0,1\} \qquad \forall i \in I \tag{1.19}$$

$$u_i \in \{0,1\} \qquad \forall i \in I \tag{1.20}$$

Objective function (1.14) includes the balance between demand covered once and twice. A value of θ near to one means that single covering of demand is preferred over double covering. Decreasing the value of this parameter makes the double coverage assume more importance in the balance. When θ is set to a null value, the objective function is exactly the same as BACOP1's.
Constraint (1.15) means that every demand zone must be covered at least once by an ambulance. A hierarchy between variables y_i and u_i is set by constraint (1.16), so that a point cannot be covered twice if it is not covered once. The maximum number of deployable ambulances is stated by constraint (1.17).

1.2 Probabilistic Models

Deterministic models represent a good instrument for approaching an ambulance location problem. They are able to provide within reasonable time

some relevant information, useful to get an insight of the problem. For example, it is possible to quickly evaluate the minimum number of ambulances needed to cover the entire territory.
However, these models turn out to be quite rough when applied to real systems. As already mentioned, deterministic models do not take into account the possibility that a vehicle could be busy when called for a service.
The aim of probabilistic models is to raise the quality of models' solutions, so as to overwhelm an approach which turns out to be quite naive.

One of the first probabilistic models was proposed by Daskin in 1983, the Maximum Expected Covering Location Problem (MEXCLP) [10].
The author assumed that all the vehicles in the system has the same probability of being busy when called for service. This probability is referred to as q. This idea was mutuated from a work published by Chapman and White in 1974 [8], in which the authors proposed to consider a system-wide busy fraction. The important underlying hypothesis is that ambulances are independent of each other.
Thanks to this assumption, it is possible to derive an expression for the expected covered demand of each point in the system. If demand point i is covered by k ambulances (i.e. there are k ambulances in its neighbourhood J_i) its expected covered demand is $E_k = d_i(1 - q^k)$, where d_i is the total demand coming from i. The marginal contribution of the k-th ambulance to this expected value is $E_k - E_{k-1} = d_i(1 - q)q^{k-1}$.
MEXCLP model aims at maximizing the expected covered demand over the entire system, with a given number of p vehicles.
In this formulation the integer variables x_j represent the number of ambulances located in point j, while the binary variables y_{ik} state if the demand point i is covered by at least k ambulances.

The model is as follows:

$$\text{(MEXCLP)} \quad \text{maximize} \quad \sum_{i \in I} \sum_{k=1}^{p} d_i (1-q) q^{k-1} y_{ik} \tag{1.21}$$

$$\text{subject to} \quad \sum_{j \in J_i} x_j \geq \sum_{k=1}^{p} y_{ik} \qquad \forall i \in I \tag{1.22}$$

$$\sum_{j \in J} x_j \leq p \tag{1.23}$$

$$x_j \geq 0, \text{ integer} \qquad \forall j \in J \tag{1.24}$$

$$y_{ik} \in \{0,1\} \qquad \forall i \in I,\ k = 1, \ldots, p \tag{1.25}$$

As already mentioned, the objective function (1.21) maximizes the expected covered demand, given the ambulance busy probability q. Constraints (1.22) state that the vehicles covering zone i cannot be more than vehicles located in its neighbourhood J_i. Since the objective function is concave in k, these constraints will be satisfied as equalities in an optimal solution. Contraint (1.23) limits the number of deployable ambulances.

Clearly, this model represents an improvement with respect to the previous models, since it considers the possibility of system congestion. However, its assumptions are quite simplifying.
Actually it is not true that ambulances are independent of each other: the amount of time in which they are not available strongly affects the availability of other vehicles. When an ambulance is called for service, it causes a change in the system configuration and an increase in the workload of nearby ambulances. In fact, they are supposed to respond to the potential calls coming from the uncovered zone. Thus, the independence assumption is only an approximation of the behaviour of real systems.
Despite of that MEXCLP has to be mentioned because it, together with Chapman and White's work [8], led the way to the formulation of many other probabilistic models. Furthermore, the concept of expected covered demand represents a good indication about the effectiveness of a location configuration.

Chapman ad White's work also introduced the idea of redefining the deterministic location set covering problem (LSCP) in order to formulate its probabilistic version. The authors wanted to explicitely take into account a desired reliability level for the system.
Their purpose was to obtain a configuration such that the probability that each demand area has an available ambulance within the maximum time r is greater or equal to a value α. This formulation of the problem implicitely contains the definition of reliability level α for a system: a system is said to be α-reliable if

$$\mathbb{P}[i \,\text{is covered}, \forall i \in I] \geq \alpha.$$

On the basis of this definition, an interesting work was developed by ReVelle and Hogan in 1989 [27]. They analyzed the deterministic models LSCM and MCLP and translated them into their probabilistic versions. These probabilistic models consider a different busy fraction for each zone of the city.
PLSCP model (Probabilistic Location Set Covering Problem) was developed on the basis of LSCM, by converting its deterministic coverage constraints (1.2) into probabilistic ones.
ReVelle and Hogan reformulated the definition of α-reliable system:

$$1 - q_i^{\sum_{j \in J_i} x_j} \geq \alpha, \quad \forall i \in I$$

where the quantity q_i represents the busy probability of a vehicle covering demand point i. They also proposed a method to approximate the values q_i. After the introduction of the value F_i, which represents the average service time of an ambulance covering i, it is possible to write a relationship between the busy fraction q_i and the number of ambulances located in the neighbourhood of zone i:

$$q_i = \frac{F_i}{\sum_{j \in J_i} x_j}$$

Thus, the reliability condition for the system can be written as

$$1 - \left(\frac{F_i}{\sum_{j \in J_i} x_j} \right)^{\sum_{j \in J_i} x_j} \geq \alpha, \quad \forall i \in I.$$

This expression has no exact inverse, hence there is no analytical solution for the number of vehicles needed to cover a point i. Despite of that, there exists a numerical solution $b_i = \sum_{j \in J_i} x_j$ which can be computed through

$$1 - \left(\frac{F_i}{b_i}\right)^{b_i} \geq \alpha.$$

Thanks to the value b_i, it is possible to define a new coverage constraint for every demand point in the system:

$$\sum_{j \in J_i} x_j \geq b_i. \tag{1.26}$$

Constraint (1.26) has a probabilistic meaning since b_i depends on the chosen parameter α, but it is in fact defined as a deterministic one. In order to guarantee that a point i is covered with α-reliability, at least b_i ambulances have to be located in its neighbourhood.
The model proposed by ReVelle and Hogan is:

$$\begin{aligned}
\text{(PLSCP)} \quad \text{minimize} \quad & \sum_{j \in J} x_j & & (1.27)\\
\text{subject to} \quad & \sum_{j \in J_i} x_j \geq b_i & \forall i \in I \quad & (1.28)\\
& x_j \geq 0, \text{integer} & \forall j \in J \quad & (1.29)
\end{aligned}$$

The objective function (1.27) minimizes the number of total deployed ambulances. The reliability level is forced by coverage constraints (1.28).

ReVelle and Hogan's MALP (Maximum Availability Location Problem) [29] can be considered as the probabilistic counterpart of Church and Revelle's MCLP, since its objective is maximizing the total covered demand.
As in their previous work, the authors approached the stochastic aspect of the problem introducing the request for demand zones to be covered by at least b_i ambulances.
Two versions of this model are present in literature.

The first version, called MALP I, makes the assumption that every point has the same busy probability q. Therefore, every demand point i needs the same number of vehicles b to be covered.
In MALP I, binary variables x_j assume value 1 if a vehicle is located into site j. Variables y_{ik} state whether demand point i is covered by at least k ambulances or not. Parameter d_i represents the quantity of demand coming from zone i, and parameter p sets the maximum number of available ambulances. The model is formulated as

$$\text{(MALPI)} \quad \text{maximize} \quad \sum_{i \in I} d_i y_{ib} \tag{1.30}$$

$$\text{subject to} \quad \sum_{k=1}^{b} y_{ik} \leq \sum_{j \in J_i} x_j \quad \forall i \in I \tag{1.31}$$

$$y_{ik} \leq y_{i,k-1} \quad \forall i \in I, k = 2, \ldots, b \tag{1.32}$$

$$\sum_{j \in J} x_j = p \tag{1.33}$$

$$x_j \in \{0, 1\} \quad \forall j \in J \tag{1.34}$$

$$y_{ik} \in \{0, 1\} \quad \forall i \in I,\ k = 1, \ldots, p \tag{1.35}$$

The objective function (1.30) maximizes the demand covered with at least b ambulances. Constraints (1.31) assure that the number of vehicles covering a zone i cannot be greater than the number of vehicles that are actually located in its neighbourhood. Since the concavity property observed in MEXCLP no longer holds, constraints (1.32) are needed. In fact, a zone cannot be covered k times if it is not covered $k-1$ times. The maximum number of deployable ambulances is constrained by (1.33).

MALP II is formulated relaxing the assumption of independence between ambulances: different busy probabilities q_i are considered. Both models have the same formulation: the parameter b has simply to be changed into b_i.

ReVelle and Hogan point out that the busy probabilities should not be obtained *a priori*, like they did in their work. Actually, busy probabilities are an output of the model and not an input.

In fact, the fraction of time an ambulance is not available depends on many factors. Among the others are demand pattern, travel times, service times. Furthermore, the busy fraction of a vehicle depends on the number and the position of nearby ambulances as well. Clearly, two ambulances improve reciprocally their performances if they work in the same area, covering the same fraction of demand.
Given an ambulance location plan, busy probabilities can be easily calculated, but these quantities vary if the configuration is chaged.

Busy probabilities can be obtained by considering ambulances as servers.
Thanks to queuing theory, it is possible to take into account the relationship between nearby located ambulances.
The key assumption is that every zone is modeled as a $M/M/S$-loss queuing system. The whole territory is divided into a number of neighbourhood zones, each of them containing s servers. Each server can handle at most one call at a time, and no queue is allowed: if a call is received when all servers are busy, it is permanently lost. Both calls interarrival and service times can be modeled as a Poisson distribution.
Thanks to this approach, it is possible to calculate the busy probabilities p_s of s servers located in a neighbourhood. Afterwards, parameters b_i can be obtained and used in the models previously presented. In their 1994 work [25], Marianov and Revelle apply queuing theory to develop an improved version of PLSCP, called Q-PLSCP.

The problem of calculating busy probabilities of vehicles in a system was also studied by Larson, during the early 70s. He developed a spatially distributed queuing model, which is capable of handling these quantities in different configurations: the hypercube model [23].
This model still considers the system as a M/M/s queuing system, but differs from the previous models since it does not make the assumption that neighbourhoods are independent. Servers are located on the territory, and each of them affects the busy probability of the other ones.
The hypercube model is a descriptive model; in order to determine the busy

probabilities, it evaluates every possible state that the system can assume. A binary variable is assigned to each ambulance to indicate if it is busy or not. Thanks to these variables it is possible to generate a hypercube, in which each edge is related to an ambulance. Each vertex of the hypercube represents a possible state of the system. For example, if there are only 2 vehicles in the system, the hypercube is simply a unit square. The configuration such that both ambulances are busy is represented by the vertex $(1, 1)$.
While solving the model, the algorithm moves through the entire hypercube examining every potential state of the system. This way, it is possible to generate a system of linear equations, which must be solved in order to obtain the busy probabilities.
Given a system configuration, the hypercube model is also able to evaluate a variety of performance measures relevant for decision-making: server workloads, mean user response times, fraction of dispatches of each server to each region.
This model, although giving very accurate results, requires an extremely high computational work. The system to be solved consists of 2^n equations, where n is the number of vehicles in the system. Moreover, a high amount of time is needed to generate the coefficients of the linear system.
Many authors tried to integrate the hypercube approach into different location models, so as to develop improved probabilistic models. This led to good results under a theoretical point of view, but extremely useless results under a practical one. In fact, the high computational requirements of hypercube make the models not solvable when applied to real problems, due to the high number of vehicles and location points.
An extension of MEXCLP and MALP models was proposed by Galvão and Morabito in 2008 [15]. The authors considered approximate methods for solving the hypercube model, as explained by Larson in 1975 [24]. Nonetheless, they report results from very small problem instances only.

1.3 Dynamic Models

The models for ambulance location can be further developed by considering its dynamic aspects.
The previously presented models aim at determining a reliable ambulance configuration, which does not depend on time. In every case, only one set of data is considered when solving the model, and solutions are static.
Actually, ambulance location problems were proposed to solve real-life situations which are not static. System's conditions are typically very fluctuating during the time. Quantities like travel times or demand pattern can significantly change within few hours, for example due to living habitudes of people. Different situations occur during a day. Consider the peak hours, when many people move to the center of cities to work. Clearly, the demand pattern during this period is totally different from the night time, when people leave cities and return to their housess which are usually located in city suburbs. Also the traffic conditions in the system are influenced.
Despite of that, static models do not consider a time-variability in system's conditions, since quantities are averaged on a single time interval.
Therefore, static solutions can sometimes turn out to be overconservative. Moreover, it can happen that the daily-averaged data is not able to describe some hard moments during the day, typically traffic peak hours, in which stronger coverage would be required.

These features of the problem suggest that an improved formulation of the models should be considered.
A good idea for stabilizing the solution with respect to changes in the data and to stochastic events is to dinamically approach the problem. In particular, dynamic relocation of ambulances can be taken in consideration. For example, every time a call for service arrive to the central and an ambulance is dispatched, the position of idle ambulances can be changed so as not to leave areas unprotected.

The first dynamic model was presented by Kolesar and Walker in 1974 [22].

The authors considered the problem of dynamically allocating fire companies in the city of New York. They introduced a minimum coverage standard to be always satisfied for all demand points: at least a certain number of vehicles has to be available for service. In order to satisfy this standard, relocation of idle vehicles is allowed. The process is managed by a real-time algorithm which repeatedly solves optimization problems.
The objective is to choose which fire companies should be relocated so as to mantain a high service quality.
The proposed algorithm is divided into four stages:

1. determining the need for a relocation;
2. determining the empty location points to be filled, minimizing the number of changes from the actual configuration;
3. determining the companies available for relocation, minimizing the expected response time;
4. determining the relocation assignments, minimizing travel distance.

Step 1 is achieved by a program, called trigger, which runs constantly and monitors the system's behaviour. Steps 2, 3 and 4 are carried out solving integer programming problems. For example, in step 2 a Set Covering Problem is solved considering uncovered demand points.
Since the algorithm has to work in real-time, long computing time are not affordable for optimization problems. Hence, the authors proposed to solve them through heuristic algorithms.

The authors reported computational results from the application of the algorithm to a difficult scenario happened in the Bronx, a New York borough. The tests made with the help of a simulation program showed that the dynamic approach is very effective and can significantly improve the choices made without this instrument.
However, it must be underlined that the size of the presented problem is very small. 25 potential location points had been used.

Another important dynamic model was proposed by Gendreau et al. in 2001 [16].
The authors aimed at developing a real-time instrument, which can manage the redeployments of ambulances and constantly guarantee a desired level of coverage. They formulate a model called RP^t, where t represent the time variable.
The coverage constraints used in this case are slightly different from the previous ones: the entire demand has to be covered by at least one vehicle within a time standard r_2, while a portion α of demand has to be covered within a time $r_1 < r_2$. Obviously, the model might be easily modified to obtain coverage constraints of the same tipe as the ones considered in previous models.
RP^t model includes a parameter M_{jl}^t associated with the relocation of ambulances at a time istant t. This parameter penalizes, in the objective function, repeated relocations of the same vehicles. Also round trips or long relocations are penalized.
The variables used in the model are y_{jl} and x_i^k. They are both binary: y_{jl} assumes value 1 if the ambulance l is located in site j, while x_i^k is equal to 1 if and only if demand point i is covered at least k times.
The parameter d_i represents the demand coming from i, p is the maximum number of deployable ambulances and p_j is the maximum number of ambulances that can be located in point j. Binary coefficients γ_{ij} and δ_{ij} indicate whether point i is reachable from point j within r_1 and r_2 time units, respectively. The fraction of demand to be covered within r_1 time units is represented by α.

The formulation of the model is

$$
\begin{aligned}
(\mathrm{RP}^{\mathrm{t}}) \quad \text{maximize} \quad & \sum_{i\in I} d_i x_i^2 - \sum_{j\in J}\sum_{l=1}^{p} M_{jl} y_{jl} && && (1.36)\\
\text{subject to} \quad & \sum_{j\in J}\sum_{l=1}^{p} \delta_{ij} y_{jl} \geq 1 && \forall i \in I && (1.37)\\
& \sum_{i\in I} d_i x_i^1 \geq \alpha \sum_{i\in I} d_i && && (1.38)\\
& \sum_{j\in J}\sum_{l=1}^{p} \gamma_{ij} y_{jl} \geq x_i^1 + x_i^2 && \forall i \in I && (1.39)\\
& x_i^2 \leq x_i^1 && \forall i \in I && (1.40)\\
& \sum_{j\in J} y_{jl} = 1 && l = 1,\ldots,p && (1.41)\\
& \sum_{l=1}^{p} y_{jl} \leq p_j && \forall j \in J && (1.42)\\
& x_i^k \in \{0,1\} && \forall i \in I,\ k \in \{1,2\} && (1.43)\\
& y_{jl} \in \{0,1\} && \forall j \in J,\ l = 1,\ldots,p && (1.44)
\end{aligned}
$$

Objective function (1.36) maximizes double covered calls while penalizing ambulance relocations. Constraints (1.37) and (1.38) express the single and the double coverage requirements, respectively. In particular, all the demand must be covered within r_2 time units and a fraction α of demand within r_1 time units. (1.39) denote that the number of ambulances locates within r_1 time units from point i must be at least one if $x_i^1 = 1$ and at least two if $x_i^2 = x_i^1 = 1$. Constraints (1.40) state that a zone cannot be covered twice, if it is not covered once. Constraint (1.41) indicates that all the ambulances must be assigned to a location point. An upper bound on the number of vehicles located in a point j is set by (1.42).

The method proposed by the authors is a sequential tabu search algorithm. The model should be solved for each istant t in which a call appears in the system, so as to give indications on the relocations to be made. These indica-

tions should be obtained as fast as possible, in order to execute them rapidly. A sequential tabu search algorithm is perfectly adequate for a static location problem, but in the dynamic case a faster and more powerful instrument is highly required. The authors propose an innovative approach, based on a parallelization of the code, which tries to take advantage from the available time between consecutive calls.
The strategy is to precompute a relocation plan for all possible future scenarios. The model is solved repeatedly, each time assuming that a different ambulance will be called for service. When one of these scenarios appears, a precomputed solution is already present. Instantaneous instructions can then be given to the ambulance crews.
It can also happen that a new call is received when the computation of all scenarios is not complete. In this case, which should rarely happen, no redeployment takes place.
The authors developed a solid algorithm working in parallel on different CPUs. During the execution of the code it is also possible to update the model data, i.e. the demand pattern or the travel times.

