Chapter 2

Literature Review

The aim of the review is to establish the broader environment within which the optimization of logistic operations occur. The process of *City Logistics* is introduced. The multi-actor environment is emphasized, where a number of non-transport influences affect the transport activity within cities. The need exists to describe and predict influencing factors in the urban network. For this purpose a new modelling approach is introduced that takes the various perceptions of stakeholders into account.
Vehicle routing and scheduling procedures are formalized as core techniques to model the logistic system in urban areas. Realistic variants of the vehicle routing problem are introduced, and their impact on the modelling task is highlighted.

The complex nature of vehicle routing problems requires the introduction of heuristic solution algorithms. The second half of the chapter reviews the process of finding a good initial solution – a critical parameter in the solution quality of the optimization process. Finally, three improvement heuristics are introduced that proved successful in vehicle routing problems.

2.1 Fundamental concepts

Transportation, as the business of conveying passengers and/or goods [2], has evolved tremendously in South Africa over the past century [1]. Transportation occurs even in the absence of infrastructure: nature transports seeds and natural materials through forces, such as wind and water; people in remote areas travel by foot in the absence of accessible road surfaces, and get from point A to point B, even across inhospitable areas. It is, however, as a result of the frequent use of transport routes that the users of the routes collectively demand improvements. The improvements apply to both travel conditions, such as shorter journeys and smoother rides, and travel reliability, for example the accessibility of roads due to weather conditions.

It is also not uncommon for users to pay for improvements. The Chair of Transportation Engineering elaborates on the evolution of transportation engineering as a discipline [38]. The demand for improvements gave rise to the development of transport infrastructure such as hardened road surfaces, bridges, ports, etc., and established disciplines such as structural, pavement, railway, and traffic engineering.

The discipline of traffic engineering is based on developing a mathematical understanding of the operation of vehicles on sections of the road network. This ability to model vehicular movement is extended when the need emerge to forecast *what* transport facilities would be required in the future. The discipline of transport planning, that applies to single and multi-modal problems, emerged. It can also be used to understand the inter-relationship of land-use and the demographic aspects that generate the need for transportation.

2.2 Modelling City Logistics

It is appropriate to elaborate on a few basic transport concepts to emphasize the complex relationships between various stakeholders in society.

2.2.1 Transport Concepts

The function of transport is to move passengers and goods from where they are to where they want to be or where their relative value is greater. The demand for transport arises from the fact that not all places are equally endowed with resources, and surpluses are then moved to areas experiencing shortages [39]. Transport is a service and does not occur for its own purpose, and is therefor referred to as a *derived demand*. In the development of industries, transport plays a vital part in linking the sources of raw material, manufacturing or processing centers, and the markets. Raw material are moved to, and between processing centers, while finished goods are moved via wholesalers and retailers to the point of consumption or utilization. Transport is essential to enable people to travel between their homes and places of employment. This is even more true in South Africa where the average commuting time for residents is in excess of one hour (Table 1.1).

Transport and development are closely linked, and effective transport is a prerequisite for the development of a country. Investment alone, however, does not guarantee prosperity. Whilst *transport* is focussed on the physical movement of objects, be it freight or passengers, *logistics*, in the context of this dissertation, is concerned with the activity of transport within a larger environment. The environment of the logistics system is illustrated in

figure 1.2. The *City Logistics* process has numerous interfaces with various built environment, development, engineering, and geographical disciplines. It is concerned with the *mobility*, *sustainability*, and *liveability* of cities [56]. A few non-transport aspects of logistics is explained:

Economic aspects of transport

Transport cost consists mainly of a fixed portion, C_f at a point A, and a variable portion based on the distance travelled from point A to a point B, as indicated in figure 2.1. The cost function need not be linear as indicated,

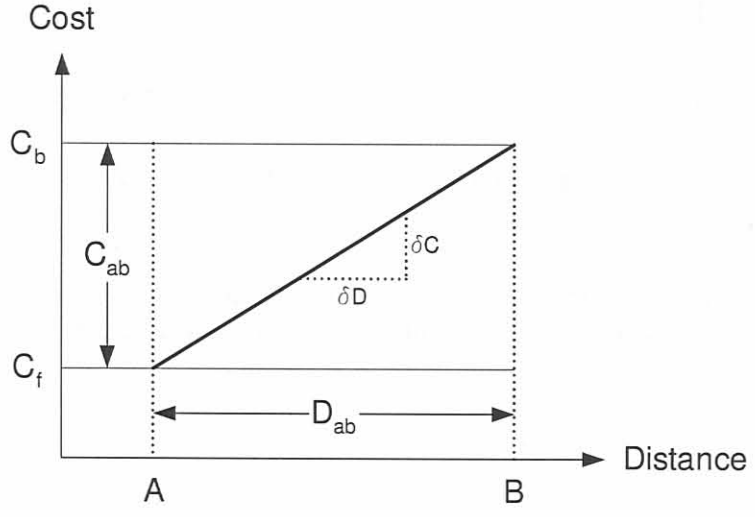

Figure 2.1: Transport cost components

and the slope of the cost function could be calculated as $\frac{\delta C}{\delta D}$. The transport cost between points A and B, C_b, can be calculated as $C_b = C_f + D_{ab} \times \frac{\delta C}{\delta D}$. The transport cost function is influenced by improvements in transport facilities and infrastructure, e.g. improved transhipment methods, more efficient vehicles, and optimized vehicle routes.

Competing carriers and shippers react differently to these advances, and could gain a *competitive advantage* by effectively implementing the advances. Consider the example illustrated in figure 2.2 where two carriers, A and B, compete on the basis of cost. At present, carriers A and B supply their services at a cost of C_{a1} and C_{b1} respectively. A technological advance becomes available to both carriers, and carrier A increases the fixed portion of its transport cost by investing in the advance – to a greater extent than carrier B. This results in