Another approach for the dynamic formulation of the model was proposed by Andersson and Värbrand in 2007 [1].
The authors introduced the concept of preparedness, as a way of evaluating the ability of the system to serve potential patients now and in the future. The level of preparedness in the system is constantly monitored, until it decreases under a safety value. Typically, this event occurs when many ambulances from the same zone are busy. In this case, a relocation problem is solved, so as to raise the level of preparedness.

The authors proposed the following model for solving the relocation problem:

$$
\begin{aligned}
\text{(DYNAROC)} \quad \text{minimize} \quad & z && && (1.45)\\
\text{subject to} \quad & z \geq \sum_{j \in J^k} \tau_j^k x_j^k && k = 1, \ldots, A && (1.46)\\
& \sum_{j \in J^k} x_j^k \leq 1 && k = 1, \ldots, A && (1.47)\\
& \sum_{k=1}^{A} \sum_{j \in J^k} x_j^k \leq M && && (1.48)\\
& \frac{1}{c_j} \sum_{l=1}^{L_j} \frac{\gamma^l}{t_j^l(\mathbf{x})} \geq P_{min} && j = 1, \ldots, N && (1.49)\\
& \mathbf{x} \in \{0, 1\} && && (1.50)\\
& && && (1.51)
\end{aligned}
$$

In this model binary variables x_j^k assume value 1 if ambulance k is relocated into point j. The variable z represents the maximum travel time for any of the relocated ambulances.
The objective function (1.45) minimizes the value of z. Constraint (1.46) states that z has to be greater than or equal to any of the travel times τ_j^k, which is the time required for ambulance k to reach zone j. Each ambulance k can be relocated to at most one location point in its neighbourhood J_k, as stated by constraint (1.47). Constraint (1.48) sets the maximum number of allowed relocations. The desired preparedness level for the new solution is controlled by constraint (1.49).

Dynamic models are without any doubt a powerful instrument for monitoring and managing an emergency system. Thanks to their real-time features they do not suffer the deficiencies that static models have with respect to variations in data and system configurations. Despite of that, they present some weak points that make them almost useless when applied to real situations.
Dynamic models require a high amount of computational work to be solved. For example, Geandreau's RP^t model was solved through the use of eight CPUs working in parallel. In most cases, such resources are not available in

normal operational centers.
This is often the main weakness of dynamic models: they can be solved only on small size istances.
In addition, a solid software is needed to handle a dynamic model: an easy and powerful user interface has to be implemented. Otherwise, a difficult software would be manageable only by a highly trained staff.
Moreover, the model database has to be updated in real time with all the received service calls and the current positions of ambulances.
Dynamic models have a big potential, since they are accurate and flexible. However, they are not affordable for managing real emergency systems. In the future, when much more powerful computational resources will be available, dynamics models will be exploited in every operational center.

1.4 Multiperiod Models

Static models aim at determining a time-independent configuration, able to guarantee a certain level of security in every possible situation.
Since system conditions are typically not stationary, static solutions are inadequate: they can turn out to be overconservative or not sufficiently reliable.
In the previous section, dynamic models have been introduced. They were proposed with the aim of producing time-dependent solutions, on the basis of the current conditions of the system. Dinamic models consider the possibility of changing the ambulance configuration, making vehicle relocations. Dynamic models are difficult to be exploited in real systems because of their complexity and their huge computational costs.

During the last years a new kind of models was proposed: multiperiod models. They can be considered as a connection between static and dynamic models. Multiperiod models are based on a static approach. They are formulated so as to improve static models to a dynamic concept: multiperiod models introduce the possibility of considering the solution as time-dependent. Thus, more reliable and realistic solutions can be obtained. Obviously, the accuracy of dynamic models cannot be reached, but a relevant simplification in

handling and in computational workload is guaranteed.

Multiperiod models are structured in the following way.
The problem is solved on a time horizon, which is divided into a set of T consecutive time intervals. The partitioning phase is based on statistical observations. In particular, the aim is to identify a set of time-clusters, such that the system data is homogeneous within each of them. This way, the system conditions can be considered as stationary in each time interval.
Solutions of multiperiod models are then time-dependent, although in a discrete sense. The level of fitting of the solution to the real system conditions depends on the number of considered time periods. The accurateness of the model can be improved by increasing the parameter T.

One of the first multiperiod models was proposed by Repede and Bernardo in 1994 [30]. The authors recognized that facing the variability of system's conditions may be a challenging problem, and proposed to solve it by considering a time-dependent approach.
On the basis of Daskin's MEXCLP (see Section 1.2) they proposed a model for the multi-period maximum expected coverage location problem (TIMEXCLP).
The objective of this model is to maximize the expected covered demand at various points in time, which are set by splitting the time horizon into a fixed number of sub-periods. The innovative element is the introduction of time-variant travel times and demand pattern. The model incorporates also the possibility of changing the fleet size in order to fit the demand in the best possible way.
The authors applied their model to the emergency system of Louisville, Kentucky, and showed that the average call response time was considerably decreased.

Another interesting model is represented by DACL, proposed by Rajagopalan et al. in 2008 [26]. The authors extended Marianov and Revelle's Q-PLSCP, introducing a multiperiod approach to the problem. In addition, they con-

sidered a probabilistic formulation based on the idea of reliability level. This task is carried out with Jarvis' hypercube approximation algorithm [21].
The objective of the model is to minimize the total number of ambulances deployed among the entire time horizon, given a parameter α representing the reliability level that has to be maintained in every time period.
This model incorporates the arduous feature of exploiting a hypercube algorithm, which is used to compute vehicles' specific busy probabilities; as it was already introduced, this kind of approach is very hard and it is sustainable only when little instances of the problem are considered. In order to find a solution for the model, a tabu-search meta-heuristic has been developed by the authors.
The binary variables of the model are defined this way:

$$x_{ik,t} = \begin{cases} 1 & \text{if server } i \text{ is located at node } k \text{ at time } t \\ 0 & \text{otherwise} \end{cases}$$

$$y_{j,t} = \begin{cases} 1 & \text{if node } j \text{ is covered by at least one server with} \\ & \alpha_t \text{ reliability at time } t \\ 0 & \text{otherwise} \end{cases}$$

The model can then be written as:

$$\text{(DACL)} \quad \text{minimize} \quad \sum_{t=1}^{T} \sum_{k=1}^{n} \sum_{i \in k} x_{ik,t} \tag{1.52}$$

$$\text{subject to} \quad \left\{ \left[1 - \prod_{i=1}^{m_t} P(x_{ij,t}) \right] - \alpha_t \right\} y_{j,t} \geq 0 \quad \forall j,t \tag{1.53}$$

$$\sum_{j=1}^{n} h_{j,t} y_{j,t} \geq c_t \quad \forall t \tag{1.54}$$

$$x_{ik,t} \in \{0,1\} \quad \forall i,k,t \tag{1.55}$$

$$y_{j,t} \in \{0,1\} \quad \forall j,t \tag{1.56}$$

where

$$P(x_{ij,t}) = p_{i,t}^{\sum_{k=1}^{n} a_{ik} x_{ij,t}} Q(x_{ij,t}) Q\left(m_t, p_t, \sum_{j=1}^{n} \sum_{i=1}^{m_t} a_{ik,t} x_{ij,t} - 1 \right)$$

is a factor accounting for ambulances busy probabilities, defined on the basis of Jarvis' approach.

The last example of multiperiod approach is represented by the 2010 work of Schmid and Doerner [32].
The authors choosed to face the location problem without considering a probabilistic formulation. In fact, they considered a double coverage approach analogous to that of Hogan and ReVelle for BACOP1 and BACOP2 models (see Section 1.1).
The innovative point in their work is the introduction of an aspect that was not considered yet: the relationship between solutions in consecutive time intervals.
In order to follow the discrete solution suggested by the multiperiod model, a certain number of changes in vehicles' locations have to be necessarily made during the time between consecutive time intervals. As it will be further discussed in the next chapter, this is a crucial matter in multiperiod location problems.
In order to handle this problem, the authors introduced a relocation cost β. This parameter is used to define a penalty term associated with relocations, which is added to the objective function. Thus, it is possible to control the quality of the solution and the number of relocations. Parameter β is set by the user: the higher this parameter is, the less ambulance relocations are tolerated by the model.
This kind of approach is without any doubt interesting because of the consideration of relocations between sub-periods; however, the parameter β has not a physical meaning, and it has to be set on the basis of experience and numerical tests.

Chapter 2

Multiperiod deterministic models

In the previous chapter a variety of models from the literature have been briefly described.
They can be classified according to several criteria. The first criterion is the one concerning the stochastic aspect of the problem: probabilistic models have been introduced as a way to improve the reliability of deterministic models. As to the time evolution of systems, another distinction has been proposed: static models versus dynamic and multiperiod ones.
In this chapter the time will be taken in consideration. In particular, a new multiperiod model will be proposed and its solutions will be examined.

2.1 The multiperiod approach

As previously mentioned, solutions given by static models present some weaknesses due to their independence with respect to time. The level of coverage they produce can turn out to be inadequate during some peak periods, or even to be overconservative for low demand periods.
Solutions obtained with static models are often exploited in real world's situations, because of their simplicity.
In every emergency system there has to be an operational center, which is

responsible for the coordination of the work of all the people involved. When ambulances are dispatched for services, it can happen that some parts of the territory are not covered enough and hence some changes in the vehicles configuration are needed. When an ambulance is moved from its current location to a new one, a *relocation* is said to take place. The decisions about relocations are usually made by the operational staff in real time, trying to mantain the best quality of service according to the available resources and the current conditions.
This kind of actions is extremely important to cover the effect of stochastic events (i.e., the service calls), but also to maintain the vehicles configuration as fitting as possible to the actual needs of the system.
After some time, it can happen that the many consecutive relocations lead to an overall drastical change in the vehicle's configuration with respect to the optimal one suggested by the model. It is clear that the improvement gained by following the optimal solution may be rapidly lost. As a consequence, after some time the location orders have to be given by the staff on the basis of experience.

An improvement in this direction is provided by dynamic models, at least from a theoretical point of view.
Altough dynamic models are, by definition, able to constantly adjust the ambulance configuration on the basis of current system's conditions, they are very difficult to be exploited in real systems. Their complexity and their huge computational load make their solving very expensive and sometimes almost impossible in practice. Moreover, systems relying on dynamic models involve much heavier logistics. In fact, all the indications given by the model have to be sent as quickly as possible to the operating vehicles. In addition, all the ambulance positions and dispatches have to be entered into the system database in real time. Such work has to be managed by a reliable and highly prepared staff. A deficiency at the executive level, for example due to the lack of staff or to the big workload during some periods of the day, can seriously affect the efficiency of the dynamic model.

Multiperiod models represent a good compromise between these two approaches.
Considering the time horizon as a sequence of homogeneous sub-periods makes it possible to introduce a time dependence into the problem, improving the quality of solutions without increasing too much the complexity as in dynamic models.
Multiperiod solutions are more accurate than static ones because they are time dependent, even if in a discrete sense. In addition, they are much simpler to use than dynamic models. Given a multiperiod solution over a set of time intervals, a preplanning phase can be made at a strategic level. Thus, it is possible to prescribe a location plan and a list of redeployments to be followed by the executive staff. Managers still have the possibility to make real-time decisions in order to cover unexpected events in the system, but at the same time they have a reliable indication about how to proceed at every moment of the day. Making their decisions, they have to maintain a configuration as close as possible to the solution given by the model.

As it was already mentioned in Section 1.4, a very important matter when working with multiperiod models is to plan a smart relocation policy.
Multiperiod solutions consist of a sequence of static solutions, hence they are discrete in time. When applying solutions to real emergency systems, a certain number of vehicle relocations have to be carried out between each time interval, in order to follow the indications given by the multiperiod model. If the number of relocations is too large, it can happen that some areas on the territory are temporarily left uncovered, or that ambulances are not able to respond to calls for long periods of time. Clearly, the benefits introduced by the multiperiod approach would be drastically affected.
Although the configuration produced by the discrete solution is optimal in each period of time, it can turn out to be not optimal *globally*. It is therefore desirable that the overall solution be as smooth as possible, avoiding too many changes between consecutive periods configurations.
To achieve this goal it is convenient to include some constraints in multiperiod models, so as to control and limit the differences between consecutive local solutions.

The multiperiod models presented in the literature review exhibit some weak points in their relocation policy.
Repede and Bernardo's TIMEXCLP [30] does not consider any constraint on the number and evolution of vehicle relocations. These aspects are not explicitely considered during the stage of optimization.
DACL model, by Rajagopalan et al. [26], suffers from the same problem of TIMEXCLP. In the model, time periods are considered to be completely independent of each other. No limitation on the number of allowed relocations is considered.

A relocation policy,hence, has to be smartly chosen and incorporated into the multiperiod model.
A first possibility is represented by the idea of including a relocation cost which penalizes consecutive and frequent relocations; this was explored by Schmid and Doerner in their 2010 work [32], as it was described in Section 1.4.
Another way to approach the question is to limit the number of relocations to be executed between each time interval. This is the choice pursued in this work.
Alternatively, the total number of relocations over the entire time horizon could be explicitely constrained.

2.2 A new model

In this section a new multiperiod model is presented.
The objective is to overcome the problems of the models present in literature. In particular, we want to formulate a model giving ambulance configurations that do not change much between consecutive time periods.
In order to simplify the notation and to better illustrate the choices made, only a deterministic version of the model is considered. In Chapter 3 a new approach will be introduced, which lead to probabilistic version of the model.
One of the first issues when approaching an emergency vehicle location problem is defining a reliability criterion to be satisfied by the solution. In this

deterministic case a Set Covering approach is considered: the aim is to cover all the demand zones with at least one ambulance. This criterion will be translated into a set of constraints: every demand zone i must have at least an ambulance located within its neighbourhood J_i.
The objective of the model is to minimize the total number of deployed ambulances over the entire time horizon, while guaranteeing that each demand zone is covered by at least one vehicle.

Given a set of T consecutive time periods and the relative reachability matrixes a_{ij}^t, defined as in Section , it is possible to begin building the new model.
The first binary variables needed are those relative to ambulance locations:

$$x_j^t = \begin{cases} 1 & \text{if an ambulance is located in } j \text{ during period } t \\ 0 & \text{otherwise} \end{cases}$$

The first coverage constraint can then be written as:

$$\sum_j a_{ij}^t x_j^t = \sum_{j \in J_i^t} x_j^t \geq 1, \quad \forall i, t \tag{2.1}$$

where J_i^t denotes the set of location points that can reach demand point i during time interval t.
Since the new model has to take into account the ambulance relocations occurring between consecutive time-intervals, two new sets of binary variables have to be introduced: $z_{IN,j}^t$ and $z_{OUT,j}^t$.
$z_{IN,j}^t = 1$ if an ambulance is assigned to a site j which was empty in the previous time interval $t-1$. $z_{OUT,j}^t$ has the opposite meaning: $z_{OUT,j}^t = 1$ if an ambulance is removed from site j at the beginning of t. It has to be underlined that both these variables refer to the beginning of the corresponding time-period t.
Variables x_j^t and $z_{IN,j}^t$,$z_{OUT,t}^t$ need to be linked. In particular, a balance between vehicles entering and exiting from each point has to be imposed. For each location point j and for each time interval t, the following relation must hold:

$$x_j^t = x_j^{t-1} + z_{IN,j}^t - z_{OUT,j}^t \tag{2.2}$$

(2.2) will be included in the model as a constraint.

The meaning of *relocation* has now to be exactly defined.
In this work, a relocation takes place at the beginning of time t when $z^t_{IN,j} = 1$ for any location point j.
This way of modeling a relocation has actually two meanings. In fact, a positive value of variable $z^t_{IN,j}$ indicates that an ambulance has moved into node j at the beginning of time t, but do not include any information about where the ambulance comes from. Thus, a relocation takes place also when a vehicle which was not previously working is assigned to an open location.
Since one of the aims of the model is to limit the differences between consecutive ambulance configurations, this choice does make sense: a new ambulance entering the system is without any doubt a relevant change from the previous configuration. Therefore, this event has necessarily to be considered as a relocation.
On the contrary, an ambulance which is removed from the system at the end of a time period must not be considered as a relevant change. If it is no longer needed, it should not affect the behaviour of the system in the future. The sense of this sentence is that we want to avoid situations in which unnecessary ambulances are forced to extend their service just because the maximum number of relocations has already been reached.
Then, the variable $z^t_{OUT,j}$ is not considered in the relocations count.

The relocation policy has now to be defined, in order to write the multiperiod model.
The solution chosed in this work is to explicitely limit the number of relocations, allowing at most M of them for each time period:

$$\sum_j z^t_{IN,j} \leq M, \, \forall t > 1 \tag{2.3}$$

We propose the following integer programming multiperiod model:

$$\text{minimize} \quad \sum_t \sum_j x_j^t \tag{2.4}$$

$$\text{subject to} \quad \sum_j a_{ij}^t x_j^t = \sum_{j \in J_i^t} x_j^t \geq 1 \qquad \forall i, t \tag{2.5}$$

$$\sum_j z_{IN,j}^t \leq M \qquad \forall t > 1 \tag{2.6}$$

$$x_j^t = x_j^{t-1} + z_{IN,j}^t - z_{OUT,t}^t \qquad \forall j, t \tag{2.7}$$

$$x_j^t \in \{0, 1\} \qquad \forall j, t \tag{2.8}$$

$$z_{IN,j}^t \in \{0, 1\} \qquad \forall j, t \tag{2.9}$$

$$z_{OUT,j}^t \in \{0, 1\} \qquad \forall j, t \tag{2.10}$$

The objective function (2.4) which is minimized corresponds to the total number of deployed ambulances, over the entire time horizon $t = 1, ..., T$. Constraints (2.5) ensure that all demand nodes are covered by at least one ambulance. The balance between entering and exiting ambulances for each node is controlled by constraints (2.7). The upper bound on the number of relocations is set by constraints (2.6) for each time period t.

2.2.1 Alternative relocation constraint

Note that this model is not the only way to formulate the multiperiod problem. Many variants of the model can be considered, on the basis of the executive manager needs.

For example, the relocation constraint (2.6) can be slightly modified, so as to limit the total number of relocations to be executed during the entire time

horizon. This way, more freedom is left to manage the timing of relocations.

$$\text{minimize} \quad \sum_t \sum_j x_j^t \tag{2.11}$$

$$\text{subject to} \quad \sum_j a_{ij}^t x_j^t = \sum_{j \in J_i^t} x_j^t \geq 1 \qquad \forall i, t \tag{2.12}$$

$$\sum_t \sum_j z_{IN,j}^t \leq M_{TOT} \qquad \forall t > 1 \tag{2.13}$$

$$x_j^t = x_j^{t-1} + z_{IN,j}^t - z_{OUT,t}^t \qquad \forall j, t \tag{2.14}$$

$$x_j^t \in \{0, 1\} \qquad \forall j, t \tag{2.15}$$

$$z_{IN,j}^t \in \{0, 1\} \qquad \forall j, t \tag{2.16}$$

$$z_{OUT,j}^t \in \{0, 1\} \qquad \forall j, t \tag{2.17}$$

In this new version of the model, constraints (2.13) impose an upper bound on the total number of relocations.
Note, however, that this kind of constraint can lead to undesired solutions. In the worst case, it can happen that all the allowed relocations take place at the beginning of the same time period. This would clearly affect the smoothness of the solution.
In practice, it may be interesting to consider optimal solutions obtained with different versions of the Model (2.11)-(2.17).

2.3 Results

In this section, computational results for the previously developed model are presented.
All the solutions have been obtained using CPLEX solver on an Intel Xeon 2.8 GHz CPU with 2GB of RAM memory.

2.3.1 Single period solutions

In order to properly analyze the solutions of the new multiperiod model (2.4)-(2.10), it is interesting to first consider the results obtained with a single time interval.

As previously mentioned, when a deterministic formulation is considered, the model is similar to a Set Covering Problem. The objective is to minimize the total number of deployed ambulances while guaranteeing that all demand nodes are covered by at least one vehicle. Due to the single time period, relocation constraints can be deleted and they do not affect the model solution.

An example of a deterministic solution on a single time period ($T = 1$) is plotted in Fig. 2.1.

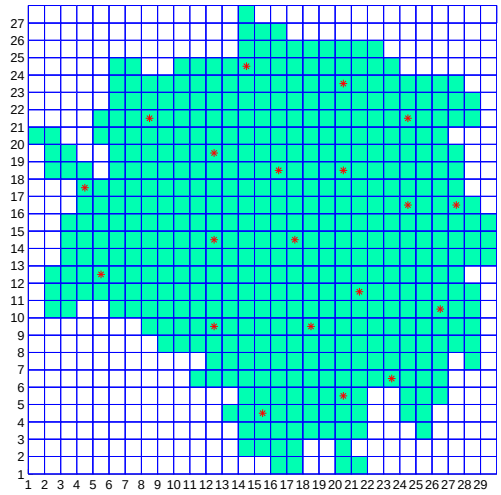
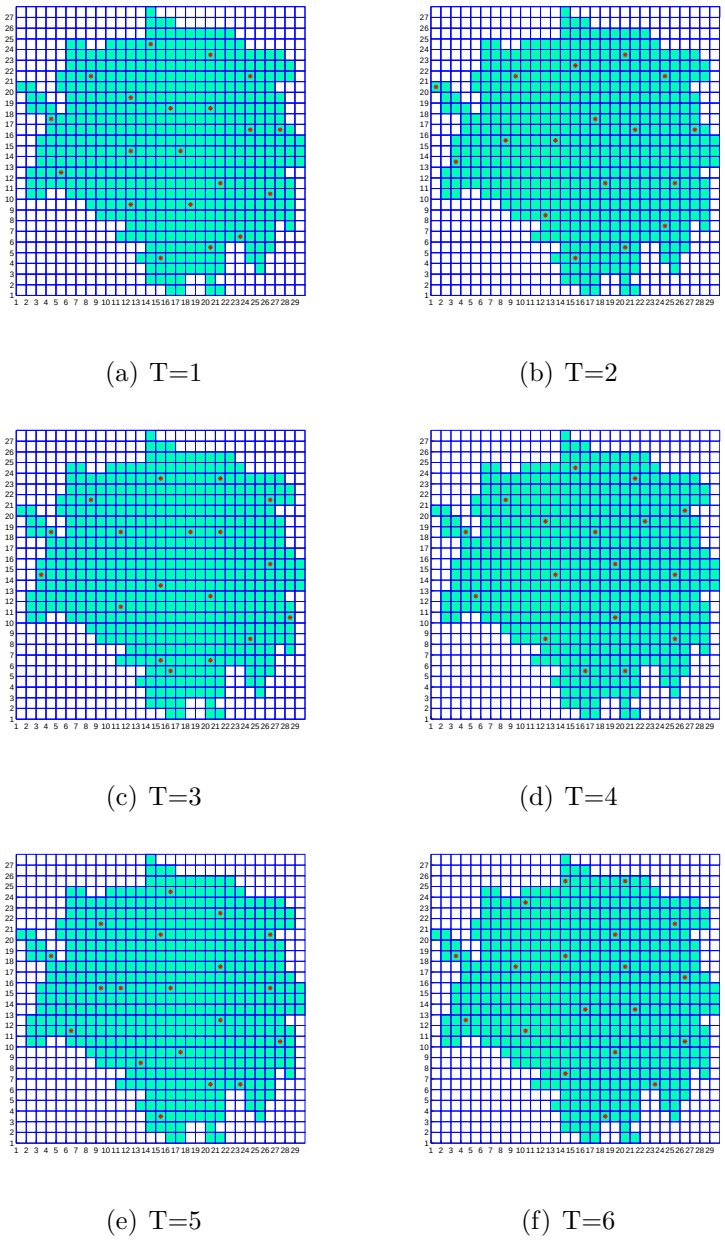


Figure 2.1: Optimal solution of deterministic problem with single time interval ($T = 1$). Deployed ambulances: 20.

Note that the ambulances, represented by a red star, are homogeneously distributed over the entire territory; each of them provide coverage with probability 1 since ambulances are considered to be always available.
This is the optimal set covering solution for time period $T = 1$. Solutions relative to other time periods are displayed in Fig. 2.2.
Clearly the ambulance configurations differ substantially among time periods. This is not a surprise since the model does not consider any relationship between consecutive intervals. The overall location plan is far from being smooth, and a high number of relocations has to be carried out in order to



(a) T=1 (b) T=2

(c) T=3 (d) T=4

(e) T=5 (f) T=6

Figure 2.2: Optimal solutions of deterministic problem with single time interval ($T = 1, 2, 3, 4, 5, 6$). Deployed ambulances: (a) 20 (b) 17 (c) 18 (d) 17 (e) 19 (f) 19.

cover the differences and move from a configuration to the next one.
As an example, consider periods $T = 2$ and $T = 3$. In the first period 17 ambulances are deployed, while in the second one 18 ambulances are needed. Altough the number of deployed ambulances does vary considerably between the two periods, their location does; many vehicles must change their position at the end of interval $T = 2$. In the lower part of the city a larger number of ambulances is needed during period $T = 3$. This requires the introduction of additional ambulances, and the relocation of at least 3 other nearby ambulances.
This is just an example of the many differences that solutions of consecutive time periods have when relocations are not considered in the model. Such solutions are *myopic*: they are accurate when focusing on a single time period, but overall they turn out not to be optimal.

We now present an overall static solution, that is a solution sharing the same deployment configuration for all the time periods; this solution must simultaneously satisfy the covering constraints of each period, considering the relative system conditions. The obtained configuration is shown in Fig.2.3.
This static location plan was developed with the intent of guaranteeing a full coverage over the entire time-horizon, without allowing any difference between solutions on consecutive time periods. This is equivalent to set $M = 0$ in the model (2.4)-(2.10).
As expected, the static approach leads to the deployment of a larger number of vehicles than in the myopic cases. For example, during the second time interval 23 ambulances are dispatched, while only 17 of them were used in the myopic solution. Such a solution is clearly overconservative.

2.3.2 Multiperiod solutions

After considering single-period solutions of the deterministic model, we now analyze multiperiod ones.
First, it must be reminded that the model is still a Set Covering Problem

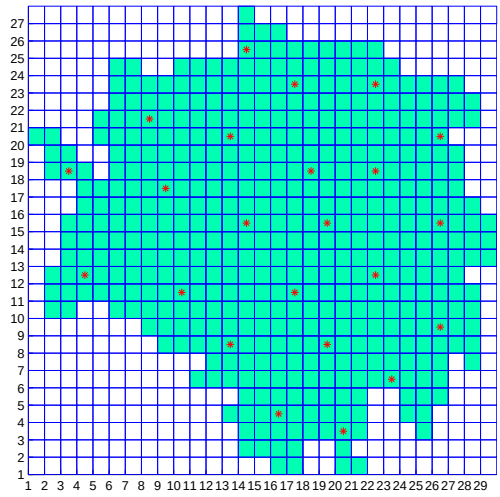


Figure 2.3: Optimal solution of deterministic problem over the entire time horizon ($T = 1..6$). No relocations allowed. Deployed ambulances: 23.

but with constraints on the maximum number of relocations allowed between consecutive periods. For this reason, smoother solutions are expected.
In Fig. 2.4 the location plan for two consecutive time intervals is plotted. The ambulances deployed in time period $T = 1$ are represented by red stars, while blue circles are used for $T = 2$. This kind of plot is useful to visualize the solutions and their differences.
Note that a different number of ambulances are deployed; in particular 20 ambulances are located during period $T = 1$ and 17 during period $T = 2$. The trend noticed in the case of myopic solutions is confirmed: the first time period is harder to manage and requires a bigger fleet of ambulances.
The very important difference with respect to the myopic case is that almost all ambulances have a static position among the two consecutive periods: although the system conditions vary with time, ambulances don't have to be relocated to guarantee the complete coverage of territory.
Some of the ambulances present in the first configuration are no longer needed during the second one. Therefore, they can end their working shift and return to the headquarters, if necessary. Since the removal of an ambulance is not

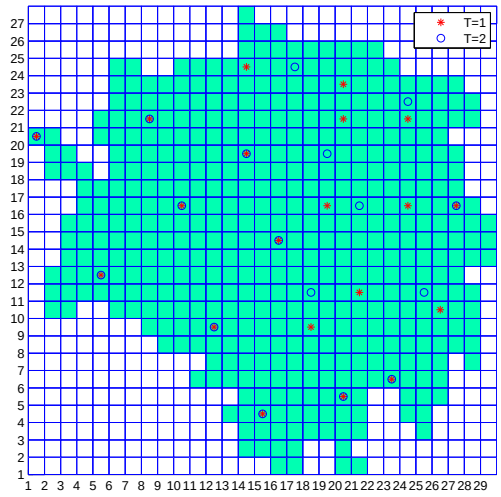


Figure 2.4: Optimal solution of multiperiod deterministic problem, two intervals. Deployed ambulances: (a) 20 (b) 17. Relocations: (b) 6.

considered as a remarkable change in the vehicle configuration, this kind of actions does not affects the smoothness of the solution.
The maximum number of relocations admitted between each time period was set to $M = 6$. Observing the picture, it is clear that the relocation constraint is respected: 6 ambulances change their location to a new one. For example, the ambulance initially located in $(26, 10)$ is relocated to $(25, 11)$ during the second time period.

It is interesting to evaluate the effect of a change in the relocation constraint parameter M. The ambulance configuration obtained with $M = 2$ is represented in Fig. 2.5.
In this case, a lower number of relocations takes place at the beginning of $T = 2$; in particular, the model spends both the two allowed relocations.
This is without any doubt an improvement with respect to the case of 6 relocations; however, note that this gain can only be obtained by increasing the value of the objective function: in the first time period 21 ambulances are deployed.

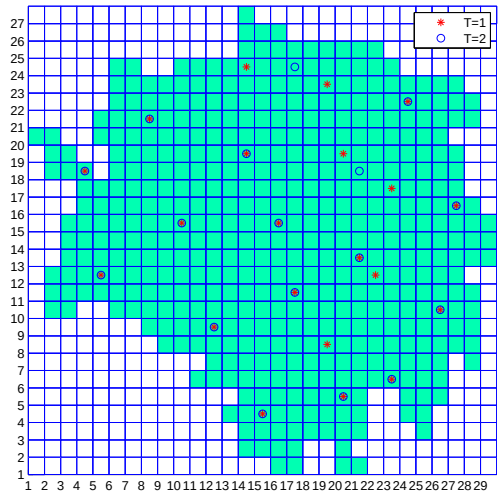


Figure 2.5: Optimal solution of multiperiod deterministic problem, two intervals. Deployed ambulances: (a) 21 (b) 17. Relocations: (b) 2.

As in the previous case, a certain number of vehicles are removed after the end of $T = 1$.

The previously exposed solutions, relative to the single and double period model, were solved exactly by CPLEX within acceptable computational times (about 1 second for one period, about 400 seconds for two periods); when more than two periods are considered, the problem requires a larger time to be solved to optimum.
The solution of the problem with 3 periods was obtained by means of CPLEX solver, after setting a maximum solve time of 6000 CPU seconds.
The obtained configuration is represented in Fig.2.6.
Since the problem is not solved to optimum, a bound on the solution is given by CPLEX: the value of the objective function, that is the total number of deployed ambulances, is 60, with a bound of 4.23 (about 7% of objective function).
Many ambulances keep their position throughout the entire time horizon. A certain number of ambulances, for example those located in $(19, 17)$ or

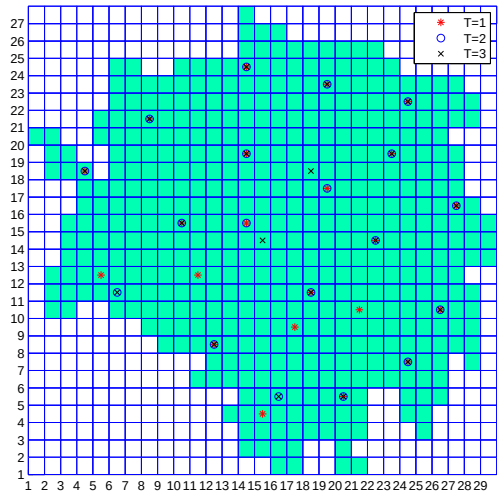


Figure 2.6: Sub-optimal solution of multiperiod deterministic problem, three intervals. Deployed ambulances: (a) 22 (b) 19 (c) 19. Relocations: (b) 2 (c) 2.(M=2)

$(14, 15)$, are relocated only at the beginning of the third time period, after standing for two periods in the same location. Two ambulances, on the contrary, move just at the end of the first period and then maintain their position during the third one.

Considered the large amount of time needed to solve the model, we propose two smaller instances of the problem. Solutions shown in the next sections will often refer to those problems.
Smaller instances are defined considering two subsets of the original sets I and J, respectively for demand and location nodes. In particular, the western part of the city has been selected. We propose a *small* instance (100 potential location nodes) and a *medium* instance (200 potential location nodes). When we consider the whole territory of Milan (492 potential location nodes), we talk about *large* instance.
Solutions relative to a two-periods time horizon are plotted in Fig. 2.7 and Fig. 2.8. In both cases a unique relocation is allowed between periods: $M = 1$.

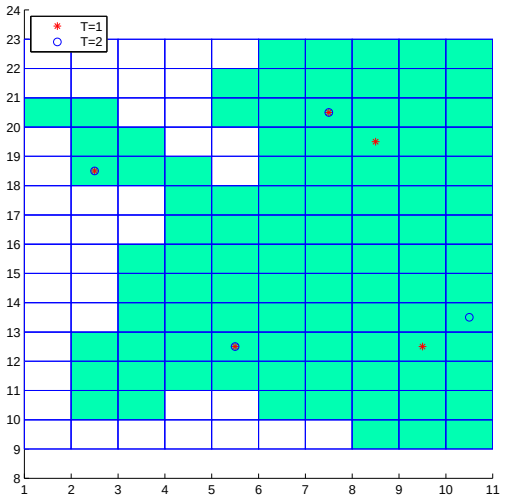


Figure 2.7: Optimal solution of multiperiod deterministic problem, two intervals, small instance. Deployed ambulances: (a) 5 (b) 4. Relocations: (b) 1. ($M = 1$)

In the first case, one ambulance is removed at the beginning of the second period. Thanks to the easier traffic conditions in period $T = 2$, the redeployment of the ambulance in $(9, 12)$ is enough to create a full covering configuration.
The easier conditions of period $T = 2$ are confirmed by the medium instance solution: two out of seven ambulances are no longer needed, thanks to the relocation of a single ambulance to an adiacent location node.

The solution obtained after increasing the number of periods to the value of 3 is reported in Fig.2.9. Also in this case, two ambulances stop their service after the end of period $T = 1$. Five ambulances maintain their position through the entire time horizon, while two of them are relocated to a nearby waiting site.

The optimal solution of the medium instance of the problem with 6 time-

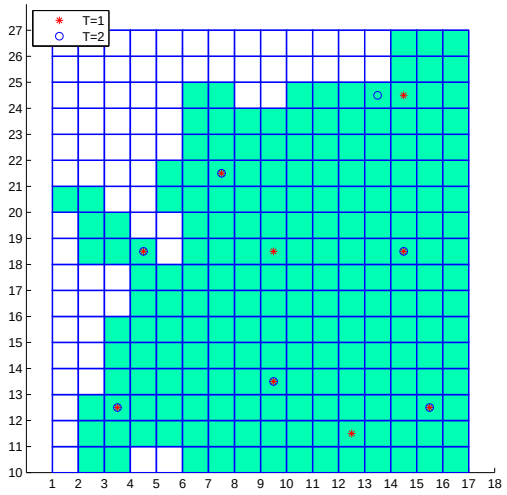


Figure 2.8: Optimal solution of multiperiod deterministic problem, two intervals, medium instance. Deployed ambulances: (a) 9 (b) 7. Relocations: (b) 1. ($M = 1$)

intervals is plotted in Fig.2.10.
A comparison between the initial and the final period is proposed in Fig.2.11.

In both $T = 1$ and $T = 6$, nine ambulances are deployed. 6 of them are stationary throughout all the periods. The remaining 3 ambulances are subject to some changes during the time. In particular, one of them is removed at the beginning of $T = 2$ and reintroduced in the system during $T = 4$. All their relocations took place between neighbourhood nodes.

SI PUO TOGLIERE FORSE:
We now present the 6 periods solution for the small instance of the problem. The differences between the first and the last configurations are appreciable in Fig.2.12.

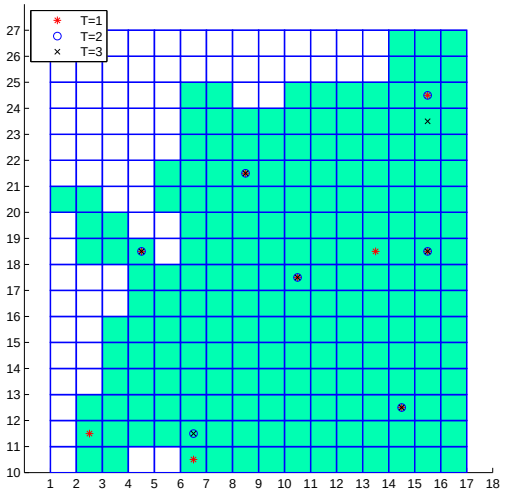
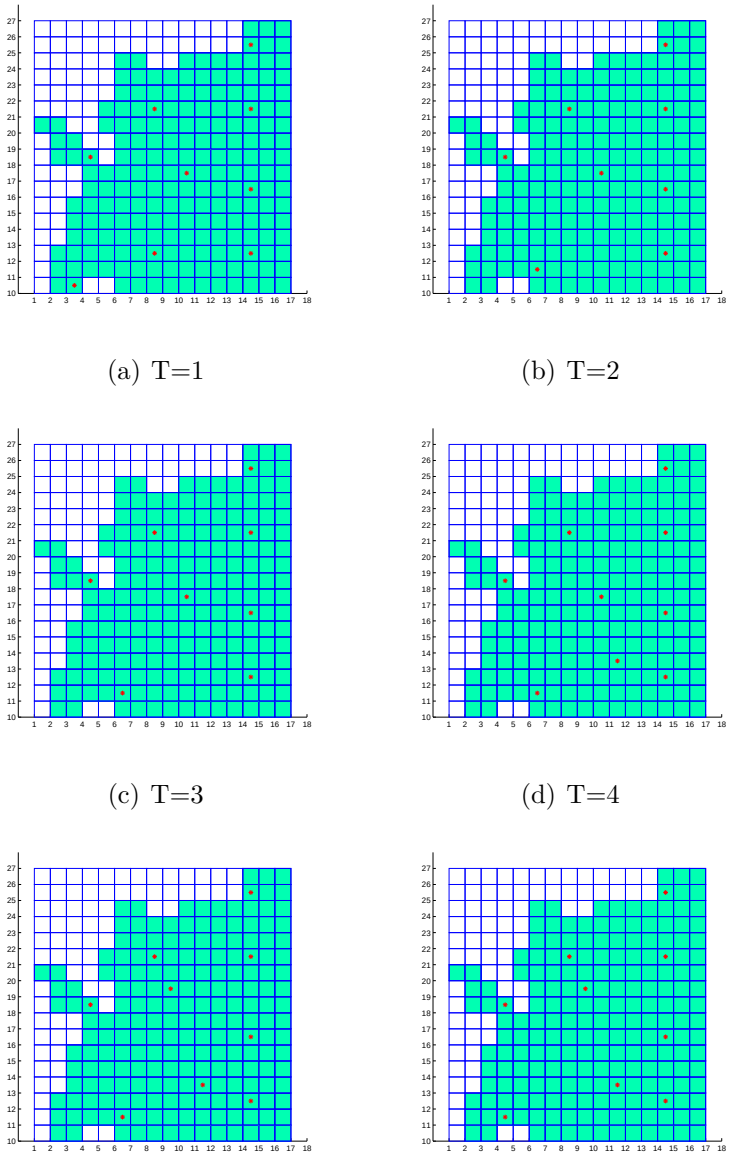


Figure 2.9: Optimal solution of multiperiod deterministic problem, three intervals, medium instance. Deployed ambulances: (a) 9 (b) 7 (c) 7. Relocations: (b) 1 (c) 1. ($M = 1$)

2.3.3 Alternative relocation constraint

We now test the model (2.11)-(2.17), that is the alternative version of (2.4)-(2.10). It was formulated using the maximum total relocations constraint.
The expected result is to obtain better solutions in terms of total deployed ambulances, since the number of relocations is not limited in every time period. However, we also take into account a potential worsening in solution smoothness.
The first presented solution is relative to the small instance of the problem, with a 6 intervals time horizon. The initial and the final configurations are plotted in Fig.2.13.
The total number of deployed ambulances over the six periods is 26. The guess on lower total number of ambulances is confirmed, since 27 ambulances were deployed in the case of a single relocation for each time period (Fig.2.12). Also the second expected consequence is confirmed, though: all the 5 allowed relocations take place during the last time periods. In this case,



(a) T=1 (b) T=2

(c) T=3 (d) T=4

(e) T=5 (f) T=6

Figure 2.10: Optimal solution of multiperiod deterministic problem, six time intervals, medium instance. Deployed ambulances: (a) 9 (b) 8 (c) 8 (d) 9 (e) 9 (f) 9. Relocations: (b) 1 (c) 0 (d) 1 (e) 1 (f) 1. ($M = 1$)

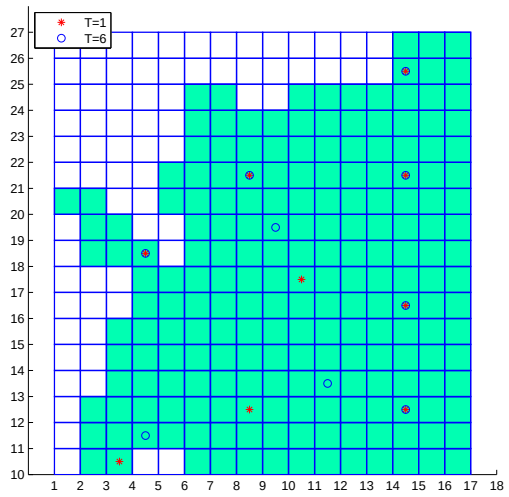


Figure 2.11: Optimal solution of multiperiod deterministic problem, six time intervals, medium instance. Deployed ambulances: (a) 9 (b) 9. Sequence of relocations: (1-0-1-1-1). ($M = 1$)

however, solution smoothness is not much worse.
The solution of medium instance is showed in Fig.2.14.
Also in this case, a small improvement in the total number of deployed ambulances is achieved (52 ambulances for the standard model, 51 ambulances for the alternative model). However, at the beginning of period $T = 5$, 3 ambulances are relocated. When we solved the problem imposing a single relocation for each time period ($M = 1$), only 4 relocations took place in the whole time horizon.

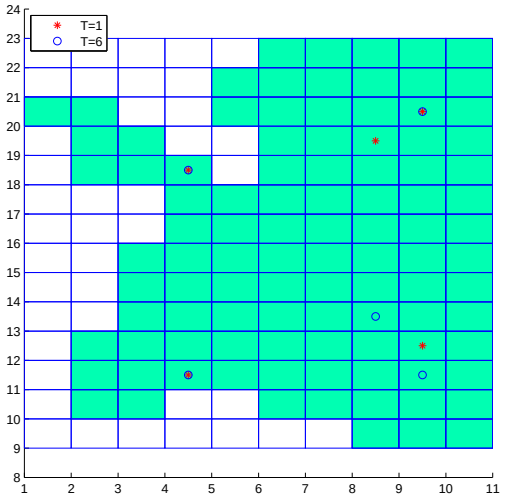


Figure 2.12: Optimal solution of multiperiod deterministic problem, six intervals, small instance. Deployed ambulances: (a) 5 (b) 5. Sequence of relocations: (1-1-1-1-1). ($M = 1$)

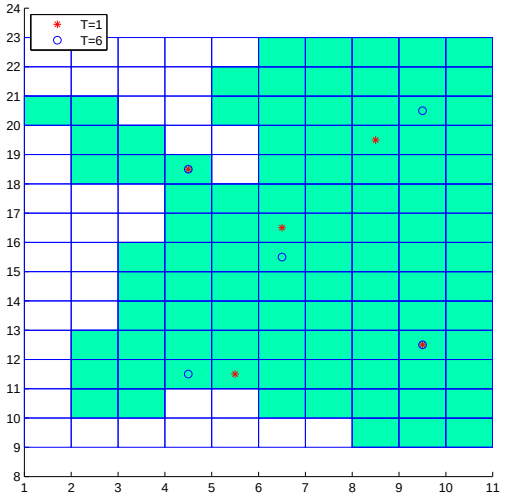


Figure 2.13: Optimal solution of multiperiod deterministic problem, six time intervals, small instance. Deployed ambulances: (a) 5 (b) 5. Sequence of relocations: (0-0-1-2-2). ($M_{TOT} = 5$)

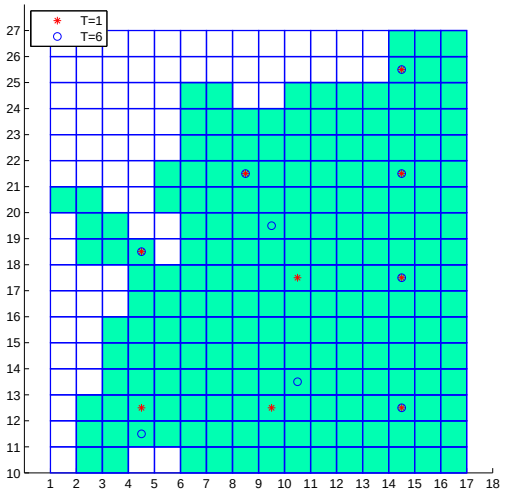


Figure 2.14: Optiumal solution of multiperiod deterministic problem, six time intervals, medium instance. Deployed ambulances: (a) 9 (b) 9. Sequence of relocations: (1-0-0-3-1). ($M_{TOT} = 5$)

Chapter 3

A Multiperiod Probabilistic Model: MPAL

In this chapter we consider the problem of finding a location plan for ambulances which takes into account the possibility of congestion in the system.
A probabilistic feature is added to the multiperiod problem presented in the previous Chapter, thanks to an alternative approach mutuated from robust optimization.
Robust optimization is a way to face the uncertainty affecting the data in a problem. Following an approach based on cutting planes, it is possible to improve the deterministic formulation of the multiperiod model and obtain its probabilistic version.

3.1 Robust optimization

One of the main issues when facing real-world optimization problems is the determination of solutions which are stable with respect to variations in the input or data. This kind of solutions is usually referred to as *robust*.
The uncertainty in the problem, caused for example by variable data, can be approached in two ways: stochastic programming or robust optimization.

Stochastic programming is based on the introduction of additional variables

into the problem and the shrinking of the feasible region, so that solutions which are most likely to become infeasible due to the uncertain data are eliminated. This way of addressing the problem leads to problems which can be very difficult to solve. Moreover, a knowledgde of how the uncertainty works is needed.
Robust optimization approach is based on the introduction of a certain number of hard constraints into the problem. It is then possible to obtain a solution which is feasible even if the worst-case conditions occur in the system. This way of modeling the uncertainty is easier to approach, but may lead to overconservative solutions which are very expensive from the point of view of costs.

In a work published in 2009 [11], Fischetti and Monaci proposed a *cutting planes* approach to manage the uncertainty underlying in robust optimization problems.
They also introduced a practical application concerning the Set Covering Problem, called Uncertain Set Covering Problem (USCP).
In the next section, USCP is presented and analyzed; it is then applied to the previously presented model in order to obtain its *uncertain* counterpart.

3.2 Uncertain Set Covering Problem

The Set Covering Problem is a famous Integer Linear Programming problem, and it has been exploited to solve a great number of practical problems.
SCP can be formulated as follows. Given two sets $I = \{1, \dots, m\}$ and $J = \{1, \dots, n\}$, let $A = (a_{ij})$ be a $m \times n$ matrix and c_j an n-dimensional integer vector.
A row $i \in I$ of A is said to be covered by a column $j \in J$ if $a_{ij} = 1$. The value c_j is called the cost of column j. Without loss of generality it can be assumed that $c_j > 0$ for all $j \in N$.
The problem consists of finding a minimum-cost subset $S \subseteq J$ of columns, such that each row $i \in I$ is covered by at least one column $j \in S$.
After the introduction of the binary variables x_j, such that $x_j = 1$ if and

only if $j \in S$, SCP model can be written as

$$\text{(SCP)} \quad \min \quad \sum_{j \in J} c_j x_j \tag{3.1}$$

$$\text{s.t.} \quad \sum_{j \in J} a_{ij} x_j = \sum_{j \in J_i} x_j \geq 1 \quad \forall i \in I \tag{3.2}$$

$$x_j \in \{0, 1\} \qquad \forall j \in J \tag{3.3}$$

where J_i is the set of columns covering row i (i.e. $J_i = \{j \in J : a_{ij} = 1\}$). The objective function (3.1) minimizes the total cost of covering the rows, while constraints (3.2) assure that all the rows are covered by at least one column.

The idea proposed in Fischetti and Monaci's work [11] is to reformulate SCP employing their cutting planes approach, in order to make it possible to handle uncertainty in data.
In particular, they consider the case in which each column $j \in J$ has a probability p_j of disappearing from the matrix A, i.e., all the coefficients a_{ij} in that column j become zero. For each row $i \in I$, a positive value $\overline{P_i}$ is introduced: this parameter is defined as the probability that row i will actually be covered by at least one of the columns selected in a certain solution.
Given the entire set of probabilities p_j, which are assumed to be independent, and the pattern of desired coverage probabilities $\overline{P_i}$, it is possible to write an uncertain version of SCP model. The objective is to minimize the costs associated with columns selection and to satisfy the i-th coverage constraint (3.2) with probability $\overline{P_i}$.
Hence, USCP model can be written as:

$$\text{(USCP)} \quad \min \quad \sum_{j \in J} c_j x_j \tag{3.4}$$

$$\text{s.t.} \quad \mathbb{P}\left\{ \sum_{j \in J_i} x_j \geq 1 \right\} > \overline{P_i} \quad \forall i \in I \tag{3.5}$$

$$x_j \in \{0, 1\} \qquad \forall j \in J \tag{3.6}$$

The authors propose to approach the uncertain problem defining as feasible a solution which satisfies a robustness criterion: each row i has to be covered at least by a set of columns $S_i \subseteq J_i$, which have a small probability of disappearing together. In particular, the probability that all the columns belonging to the set S_i be unavailable has to be less than or equal to the value $1 - \overline{P_i}$. This condition can be written as follows:

$$\sum_{j \in J_i} x_j - \sum_{j \in S_i} x_j \geq 1, \quad S_i \subseteq J_i : \mathbb{P}\{\sum_{j \in S_i} x_j\} \leq 1 - \overline{P_i}, \quad \forall i \in I \tag{3.7}$$

Since these constraints are defined using probabilities, they have to be reformulated in order to obtain a linear expression.
Recalling that the values p_j are independent of each other, the probability that a node i is not covered by any row (i.e. constraint (3.2) is violated) can be easily obtained. Defined as J_i^* the set of columns covering node i in a given solution x^* (i.e., $J_i^* = \{j \in J_i : x_j^* = 1\}$), this probability is equal to:

$$\mathbb{P}\{\sum_{j \in J_i} x_j^* < 1\} = \prod_{j \in J_i^*} p_j. \tag{3.8}$$

Then, (3.7) becomes:

$$\sum_{j \in J_i} x_j - \sum_{j \in S_i} x_j \geq 1, \quad S_i \subseteq J_i : \prod_{j \in S_i} p_j \leq 1 - \overline{P_i}, \quad \forall i \in I \tag{3.9}$$

Defining the nonnegative quantities $w_j = -\ln p_j$ and $\overline{W_i} = -\ln(1 - \overline{P_i})$ it is possible to write a linear condition:

$$\sum_{j \in J_i} x_j - \sum_{j \in S_i} x_j \geq 1, \quad S_i \subseteq J_i : \sum_{j \in S_i} w_j \leq \overline{W_i}, \quad \forall i \in I \tag{3.10}$$

This is a convenient expression for the covering constraint to be used in a

linear formulation of the USCP model:

$$(\text{USCP, M1}) \quad \min \quad \sum_{j \in J} c_j x_j \tag{3.11}$$

$$\text{s.t.} \quad \sum_{j \in J_i} x_j - \sum_{j \in S_i} x_j \geq 1, \quad S_i \subseteq J_i : \sum_{j \in S_i} w_j < \overline{W_i} \quad \forall i \in I \tag{3.12}$$

$$x_j \in \{0,1\} \qquad \forall j \in J \tag{3.13}$$

The objective function (3.11) minimizes the cost of selected columns, while constraints (3.15) guarantee that the coverage level is satisfied.
We underline again that the solution of this model is robust with respect to uncertainty in columns availability. This property is guaranteed since each row i must be covered by a set of colums having a small probability to disappear together. A subset S_i which does not satisfy constraint $\sum_{j \in S_i} w_j \geq \overline{W_i}$ is not enough for covering the demand node i. Additional columns have to be selected in order to obtain a robust solution.

Model (3.11)-(3.13) is written in a noncompact formulation. The feasibility conditions concerning the coverage of rows lead to an exponential number of constraints.
It can be proved that the feasible set induced by (3.12) has an alternative expression [11]. A compact version of USCP model can then be written as follows:

$$(\text{USCP, M2}) \quad \min \quad \sum_{j \in J} c_j x_j \tag{3.14}$$

$$\text{s.t.} \quad \sum_{j \in J_i} w_j x_j \geq \overline{W_i} \quad \forall i \in I \tag{3.15}$$

$$x_j \in \{0,1\} \qquad \forall j \in J \tag{3.16}$$

In their paper, the authors showed results from many different instances of USCP, claiming that (USCP,M2) is much easier to be solved than (USCP,M1).

In some cases, however, solving USCP model in its (USCP,M2) formulation can turn out to be challenging for MIP solvers. In fact, due to the definition of parameters w_j, coverage constraints (3.15) typically lead to nasty knapsack conditions.
In order to overcome this problem, w_j can be redefined. The covering constraints (3.15) can be reformulated as well, by exploiting the integrality of x_j.
The authors propose to replace w_j with $\overline{w}_j = \min\{w_j, \overline{W_i}\}$. In addition, they rewrite the nasty coverage constraints (3.15) in an alternative formulation, so as to make them numerically more stable. The following inequality is considered:

$$\sum_{j \in J_i} \lceil \frac{k-1}{\overline{W_i} - \epsilon} \overline{w}_j \rceil x_j \geq \lceil \frac{k-1}{\overline{W_i} - \epsilon} \overline{W_i} \rceil = k, \tag{3.17}$$

where $k \geq 2$ is an integer parameter, and ϵ is a small positive value.
This expression is similar to that used to derive Gomory's fractional cuts. It has the capacity of narrowing the feasible region without eliminating any integer solution. Thus, (3.17) can be used to replace (3.15) and strenghten (USCP,M2) formulation.
We tested the effect of covering constraints (3.17) by solving to optimality different instances of the problem. The strenghtened formulation of the problem always led to a substantial improvement in terms of computing time. In some cases, a 20% gain was obtained when considering the above formulation.

3.3 Adaptation of USCP to the ambulance location problem: MPAL model

There are strong analogies between the hypothesis of USCP and those of emergency vehicle location problem.

As already mentioned in Section 1.1, SCP was exploited by Toregas to solve the ambulance location problem. Given a set I of demand points and a set J of available location points, the aim of LSCP is to find a minimum cost set $S \subset J$ of ambulance locations such that all the demand points are covered.

In USCP problem, the set of available location points J is represented by the set of columns. The demand points set I is analogous to that of USCP rows. Since our location problem aims at minimizing the total number of deployed ambulances without any preference between location sites, the costs associated with each column $j \in J$ is unitary.

In the probabilistic version of the ambulance location problem, the main issue is the introduction of the possibility that an ambulance is not available to respond to received calls. Ambulances must not be considered as being always available: a busy probability for each vehicle has to be taken into account.
The possibility of a column j disappearing from the system is contemplated in USCP: this event has an assigned probability p_j. There is an evident analogy with the ambulance location problem: when an ambulance is not available for service, it is no longer able to cover any demand node. Thus, it can be considered as disappearing from the system.
Since there is a possibility that ambulances are not available for service, it is not reasonable to ask that all the calls are satisfied with certainty within the maximum response time r. It is then convenient to introduce a value for the probability of a demand node to be covered by at least one available ambulance. This probability is similar to the value $\overline{P_i}$ of USCP.

Starting from USCP we can formulate a probabilistic ambulance location problem. Given a set of busy probabilities p_j relative to location nodes $j \in J$, we want to find an ambulance configuration such that each demand node i is covered by an available ambulance with probability greather than or equal to $\overline{P_i}$.
Reasonable values for busy probabilities p_j can be obtained in many ways, as mentioned in Chapter 1. They can also be set by the managers of an EMS system, on the basis of historical data and direct experience.
Minimum coverage probabilities $\overline{P_i}$ can be set by the managers as well. They have to be chosen on the basis of the required level of coverage of each zone. Many factors affect this choice, for example the number of inhabitants or the

frequence of received calls.
The chance to set a heterogeneous pattern in desired coverage makes the model very flexible.

The analogies between USCP and the emergency vehicle location problem have been fully analyzed. We now apply the new probabilistic approach to the deterministic multiperiod model proposed in Chapter 2.

Since our model is defined on multiple periods, all the variables and parameters of USCP have to be adapted so as to depend explicitely on the time variable t. For example, our model includes the parameters w_j^t and $\overline{W_i^t}$. This way, variable busy probabilities and demand patterns can be considered.
The objective function (2.4) of the deterministic multiperiod model, as well as the relocation constraint (2.6) and the ambulance balance (2.7), do not need to be modified since they do not contain any probabilistic parameter.
The static USCP covering constraints (3.15) can be adapted to our multiperiod formulation:

$$\sum_{j \in J} a_{ij}^t w_j^t x_j^t = \sum_{j \in J_i^t} w_j^t x_j^t \geq \overline{W_i^t} \quad \forall i, t \tag{3.18}$$

The Multiperiod Probabilistic Ambulance Location model (MPAL) is formulated as follows:

$$\text{(MPAL)} \quad \text{minimize} \quad \sum_t \sum_j x_j^t \tag{3.19}$$

$$\text{subject to} \quad \sum_{j \in J} a_{ij}^t w_j^t x_j^t = \sum_{j \in J_i^t} w_j^t x_j^t \geq \overline{W_i^t} \quad \forall i, t \tag{3.20}$$

$$\sum_j z_{IN,j}^t \leq M \quad \forall t \tag{3.21}$$

$$x_j^t = x_j^{t-1} + z_{IN,j}^t - z_{OUT,j}^t \quad \forall j, t \tag{3.22}$$

$$x_j^t \in \{0, 1\} \quad \forall j, t \tag{3.23}$$

$$z_{IN,j}^t \in \{0, 1\} \quad \forall j, t \tag{3.24}$$

$$z_{OUT,j}^t \in \{0, 1\} \quad \forall j, t \tag{3.25}$$

Note that the demand pattern, altough not explicitely considered in the formulation, is implicitely included into the model data. In fact, it is exploited to evaluate the required reliability level of each zone. A zone with a high probability of receiving a call might be covered with a higher $\overline{W_i^t}$. In the same way, we can decrease the required reliability level in zones with a small amount of requests, without obtaining a worse quality of service.

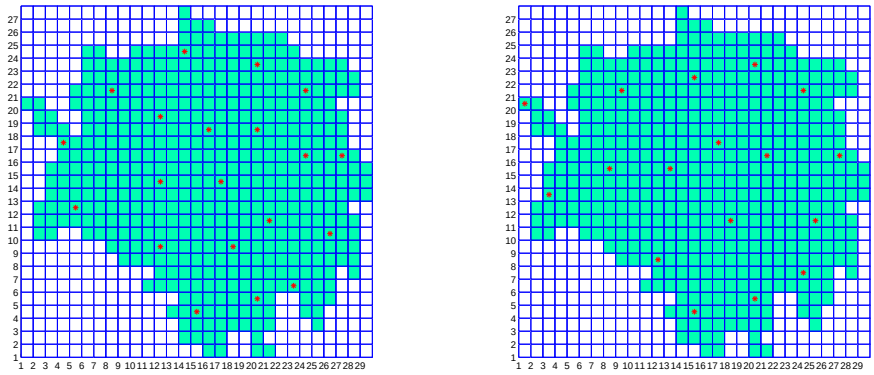
3.4 Results

In this section, we present computational results for MPAL model.
All the solutions have been obtained using CPLEX solver on an Intel Xeon 2.8 GHz CPU with 2GB of RAM memory.

3.4.1 Single period solutions

We start by analyzing solutions obtained with a single time period. This way, we can properly evaluate the effect caused by the introduction of the probabilistic feature.

Since the values p_j introduced in MPAL represent the probabilities that an ambulance located in j is not available to respond to calls, we expect that a null value for p_j lead to results analogous to those obtained in the deterministic case. In fact, the hypothesis of the deterministic model is that ambulances are always available for services, i.e. their busy probability is zero.
In order to verify this property, we compare MPAL solutions with those of the deterministic model. MPAL formulation does not admit null values for busy probabilities, because of the definition of parameters w_j, which contains a logarithm: $w_j = -\ln p_j$. Anyway, we can choose small values for p_j, so as to have $p_j \simeq 0$.
In Fig. 3.1 we show two solutions of MPAL with single time interval ($T = 1, 2$).
Comparing these solutions with their deterministic counterparts (Fig. 2.2), it is immediate to see that they coincide. The first results of probabilistic

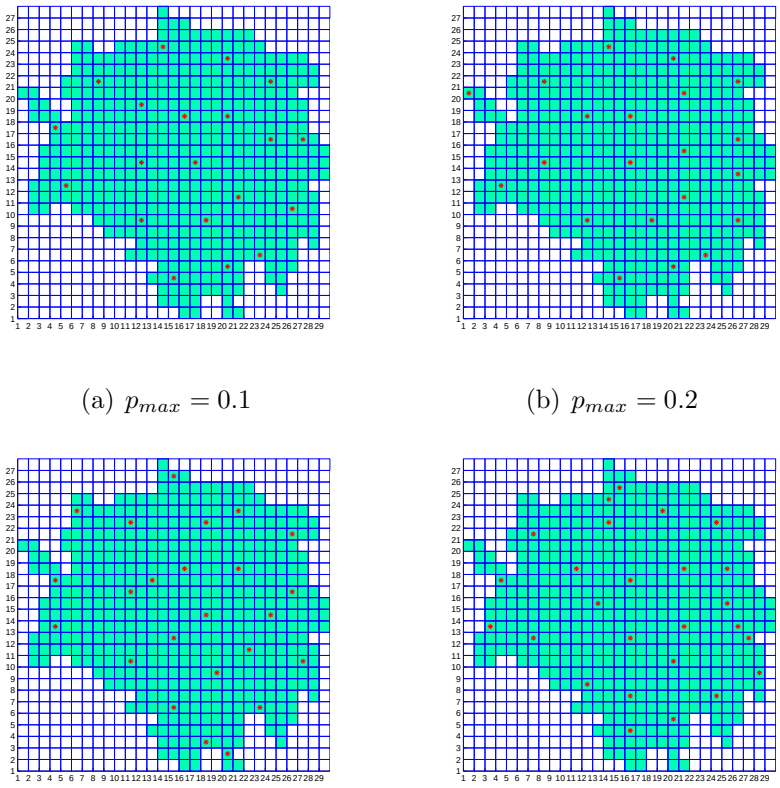


(a) T=1 (b) T=2

Figure 3.1: Optimal solutions of MPAL, with single time interval ($T = 1, 2$). $p_j \simeq 0$, $\overline{P_i} = 0.85$. Deployed ambulances: (a) 20 (b) 17.

formulation are in agreement with the expected model behaviour.
The previous results were obtained considering a coverage probability $\overline{P_i} = 0.85$ for all demand nodes $i \in I$. Since ambulance are always available, the coverage probability of nodes is actually 1. Therefore, the reliability value $\overline{P_i}$ is not relevant in the case of null busy probabilities.

Increasing the value of p_j, we can observe the effect produced by the introduction of ambulances unavailability. In Fig. 3.2 four solutions of single-period MPAL are displayed. Each of them was obtained using a sequence of busy probabilities randomly generated from a uniform distribution: $p_j \sim U(0, p_{max})$, with different values of p_{max}. The coverage probability $\overline{P_i}$ was set in all cases to the value of $\overline{P_i} = 0.85$.
Observe that increasing the value of busy probabilities leads to the deployment of a higher number of ambulances. In particular, when $p_{max} = 0.6$, 26 ambulances are deployed. In the deterministic case, only 20 vehicles were necessary.
This kind of results are reasonable, since a backup coverage is clearly needed when ambulance unavailability is considered. This is a good feedback about the good performance of the model.



(a) $p_{max} = 0.1$ (b) $p_{max} = 0.2$

(c) $p_{max} = 0.4$ (d) $p_{max} = 0.6$

Figure 3.2: Optimal solutions of MPAL, with single time interval ($T = 1$). Different values of p_{max}, $\overline{P_i} = 0.85$. Deployed ambulances: (a) 20, (b) 21, (c) 24, (d) 26.

In Table 3.1 we report results relative to small and medium instances of the problem (respectively 100 and 200 potential location zones). The optimal value of the objective function, i.e. the minimum number of needed ambulances, is denoted as $z = \sum_t \sum_j x_j^t$.

Also in this case, observe the effect of the increasing of busy probabilities. 5 ambulances are needed to cover the entire territory when we consider 100 potential location zones and $p_j \simeq 0.7$ ambulances are needed in the worst case we propose ($p_{max} = 0.9$).

The same situation occurs with the medium size instance. The number of

(a) Small instance

Nodes	p_{max}	$\overline{P_i}$	z
100	0.001	0.85	5
100	0.1	0.85	5
100	0.2	0.85	5
100	0.3	0.85	5
100	0.4	0.85	5
100	0.5	0.85	5
100	0.6	0.85	6
100	0.7	0.85	6
100	0.8	0.85	7
100	0.9	0.85	7

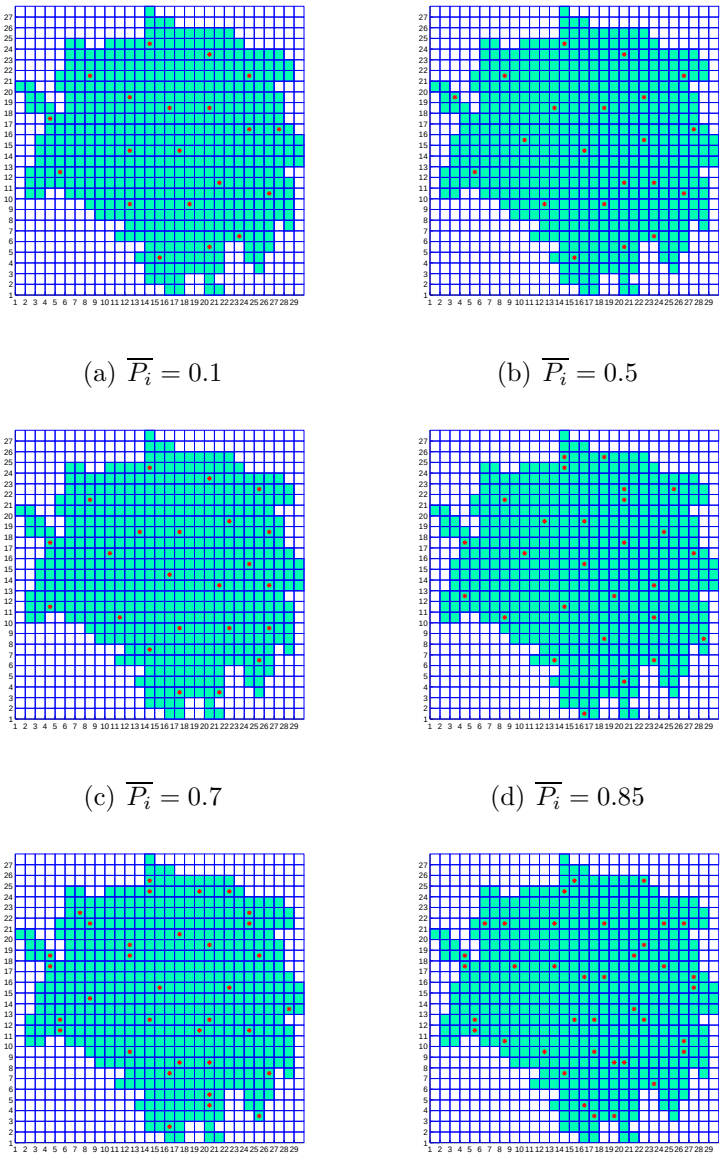
(b) Medium instance

Nodes	p_{max}	$\overline{P_i}$	z
200	0.001	0.85	8
200	0.1	0.85	8
200	0.2	0.85	8
200	0.3	0.85	9
200	0.4	0.85	9
200	0.5	0.85	9
200	0.6	0.85	10
200	0.7	0.85	10
200	0.8	0.85	11
200	0.9	0.85	12

Table 3.1: Sentitivity of solution with respect to busy probability. Small and medium instances of the problem. $T = 1$.

deployed ambulances shifts from 8 to 12 when p_{max} increases to $p_{max} = 0.9$. Results are still in perfect agreement with the expected behaviour of the model.

Another relevant question is the analysis of the sensitivity of solutions with respect to coverage probability. In order to evaluate the impact of this parameter on MPAL solutions, we fixed the value of p_{max} to $p_{max} = 0.7$, and we considered different values of $\overline{P_i}$ in the interval $[0.1, 0.99]$. The obtained results are plotted in Fig. 3.3.
Note that, when the reliability level is very low ($\overline{P_i} \leq 0.5$), MPAL solution is almost equivalent to that of deterministic case. During the first time period 20 ambulances are deployed. This is the same result as in the case of $p_{max} \simeq 0$. When $\overline{P_i} = 0.5$, only one additional ambulance is added.
When we choose values greater than $\overline{P_i} = 0.5$, we observe a considerable increase in the number of deployed ambulances. In Fig. 3.3(f) (referring to the case with $\overline{P_i} = 0.99$), 38 ambulances are needed to obtain the prescribed coverage.



(a) $\overline{P_i} = 0.1$ (b) $\overline{P_i} = 0.5$

(c) $\overline{P_i} = 0.7$ (d) $\overline{P_i} = 0.85$

(e) $\overline{P_i} = 0.95$ (f) $\overline{P_i} = 0.99$

Figure 3.3: Optimal solutions of MPAL, with single time interval ($T = 1$). $p_j \sim U(0, 0.7)$. Different values of $\overline{P_i}$. Deployed ambulances: (a) 20, (b) 21, (c) 23, (d) 27, (e) 34 (f) 38.

Observe that the number of ambulances is not linear in the reliability level. When large values of $\overline{P_i}$ are considered, a small increase in the reliability level is enough to require much more ambulances.
Reasonable values for $\overline{P_i}$ are in the range $\overline{P_i} \in [0.85, 0.95]$ when a big city like Milan is considered. Values higher than $\overline{P_i} = 0.9$ are difficult to be reached in rural zones.
Results relative to small and medium instances of the problem are presented in Table 3.2.

(a) Small instance

Nodes	p_{max}	$\overline{P_i}$	z
100	0.7	0.001	5
100	0.7	0.1	5
100	0.7	0.2	5
100	0.7	0.3	5
100	0.7	0.4	5
100	0.7	0.5	5
100	0.7	0.6	5
100	0.7	0.7	5
100	0.7	0.8	6
100	0.7	0.85	6
100	0.7	0.9	7
100	0.7	0.95	7
100	0.7	0.99	8

(b) Medium instance

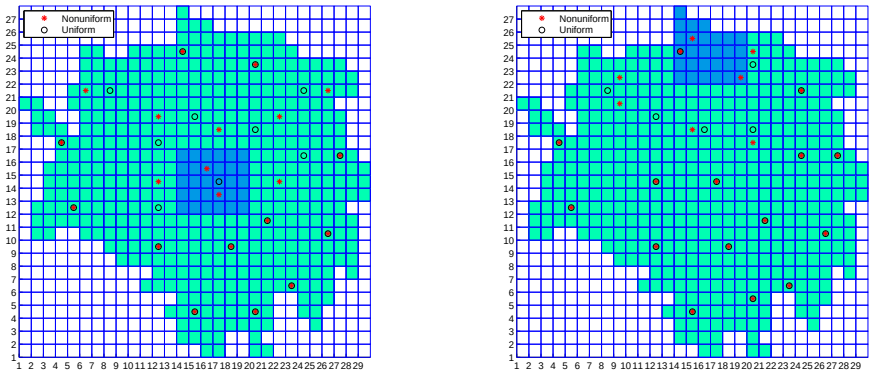
Nodes	p_{max}	$\overline{P_i}$	z
200	0.7	0.001	8
200	0.7	0.1	8
200	0.7	0.2	8
200	0.7	0.3	8
200	0.7	0.4	8
200	0.7	0.5	8
200	0.7	0.6	8
200	0.7	0.7	9
200	0.7	0.8	10
200	0.7	0.85	11
200	0.7	0.9	12
200	0.7	0.95	13
200	0.7	0.99	15

Table 3.2: Sentitivity of solution with respect to reliability value. Small and medium instances of the problem. $T = 1$.

As expected, when we require a higher reliability level, new ambulances has to be added in order to provide more backup coverage.

Until now, we always set a uniform pattern of $\overline{P_i}$ on the whole territory. One of the features of MPAL model is the possibility to define a nonuniform coverage pattern.

Thanks to this property, the flexibility of the model is increased: different levels of reliability can be required by the EMS manager for each part of the city, or even for each demand node. In some situations it can be convenient to plan a strong coverage in zones with high probability of call arrival, like the city center. On the contrary, in the periphery of the city the coverage level can be reasonably decreased because of the low population.
In Fig. 3.4 we present solutions obtained with different coverage patterns. In each case we compare the location plan for a uniform coverage pattern ($\overline{P_i} = 0.6$) with the plan induced by a nonuniform pattern, in which the reliability level of a small part of the city is set to an high value ($\overline{P_i} = 0.99$). In particular, Fig. 3.4(a) refers to the center of Milan, that is the zone of the Milan Cathedral. The zone of a big important hospital in the northern part of Milan is considered in Fig. 3.4(b).



(a) City center (b) Northern zone

Figure 3.4: Optimal solutions of MPAL, with single time interval ($T = 1$). $p_j \sim U(0, 0.7)$. Different patterns for $\overline{P_i}$: light blue = 0.6, dark blue = 0.99. Deployed ambulances: (a) Nonuniform-21 Uniform-20, (b) Nonuniform-22 Uniform-20.

In both cases, the different reliability pattern produces different solutions. Although most ambulances, expecially those further from the high-reliability zones, mantain their location when the pattern is changed, a certain number of ambulances change their position in order to cover the critical zone with

high reliability.
Note that in Fig. 3.4(a), two ambulances are located within the city center zone, while only one is positioned when a uniform pattern is considered. In Fig. 3.4(b), three ambulances are located in the high-reliability zone, with only one ambulance in the standard case.
This behaviour of MPAL was expectable. A high reliability level can be ensured only by the coverage of more than one vehicle. Since it is very probable that ambulances are busy, there is a large need of backup coverage.

3.4.2 Multi-period solutions

The effect of the probabilistic feature of MPAL has been analyzed considering single period solutions of the model. In this section we present some results concerning multiperiod solutions.
As for the deterministic model, we show solutions of small and medium instances of the problem.

The first solution we propose is obtained considering two time periods ($T = 1, 2$) and different coverage patterns. In particular, we require a high reliability level ($\overline{P_i^t} = 0.99$) in northern part of the city during $T = 1$ and in the south-eastern part during $T = 2$. In both periods, the reliability of remaining demand nodes is set to $\overline{P_i^t} = 0.6$. Busy probabilities p_j are obtained randomly: $p_j \sim U(0, p_{max})$ with $p_{max} = 0.7$. The maximum number of allowed relocations is $M = 6$.
The solution is plotted in Fig. 3.5. High reliability zones are indicated with different colours: grey for $T = 1$, dark blue for $T = 2$.
Note that, as in the single period case, the solution follows the nonuniform reliability pattern. During the first time period, two ambulances are located in the critical northern part of the city, and two additional ones are located in its neighbourhood. During $T = 2$, when the desired reliability is decreased, only one ambulance is located in this zone. There is no longer need for backup coverage.
The same behaviour occurs in the south-eastern part of the city: two ambu-

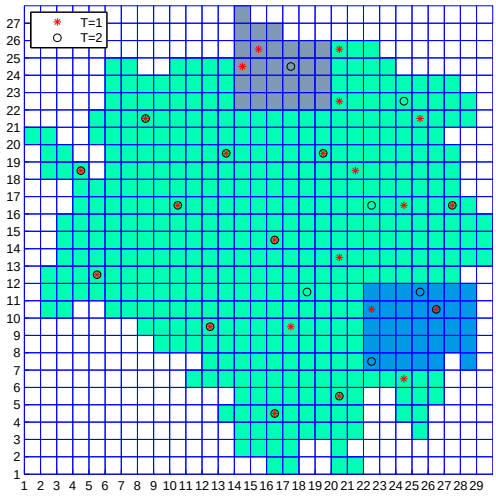
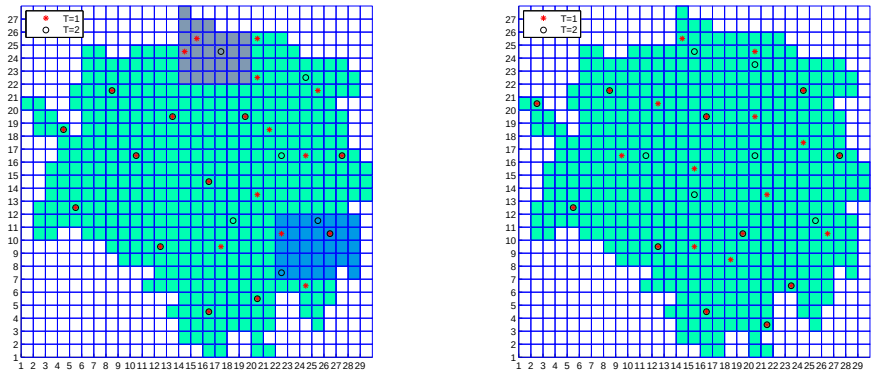


Figure 3.5: Optimal solution of MPAL, with two time intervals ($T = 1, 2$). $p_{max} = 0.7$. Different patterns for $\overline{P_i^t}$: light blue = 0.6, dark blue = 0.99, grey = 0.99.

lances are dispatched during $T = 1$, while three ambulances are dispatched during $T = 2$.
Relocations of ambulances occur mainly in the eastern part of the city, between the two high-reliability zones. No ambulance is relocated in the western part of the city.

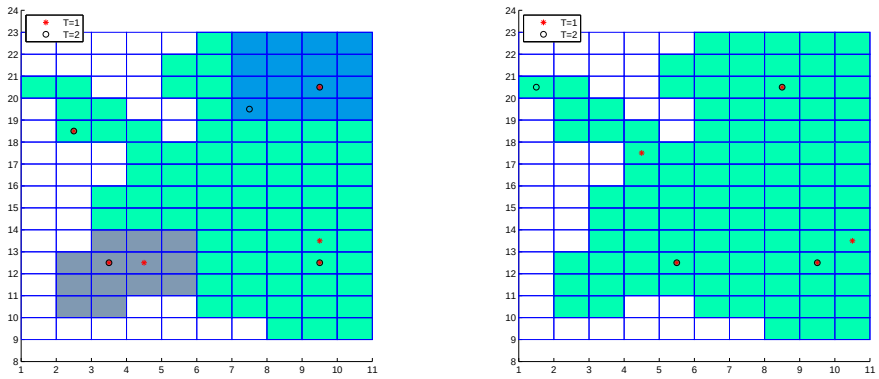
In Fig. 3.6 we show a comparison between nonuniform and uniform pattern solutions.
Note again the effect of the introduction of the nonuniform patter. In Fig. 3.6(a), ambulances are located homogeneously on the territory. Relocations occur in every part of the city.
On the contrary, in Fig. 3.6(b) all the allowed relocations are spent so as to follow the change in $\overline{P_i}$ pattern. No relocation takes place in the western part of Milan, while in the eastern part ambulances are relocated from the first critical zone to the second one.



(a) Nonuniform pattern (b) Uniform pattern

Figure 3.6: Comparison between optimal solutions of MPAL, with two time intervals and different pattern for $\overline{P_i^t}$. $p_{max} = 0.7$.

Results relative to small and medium instances are shown in Fig. 3.7 and Fig. 3.8. Two time intervals ($T = 1, 2$) were considered, with a maximum number of relocations $M = 1$.

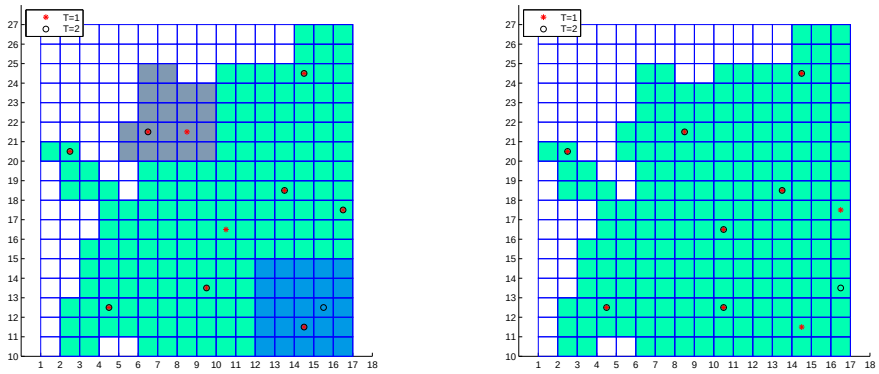


(a) Nonuniform pattern (b) Uniform pattern

Figure 3.7: Comparison between optimal solutions of MPAL, with two time intervals and different pattern for $\overline{P_i^t}$. $p_{max} = 0.7$. Small instance of the problem.

In both cases, the effect produced by the nonuniform pattern is the deploye-

ment of two ambulances to cover the high-reliability zone, while in the uniform pattern solution only one ambulance is deployed.
The unique allowed relocation is used to introduce a new ambulance in the critical zone. The other vehicles are not relocated between the first and the second period.



(a) Nonuniform pattern (b) Uniform pattern

Figure 3.8: Comparison between optimal solutions of MPAL, with two time intervals and different pattern for $\overline{P_i^t}$. $p_{max} = 0.7$. Small instance of the problem.

We now consider solutions with a time horizon of 6 periods. In Fig. 3.9 we consider a medium instance with 200 potential location points. The desired coverage pattern varies in time. In particular, a high-reliability zone ($\overline{P_i^t} = 0.99$) is set in the north-western part of the map during the first three periods ($T = 1, 2, 3$) and in the south-eastern part during the other periods ($T = 4, 5, 6$). The chosen value for p_{max} is 0.7, and the maximum number of relocations is $M = 1$.
The first high-reliability zone is represented in grey, while the second one in dark blue. Only the initial and final solutions are shown.
Note tht a double ambulance coverage is planned for high-reliability zones during critical periods. After $T = 3$, one ambulance is removed from the first zone and relocated in order to improve the reliability level of the second zone.

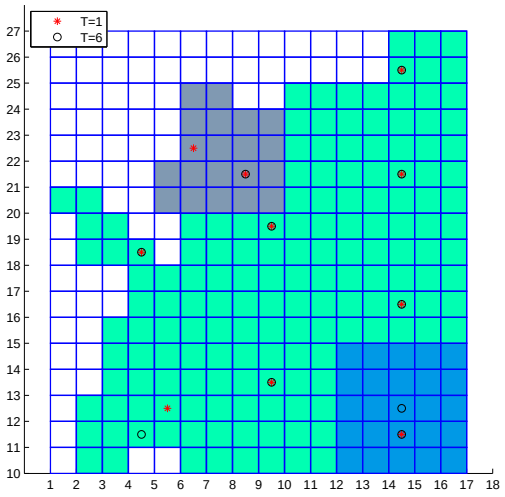


Figure 3.9: Optimal solution of MPAL, with six time intervals ($T = 1, .., 6$). $p_{max} = 0.7$. Different patterns for $\overline{P_i^t}$: light blue = 0.6, dark blue = 0.99. Medium instance of the problem.

Chapter 4

Lagrangian-based heuristic for MPAL

Since the MPAL formulation (3.19)-(3.25) is very challenging to solve to optimality with a state-of-the-art solver like CPLEX, in this chapter we present a heuristic algorithm capable of producing near-optimal solutions in a reasonable amount of time.
Our heuristic algorithm is based on a Lagrangian Relaxation (LR) approach. This is motivated by the excellent results obtained for the classical Set Covering Problems with Lagrangian-based heuristics; see, for example [5], [7] or [6].
After recalling the basic ideas of Lagrangian Relaxation, the algorithm is described and the results obtained for different instances of the problem are reported and discussed.

4.1 Lagrangian relaxation

In this section the Lagrangian Relaxation technique is outlined. A more exahustive exposition can be found in [13], [14] and [34].

Many hard integer programming problems can be viewed as *easy* problems complicated by a relatively small set of side constraints. This observation is

at the basis of the Lagrangian Relaxation method.
Consider an integer linear programming problem and suppose that its constraints can be divided into two sets of constraints, so that the problem can be written as

$$\begin{aligned} \text{(P)} \quad z_P = \quad & \min \quad c^T x && (4.1)\\ & \text{s.t.} \quad Ax \geq b && (4.2)\\ & \qquad Dx \geq e && (4.3)\\ & \qquad x \in \{0,1\}, && (4.4) \end{aligned}$$

where A and D are $m \times n$ matrixes, b and e are vectors of dimension m, c and x are vectors of dimension n.
Suppose that (4.2) represent the *difficult* constraints of the problem, while (4.3) stand for the *easy* constraints.
The Lagrangian Relaxation of problem (P) can be obtained by relaxing constraints (4.2) and introducing them into the objective function as a penalty term. The resulting problem is then

$$\begin{aligned} \text{LR}(u) \quad z_D(u) = \quad & \min \quad c^T x + u(b - Ax) && (4.5)\\ & \text{s.t.} \quad Dx \geq e && (4.6)\\ & \qquad x \in \{0,1\} && (4.7) \end{aligned}$$

where u is a nonnegative Lagrangian multiplier vector of dimension m. $z_D(u)$ is called Lagrangian function.
Given a Lagrangian multiplier vector u, the problem LR(u) is easier to solve since it contains only (4.6) constraints.
Since it is easy to verify that

$$z_P \geq z_D(u), \quad \forall u \geq 0, \tag{4.8}$$

the Lagrangian dual problem associated with (P) consists of

$$\text{(D)} \quad z_D = \max_{u \geq 0} z_D(u) \tag{4.9}$$

For some problems, $z_D = z_P$. Otherwise, a duality gap exists.

A near-optimal Lagrangian multiplier vector is usually determined by applying a *subgradient method* [34].
Recall that a vector s is called subgradient of $z_D(u)$ in $\overline{u}$ if it satisfies:

$$z_D(u) \leq z_D(\overline{u}) + s(u - \overline{u}). \tag{4.10}$$

The near-optimal multiplier vector can be obtained by means of an algorithm generating a sequence $\{u_k\} = \{u_1, u_2, ...\}$ of nonnegative Lagrangian multipliers, possibly converging to the optimal vector u^*.
The problem is now to determine an expression for the subgradient of $z_D(u)$ and to develop an algorithm that generates a good sequence $\{u_k\}$.

4.2 Adaptation to MPAL model

Starting from the original formulation of MPAL:

$$\begin{aligned}
\text{(MPAL)} \quad \text{minimize} \quad & \sum_t \sum_j x_j^t & & (4.11)\\
\text{subject to} \quad & \sum_{j \in J_i^t} w_j^t x_j^t \geq \overline{W_i^t} & \forall i,t & (4.12)\\
& \sum_j z_{IN,j}^t \leq M & \forall t & (4.13)\\
& x_j^t = x_j^{t-1} + z_{IN,j}^t - z_{OUT,j}^t & \forall j,t & (4.14)\\
& x_j^t \in \{0,1\} & \forall j,t & (4.15)\\
& z_{IN,j}^t \in \{0,1\} & \forall j,t & (4.16)\\
& z_{OUT,j}^t \in \{0,1\} & \forall j,t & (4.17)
\end{aligned}$$

and relaxing the covering constraints (4.12), we obtain

$$z_D(u) = \min \quad \sum_t \sum_j x_j^t + \sum_t \sum_i u_i^t \left(\overline{W_i^t} - \sum_{j \in J_i^t} w_j^t x_j^t \right) \tag{4.18}$$

$$\text{subject to} \quad \sum_j z_{IN,j}^t \leq M \qquad \forall t \tag{4.19}$$

$$x_j^t = x_j^{t-1} + z_{IN,j}^t - z_{OUT,j}^t \qquad \forall j,t \tag{4.20}$$

$$x_j^t \in \{0,1\} \qquad \forall j,t \tag{4.21}$$

$$z_{IN,j}^t \in \{0,1\} \qquad \forall j,t \tag{4.22}$$

$$z_{OUT,j}^t \in \{0,1\} \qquad \forall j,t \tag{4.23}$$

The constraint violations are weighted in the objective function by the non-negative Lagrangian multipliers u_i^t.

Note that the relaxed problem (4.18)-(4.23) is much easier to solve than (4.11)-(4.17). Using an efficient solver, solutions of large instances of the problem can be obtained in a very small amount of time. For example, it takes 0.34 seconds for CPLEX to solve a large istance with 492 potential location points and 6 time intervals[1].

Other possibilities have been considered before choosing the above Lagrangian Relaxation of MPAL. Dualizing the constraints (4.13) or (4.14) did not lead to easier problems. Solving a medium instance with 200 potential location points and 4 time intervals took 31.12 seconds when we relaxed the maximum relocation constraints (4.13). Negative results were also obtained when we relaxed the location points balance constraints (4.14).

This is not a surprise since typically Set Covering Problems are solved through Lagrangian Relaxation by dualizing their covering constraints. Although MPAL model has a different formulation including additional constraints and variables, the covering constraints are STILL the more problematic to satisfy.

4.2.1 Solving the Lagrangian Relaxation

As already mentioned, problem (4.18)-(4.23) can be solved to optimality in a very small amount of time by using CPLEX solver. Thus, for each iteration

[1] on our Intel Xeon 2.8 GHz CPU with 2GB of RAM memory.

of the subgradient method, we can obtain the optimal value $z_D(u)$, which is a valid Lower Bound for the primal problem (P).
Solving problem LR(u) for different values of Lagrangian multipliers u, it is possible to gradually increase the value of the Lower Bound. This provides some information on the optimal objective function value.
During the first iteration of the subgradient method, Lagrangian multipliers u_k are initialized to a null value: $u_1 = 0$. Afterwards, they are obtained through the following updating formula, defined on the basis of the subgradient matrix $s(u_k)$ [34]:

$$u_{k+1} = \max\{u_k + \lambda_k \frac{UB - z_D(u_k)}{\|s(u_k)\|^2} s(u_k), 0\}. \tag{4.24}$$

UB is an Upper Bound on the optimal solution, that is the objective value of the best feasible solution found. λ_k is a positive step-size along the subgradient direction.

The subgradient matrix $s(u_k)$ consists of an evaluation of the relaxed constraints' violations in the solution of (4.5)-(4.7). Considering constraints (4.12), a generic element $s_i^t(u)$ is defined as follows:

$$s_i^t(u) = \left(\overline{W_i^t} - \sum_{j \in J_i^t} w_j^t x_j^t \right) \tag{4.25}$$

Thanks to the subgradient matrix, it is possible to evaluate which constraints are violated by the relaxed solution and to adjust the corresponding Lagrangian multipliers.

In Fig. 4.1 we show the values $z_D(u)$ obtained during an execution of the subgradient method. We consider a medium instance with 200 location points and 4 time periods.
Note that the Lower Bound produced by the Lagrangian Relaxation changes during the iterations of the subgradient method. In particular, the obtained values follow a trend which is nearly monotone.

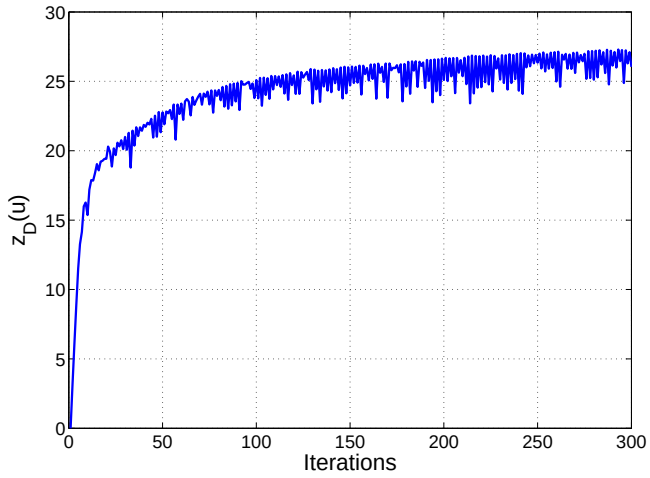


Figure 4.1: Values of the objective function of the Lagrangian Relaxation, during the iterations of the subgradient method.

Otaining a Lower Bound on the optimal value of MPAL solution is not the only intent pursued during this first phase of the algorithm. For several problems, in fact, the lower bound produced by Lagrangian or alternative relaxations is sensibly lower than the optimal solution value. This is usually the case in which a duality gap is present.
As already mentioned, the existence of this gap is not a drawback within heuristic algorithms, since the main objective is to find near-optimal feasible solutions withouth insisting in proving optimality. [6]

4.2.2 Updating the step-size parameter

The step-size parameter λ is particularly important for the good performance of the subgradient method. If a large value of λ is used, violated constraints will be heavily penalized in the objective function. On the contrary, a low value of λ implies a small correction of infeasible solutions.
Thus, the chosen values of λ strongly affects the behaviour of the algorithm. Two updating strategies for the step-size parameter λ are considered in this work. Both of them start from a value λ_1 which is set by the user, and update λ_k at every iteration.

Classical strategy

According to Held and Karp [19] the parameter λ_k has to be halved if no Lower Bound improvement occurs for p consecutive iterations. With this strategy, the algorithm detects the proximity to the optimal Lagrangian multipliers and progressively decreases λ_k in order to make smaller and more accurate steps.

Enhanced strategy

The second strategy is described in Caprara et al. [5]. We will refer to this strategy as CFT.
CFT has been showed to work well when exploited in Lagrangian-based heuristics for the classical Set Covering Problem.
The strategy consists in updating the step-size λ_k on the basis of the progresses made for the Lower Bound value during the last iterations. Every p iterations, the best and worst Lower Bounds obtained in the last p iterations are compared. If these two values differ by more than 2% the value of λ_k is halved. If, on the contrary, they are within 2% from each other the value of the step-size λ_k is multiplied by a factor 1.5.
This approach is based on the following observations. When the Lower Bound values found in the last iterations are very different (i.e., they are fluctuating or rapidly increasing), the step-size has to be reduced in order to make more accurate steps in the subgradient direction. On the contrary, a small difference in the last values indicates that the Lower Bound is not improving, perhaps due to a too small step-size. Then, parameter λ_k has to be increased in order to give the possibility to the algorithm to make a reasonable step.

4.3 Lagrangian-based heuristic: finding a feasible solution

The Lagrangian-based heuristic algorithm we developed to solve MPAL model can be outlined in three main steps.

Starting from a given Lagrangian multiplier matrix u_k, the Lagrangian Relaxation problem LR(u_k) is solved. The aim of this phase is not only to obtain a Lower Bound on the solution of primal problem, but also to drive the search of near-optimal solutions. In general, solutions of the Lagrangian Relaxation problem (4.18)-(4.23) are infeasible for primal problem (4.11)-(4.17). Solutions LR(u_k) have to be cleverly modified in order to determine a feasible solution of MPAL. This is achieved either with a greedy procedure or a refining. During the last step, Lagrangian multipliers are updated in order to execute the following iteration of the subgradient method.
The Lagrangian-based heuristic algorithm executes the above steps repeatedly, until a the best feasible solution of MPAL cannot be improved.
The overall Lagrangian heuristic algorithm can be outlined as follows:

Procedure 1 Lagrangian heuristic algorithm

$u_1 := 0$
repeat
1. Solve the Lagrangian Relaxation LR(u_k)
2. Derive a feasible solution x_k of the primal problem
3. Update the Lagrangian multipliers u_{k+1}
4. $k := k + 1$

until x_k cannot be improved
return x_k

The generation of a feasible solution of the primal problem starting from a solution of Lagrangian Relaxation problem is a delicate and important issue within Lagrangian-based heuristics. In this work we consider different methods for deriving feasible solutions. They were implemented and refined on the basis of results obtained in repeated tests.

4.3.1 Greedy approach

The first considered strategy is to derive a feasible solution by means of a greedy approach.

The Lagrangian Relaxation solution is infeasible because the dualized constraint is violated, typically because too few ambulances are deployed. Thus, additional ambulances have to be added in order to make the solution feasible.
The greedy approach we propose progressively adds ambulances, until all the covering constraints are satisfied. A similar approach has been adopted in [5] and [3] for the classical Set Covering Problem.
The greedy phase is organized as follows:

Procedure 2 Greedy phase

for $t \in \{T, T-1, ..., 1\}$ **do**

repeat

1. Choose a candidate location point j such that $x_j^t = 0$

2. Add an ambulance in j for all periods $k \leq t$ (i.e. $x_j^k = 1 \forall k \leq t$)

until all covering constraints are satisfied during period t

end for

The first step is carried out by randomly choosing a location site j. The candidate points are those which would allow to cover at least one uncovered demand point. An alternative would be to choose the location point j giving the maximum gain in terms of covered demand points. Although it would probably lead to better results, the computational workload would be considerably increased.
During the second step, an ambulance is placed in the chosen location point j, for all time periods previous to the current one. This way, no relocations are added to the solution.
Steps 1 and 2 are ciclically repeated, until the obtained configuration satisfies all the covering constraints for the current time period.
Proceeding backwards from $t = T$ to $t = 1$, the algorithm guarantees the feasibility of the ambulance configuration for all the periods.
The greedy phase is executed for a predefined number of times and the best solution found is considered.

In order to reduce the computational workload of the algorithm, in some cases the greedy phase is not performed after solving the relaxed problem. This happens when the number of deployed ambulances in the current solution is greater than the best Upper Bound obtained. In this case, in fact, the greedy algorithm will yield to a worse solution.

4.3.2 Neighbourhood search via local branching

The above greedy procedure turns out to be very effective in producing Upper Bounds on the optimal value of MPAL solution, at least during the first iterations of the Lagrangian heuristic algorithm. But, we noticed that after the first iterations the best solution is rarely improved.
Greedy ambulance configurations are clearly overconservative. A certain number of ambulances can for sure be deleted from the greedy solutions.
In order to find a near-optimal configuration, a *refining* procedure is needed. We propose a neighbourhood search, based on the interesting technique proposed by Fischetti and Lodi in 2003, called Local Branching [12].
Local Branching is a MIP technique to search for local optimum in a neighbourhood of a reference solution. The neighbourhood is defined by adding constraints to the original problem.
We outline the Local Branching for general MIPs. For more details, see [12] and [18].

Local Branching

Let (P) be a generic MIP with 0-1 variables:

$$\text{(P)} \quad \text{minimize} \quad \sum_{j \in J} c_j x_j \tag{4.26}$$

$$\text{s.t.} \quad \sum_{j \in J} a_{ij} x_j \geq b_i \quad \forall i \in I \tag{4.27}$$

$$x_j \in \{0,1\} \qquad \forall j \in J \tag{4.28}$$

and let $\overline{x}$ be a *reference solution* (feasible or infeasible) of (P). Denote as $\overline{S}$ the binary support of $\overline{x}$, i.e. $\overline{S} = \{j \in J : \overline{x}_j = 1\}$. Given a positive integer

parameter k, it is possible to define the k-OPT *neighbourhood* $\mathcal{N}(\overline{x}, k)$ of $\overline{x}$ as the set of all the feasible solutions of (P) satisfying the additional *local branching constraint*:

$$\Delta(x, \overline{x}) = \sum_{j \in \overline{S}} (1 - x_j) + \sum_{j \in J \setminus \overline{S}} x_j \leq k \tag{4.29}$$

The two terms in the left-hand side of (4.29) count the number of binary variables whose value is flipped either from 1 to 0 or from 0 to 1, respectively. Thus, the local branching constraint defines a neighbourhood $\mathcal{N}(\overline{x}, k)$ of feasible solutions within Hamming distance at most k from $\overline{x}$.

Local Branching for MPAL

We adapted the local branching approach to define a set of neighbourhood solutions for MPAL. Given a multiperiod solution $\overline{x_j^t}$ (feasible or infeasible) of MPAL and a positive integer paramter k, the k-OPT neighbourhood set is defined as:

$$\mathcal{N}(\overline{x_j^t}, k) = \{x_j^t : \Delta(x_j^t, \overline{x_j^t}) \leq k\} \tag{4.30}$$

where

$$\Delta(x_j^t, \overline{x_j^t}) = \sum_{t \in T} \left(\sum_{j \in \overline{S^t}} (1 - x_j^t) + \sum_{j \in J \setminus \overline{S^t}} x_j^t \right) \tag{4.31}$$

Similarly to the previous definition, $\overline{S^t}$ denotes the binary support of $\overline{x_j^t}$ during time period t: $\overline{S^t} = \{j \in J : \overline{x_j^t} = 1\}$.
Given a reference solution $\overline{x_j^t}$

$$\overline{x_j^t} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{4.32}$$

some examples of 2-OPT neighbourhood solutions are:

$$x_{j,1}^t = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad x_{j,2}^t = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \tag{4.33}$$

The neighbourhood search steps implemented in the Lagrangian heuristic algorithm are solved to optimality with CPLEX, considering the original MPAL problem (4.11)-(4.17) with the additional local branching constraint (4.31).
First, the algorithm generates a greedy solution on the basis of Lagrangian Relaxation solution. Then, the obtained configuration is refined by applying the neighbourhood search.

In order to improve the search of global optimal solution, the parameter k is increased during the execution of the algorithm. We start from an initial value k_1 depending on the number of time periods: $k_1 = 2T$.
At every iteration of the Lagrangian algorithm, we consider the refined solutions obtained in the last δ iterations. If the best feasible solution was never improved, we increase the value of k: $k_{i+1} = k_i + T$. This way, we explore larger neighbourhoods of greedy solutions. Reasonable values for δ are in the interval $\delta \in \{5, 15\}$.
The Lagrangian heuristic algorithm stops when a value $k = k_{max}$ is reached and the solution did not improve during the last δ iterations.

During the refining procedure we execute many consecutive neighbourhood search steps. This way, it is possible to make more than one step towards local optimal solutions.
Denote as N_{max} the maximum number of consecutive neighbourhood search steps to execute for each greedy solution. Then, the process is organized as follows:

Procedure 3 Refining procedure

1. derive a greedy solution x_k
2. $\overline{x_k} \leftarrow x_k$

for $i = 1, ..., N_{max}$ **do**

3.(a) improve $\overline{x_k}$ via neighbourhood search and derive x_k^i

3.(b) $\overline{x_k} \leftarrow x_k^i$

end for

return $x_k^{N_{max}}$

4.4 Computational results

4.4.1 Step-size parameter

In order to evaluate the effect of the step-size parameter on the Lower Bounds produced by the algorithm, we considered different values for λ_k. Typical results are plotted in Fig. 4.2.

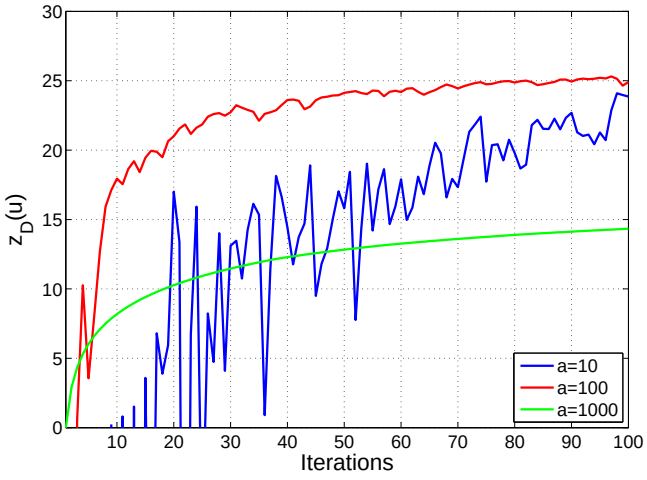
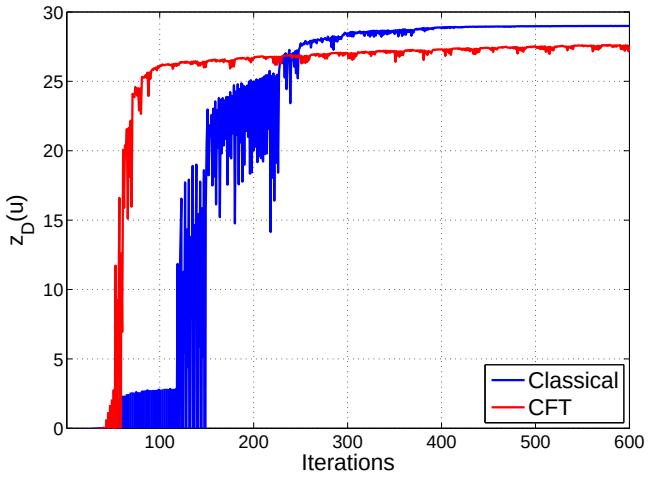


Figure 4.2: Values of objective function of Lagrangian Relaxation, with different values of step-size $\lambda_k = \frac{1}{ak}$.

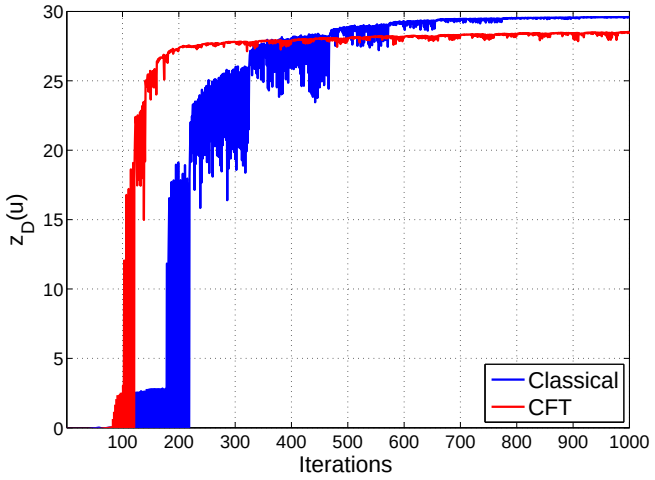
The graph represents the optimal values $z_D(u_k)$ obtained at every iteration of the subgradient method. In all cases, a medium MPAL instance with 4

time periods was considered.
The step-size parameter λ was updated with the strategy $\lambda_k = \frac{1}{ak}$, where k is the iteration number and a is one value chosen from $\{10, 100, 1000\}$.
Observe that the blue line, corresponding to the the value $a = 10$, is very fluctuating. $z_D(u_k)$ is in many cases lower than the current Lower Bound. This is because the step-size parameter is too large, and constraint violations are not properly weighted. The Lower Bound starts to considerably grow after iteration $k = 20$, even if an irregular behaviour is still present.
Consider now the results obtained with the smallest step-size ($a = 1000$). The values of $z_D(u)$ assume a very regular behaviour. The fluctuating effect caused by the large λ_ks is no more visible. However, although the growth of the lower bound is very regular, it is quite slow. The best Lower Bound is much smaller than the one obtained with $a = 10$. The step-size is clearly too small. The algorithm needs a large number of iterations before starting to produce near-optimal Lagrangian multipliers.
When we use $a = 100$, $z_D(u_k)$ grows regularly and quickly after a very small number of iterations. Moreover, a certain stability is reached after the first 50 iterations. The Lagrangian multipliers obtained by the algorithm are approaching the optimal ones.

We now consider the effects produced by the two updating strategies described in Section 4.2.2.
A comparison between classical strategy and CFT strategy is proposed in Fig. 4.3. Different values of p are considered.
The initial step-size λ_1 has been set to a large value: $\lambda_1 = \frac{1}{10}$.
In both cases, the increasing of the Lower Bound is much faster when the CFT strategy is used. With the classical strategy, the algorithm takes a large number of iterations before correcting the initial step-size. The value of λ_k is halved more rarely than in the case of CFT. This happens because a small increase in Lower Bound value is enough to stop the halving process for p iterations.
Employing CFT strategy, a stability in λ_k is quickly reached. Note that the maximum Lower Bound values produced by this strategy are slightly lower



(a) $p = 10$

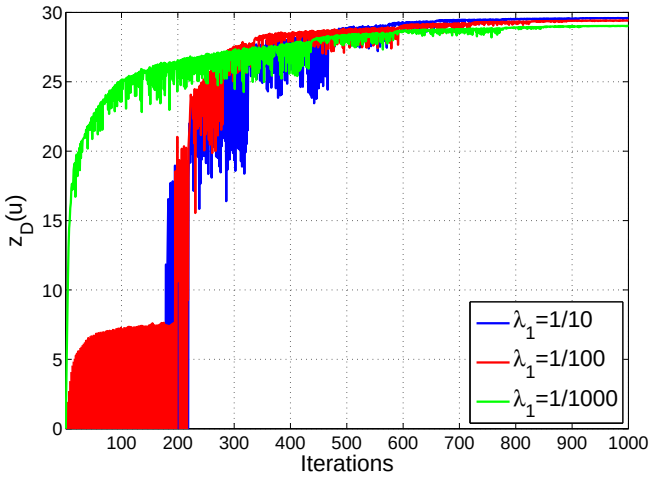


(b) $p = 20$

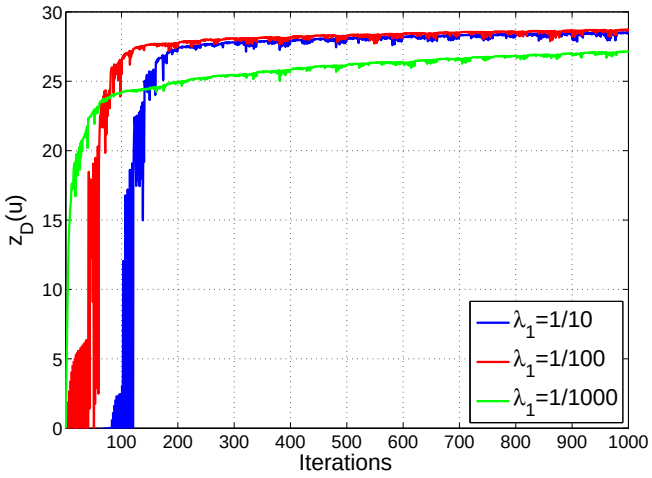
Figure 4.3: Comparison between the classical and the CFT strategies for the updating of λ_k. $\lambda_1 = \frac{1}{10}$.

than those of classical strategy. Anyway, this is not a problem within the search of near-optimal feasible solutions. The objective is not to get to the best Lower Bound value, but to obtain as quickly as possible a sequence of near-optimal Lagrangian Multipliers.

The effect of setting different initial step-size λ_1 is now analyzed. We considered 3 different values for λ_1, and we plotted the values of $z_D(u_k)$ for each iteration. The results are presented in Fig. 4.4.



(a) Classical strategy



(b) CFT strategy

Figure 4.4: Comparison between different values of λ_1. $p = 20$.

Note that in both cases the choice of a large value for λ_1 causes a late increasing of the Lower Bound. When the CFT strategy is considered, the algorithm

quickly detects this problem and corrects the value of the step-size.
CFT appears is an effective strategy for searching near-optimal Lagrangian multipliers. The value of λ_1 has to be selected on the basis of numerical tests. In the case proposed in Fig. 4.4, reasonable values are $\lambda_1 \in [\frac{1}{1000}, \frac{1}{100}]$.

4.4.2 Lagrangian heuristic with greedy

We now present results obtained by the Lagrangian heuristic algorithm. In this case feasible solutions are found with the greedy approach.

The best Lower Bounds and Upper bounds found during an execution of the Lagrangian algorithm are plotted in Fig. 4.5.
The Upper Bound value corresponds to the best feasible solution determined by the greedy algorithm: UB=$\sum_t \sum_j x_j^t$. Lower Bound values are determined by solving the Lagrangian Relaxation (4.18)-(4.23), as explained in the previous sections.
We considered a medium instance of the problem, with 4 time intervals. The optimal value, determined by CPLEX, is $\sum_t \sum_j x_j^t = 32$.

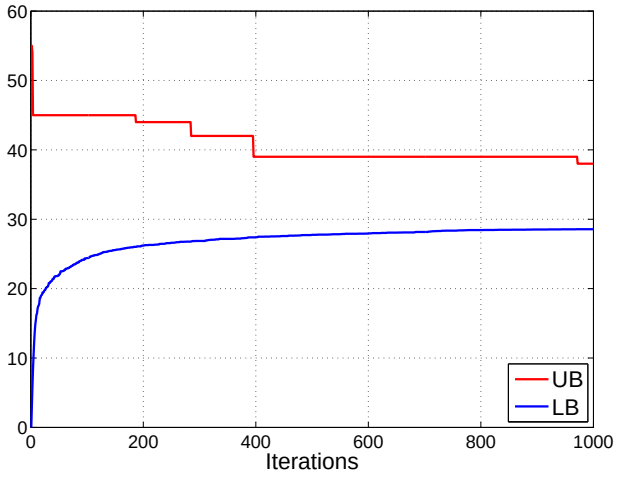


Figure 4.5: Values of Lower Bound and Upper Bound, obtained by the heuristic algorithm during the greedy phase. CFT lambda updating strategy. Medium instance, with 4 time intervals.

As expected, the Upper Bound on the optimal solution is progressively decreased during the execution of the Lagrangian heuristic algorithm.
At the first iterations, UB values are very far from the optimal value of the problem. This is due to the inaccurate Lagrangian multipliers considered by the algorithm in its early stages.
After a small number of iterations, UB value starts to improve. This is due to the effective correction of Lagrangian Multipliers executed by the algorithm. Solutions of the relaxed problem are better in terms of number of covered demand nodes, hence they are easier to be make feasible.
Observe that the algorithm improves the UB value only three times between iterations 100 and 500. The best greedy solution, corresponding to an Upper Bound value of 38, is found at iteration 972. During the previous 572 iterations, the greedy algorithm never improves the Upper Bound value.

4.4.3 Lagrangian heuristic with greedy and local search

In order to refine the solutions found during the greedy phase, we added a neighbourhood search to the Lagrangian-based heuristic algorithm.
The improvements obtained with this technique are shown in Fig. 4.6.
As before, we consider a medium instance of the problem with 4 time intervals. The black line corresponds to the Upper Bound values given by the greedy procedure, and the red line to the Upper Bounds values obtained after the neighbourhood search.
In this case, we execute three consecutive neighbourhood search steps for each greedy solution.
Note that, although the number of neighbourhood search steps is small, we obtain a substantial improvement in the number of deployed ambulances.
The best greedy Upper Bound is obtained at iteration 20 and then it is no longer changed, while the best solution given by the neighbourhood search is frequently updated.
This considerable improvement is not expensive in terms of computing time, since the local search can be executed in a fraction of second. In particular, the time needed by the refining procedure is much smaller than the time

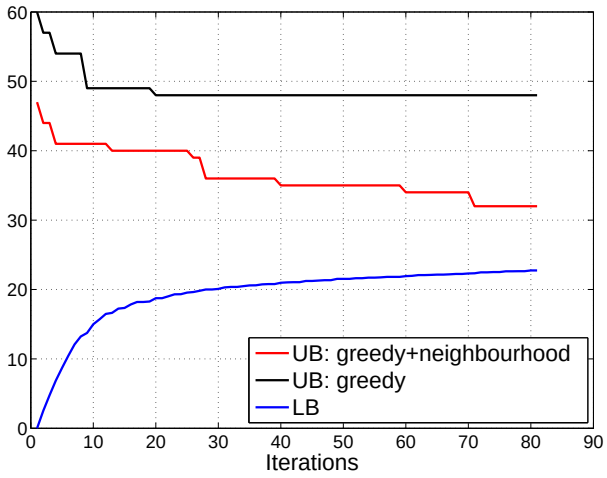


Figure 4.6: Values of the Lower and Upper Bounds obtained by the Lagrangian heuristic algorithm with 3 consecutive neighbourhood search steps for each greedy solution. CFT lambda updating strategy. Medium MPAL instance with 4 time periods.

required for the execution of the greedy procedure.
Note that in this case the algorithm is able to find the optimal solution of the problem, corresponding to the Upper Bound value UB=32.

In Table 4.1 we present results obtained with our Lagrangian-based heuristic, for different size instances of the problem.
The algorithm was coded in AMPL and executed on an Intel Xeon 2.8 GHz CPU with 2GB of RAM memory.
For each solution, we report the following information:

- Instance: size of the problem.
- $|J|$: number of potential location points; we considered $I = J$.
- T: time periods.
- UB: Upper Bound.
- LB: Lower Bound.

- Gap: relative Bound gap. Gap=$(UB - LB)/LB$.
- $t(s)$: CPU time (seconds).

Whenever possible, we report optimal solutions found by CPLEX. In those cases, neither LB nor Gap are displayed.

			Heuristic				CPLEX			
Instance	$\|J\|$	T	UB	LB	Gap	$t(s)$	UB	LB	$\%\Delta z$	$t(s)$
Small	100	3	13	10.61	0.22	29	13			1
		4	17	14.37	0.18	31	17			1
		5	22	18.89	0.16	34	22			1
		6	27	23.19	0.16	47	27			1
Medium	200	3	23	21.03	0.09	126	23			1
		4	32	28.22	0.13	312	32			15
		5	42	36.02	0.16	375	40			11
		6	56	43.96	0.27	896	52			1076
Large	492	3	59	49.04	0.20	1519	56			4818
		4	86	64.95	0.32	2958	78	7.68	0.09	5000
		5	120	82.64	0.45	3472	101	12.00	0.12	5000
		6	148	100.21	0.47	4249	125	16.8	0.13	5000

Table 4.1: Results obtained with heuristic algorithm, different instances of the problem.

Observe that our model can be solved to optimality with CPLEX for small and medium instances within a few seconds. Because of this, the model can be exploited to support the planning of ambulance locations in real EMS systems. In fact, the number of potential location points of these instances is reasonable for representing small and medium size cities.

Our Lagrangian-based heuristic turns out to be effective for solving the model (4.11)-(4.17).

Observe that when small instances of the problem are considered, the algorithm always found the optimal solution. Results for medium instances are equivalent or just slightly different from the optimal solutions given by CPLEX.
The impact of the increase in the number of time intervals is clear in this case. The medium instance with 6 periods was solved to optimality by CPLEX in 1076 seconds, while just 11 seconds were needed for the medium instance with 5 periods. This effect is observed also for the Lagrangian-based heuristic algorithm, but the computing times are not exponential as for CPLEX.

The preliminary results on large size instances are not completely satisfactory, but they indicate that there is a good margin for improvement.
If we compare the Upper Bounds found by the Lagrangian-based heuristic with the Upper Bounds given by CPLEX, we see that the latter are better, even if obtained in a comparable time.
This is due to the poor quality of solutions produced by the greedy algorithm, which are not refined enough by the neighbourhood searches. The refining procedure can be clearly enhanced by executing more consecutive local search steps, until a local optimum is reached. The parameter k, which determines the size of the solution neighbourhood, has to be set to a low value in order to obtain easier local search sub-problems. Thus, the required computing time would not be considerably affected by the increase in the number of neighbourhood search steps.

In addition, the greedy procedure can be improved by considering different strategies for the derivation of feasible solutions. For example, a hierarchy rule can be defined and used to choose the best location point in which to add an ambulance. A similar approach has been proposed in [7, 5] for solving the classical Set Covering Problem.
Clearly, a substantial improvement in terms of computating time can also be obtained by employing an efficient programming language for the coding of the algorithm, like C++. Indeed AMPL is an excellent software for modeling optimization problem, but it is not very efficient since it is interpreted.

Conclusions

In this work we have addressed the important problem of ambulance location. We have proposed and investigated a multiperiod probabilistic model, which takes into account the main aspects of the problem.
In particular, we accounted for the variability of traffic conditions and the changes in the demand pattern during the day, as well as for the possibility of system congestion due to the unavailability of ambulances.

In the first part of the work we reviewed the main models presented in the literature regarding the ambulance location problem. We discussed the positive and negative aspects of each model, in order to evaluate the best way to develop our model.

The multiperiod model relies on the time discretization of the problem. We subdivided the time horizon into a set of consecutive intervals, so that the system conditions are homogeneous within each one of them.
In order to limit the differences between consecutive ambulance configurations, we imposed constraints on the maximum number of relocations allowed at the beginning of each period. As an alternative, we can restrict the total number of relocations on the whole time horizon.

The probabilistic version of our multiperiod model explicitely considers the busy probabilities of ambulances. Considering a the definition of system reliability level, we required that the solution satisfies a predetermined reliability for each demand zone.
Our probabilistic model is an adaptation of a general method recently pro-

posed for the classical Set Covering Problem.
A strenghtend formulation of the problem was also considered, which led to a speed up in the computing time. In some cases, a 20% gain was obtained.

We showed that our models can be solved to optimality for small to medium instances (100 to 200 potential location points and 6 time periods) within a short amount of time, with a state-of-the-art solver like CPLEX.
Since the number of potential location points of a medium size instance is certainly reasonable for representing the territory of a small or medium city, the model can be used to guide the planning of ambulance locations in real EMS systems.
For larger instances (about 500 potential location points and more than 4 time periods, as in the case of Milan) the model becomes rather challenging to be solved to optimality.
Therefore, we also proposed a Lagrangian-based heuristic algorithm. Feasible solutions are initially derived with a greedy procedure, and then refined using a neighbourhood search approach based on local branching.

Solutions of the heuristic algorithm are in agreement with the optimal solutions obtained with CPLEX, at least for small and medium size instances of the problem. In many cases, the Lagrangian algorithm is able to determine the optimal solution in a reasonable amount of time, with very few local search steps.
The preliminary results on larger instances are not completely satisfactory, but there is still a good margin for improvement.
In particular, it is possible to increase the perfomance of the refining procedure by making a much larger number of small local search steps, e.g. until a local optimal solution is reached. This should not considerably affect the computational time since local search steps are not expensive to solve to optimality.
The greedy procedure can also be improved, by defining a hierarchy between the potential location points so as to choose the best zone in which to add an ambulance. A substantial improvement in computational time can clearly be

obtained also by implementing the algorithm in a more efficient programming language.

Bibliography

[1] *T. Andersson, P. Värbrand.* Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society* **58**, 195-201 (2007)

[2] *M.O. Ball, L.F. Lin.* A reliability model applied to emergency service vehicle location. *Operations Research* **41**, 18-36 (1993)

[3] *J.E. Beasley.* A Lagrangian Heuristic for Set Covering Problems. *Naval Research Logistics* **37**, 151-164 (1990)

[4] *L. Brotcorne, G. Laporte, F. Semet.* Ambulance location and relocation models. *European Journal of Operational Research* **147**, 451-463 (2003)

[5] *A. Caprara, M. Fischetti, P. Toth.* A heuristic method for the Set Covering Problem. *Operations Research* **47**, 730-743 (1999)

[6] *A. Caprara, M. Fischetti, P. Toth.* Algorithms for the Set Covering Problem. *Annals of Operations Research* **98**, 353-371 (2000)

[7] *S. Ceria, P. Nobili, A. Sassano.* A Lagrangian-based heuristic for large-scale set covering problems. *Mathematical Programming* **81**, 215-228 (1998)

[8] *S. Chapman, J. White.* Probabilistic formulation of emergency service facilities location problems. *ORSA/TIMS paper* (San Juan, Puerto Rico, 1974)

[9] *R.L. Church, C. ReVelle.* The maximal covering location problem. *Papers of the Regional Science Association* **32**, 101-118 (1974)

[10] *M. Daskin*. A Maximum Expected Covering Location Model: Formulation, Properties and Heuristic Solution. *Transportation Science* **17**, 48-70 (1983)

[11] *M. Fischetti*. Robustness by cutting planes and the Uncertain Set Covering Problem. *Technical Report DEI - University of Padova* (2009)

[12] *M. Fischetti, A. Lodi*. Local Branching. *Mathematical Programming* **98**, 23-47 (2003)

[13] *M.L. Fisher*. The Lagrangian Relaxation method for solving integer programming problems. *Management Science* **27**, 1-18 (1981)

[14] *M.L. Fisher*. An Applications Oriented Guide to Lagrangian Optimization. *Interfaces* **15**, 10-21 (1985)

[15] *R.D. Galvão, R. Morabito*. Emergency service systems: The use of the hypercube queuing model in the solution of probabilistic location problems. *International Transactions in Operational Research* **15**, 525-549 (2008)

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

[17] *J.B. Goldberg*. Operations Research Models for the Deployment of Emergency Service Vehicles. *EMS Management Journal* **1**, 20-39 (2006)

[18] *P. Hansen, N. Mladenović, Dragan Urošević*. Variable neighbourhood search and local branching. *Computer & Operations Research* **33**, 3034-3045 (2006)

[19] *M. Held, R.M. Karp*. The traveling salesman problem and minimum spanning trees: Part II. *Mathematical Programming* **1**, 6-25 (1971)

[20] *K. Hogan, C. ReVelle*. Concepts and applications of backup coverage. *Management Science* **34**, 1434-1444 (1986)

[21] *J.P. Jarvis*. Approximating the equilibrium behaviour of multi-server loss systems. *Management of Science* **31**, 235-239 (1985)

[22] *P. Kolesar, W.E. Walker*. An algorithm for the dynamic relocation of fire Companies. *Operations Research* **22**, 249-274 (1974)

[23] *R.C. Larson*. A hypercube queuing model for facility location and redistricting in urban emergency services. *Computer & Operations Research* **1**, 67-95 (1974)

[24] *R.C. Larson*. Approximating the performance of urban emergency service systems. *Operations Research* **23**, 845-868 (1975)

[25] *V. Marianov, C. Revelle*. The Queuing Probabilistic Location Set Covering Problem and some extensions. *Socio-Economic Planning Sciences* **28**, 167-178 (1994)

[26] *H.K. Rajagopalan, C. Saydam, J. Xiao*. A multiperiod set covering location model for dynamic redeployment of ambulances. *Computers & Operations Research* **35**, 814-826 (2008)

[27] *C. ReVelle K. Hogan*. A reliability constrained siting model with local estimates of busy fractions. *Environment and Planning B: Planning and Design* **15** 143-152 (1988)

[28] *C. Revelle, K. Hogan*. The maximum reliability location problem and α-realiable p-center problem: Derivatives of the probabilistic location set covering problem. *Annals of Operations Research* **18**, 155-174 (1989)

[29] *C. Revelle, K. Hogan*. The maximum availability location problem *Transportation Science* **23**, 192-200 (1989)

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

[31] *S. Perego, C. Romantini*. 118: Analisi dei dati e modelli per la dislocazione dei mezzi di soccorso. *Master of Science Thesis*, (2006)

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

[33] *C. Toregas, R. Swain, C. ReVelle, L. Bergman*. The location of emergency service facilities. *Computers & Operations Research* **19**, 1363-1373 (1971)

[34] *L.A. Wolsey*. Integer programming. *Wiley-Interscience publication* (1998)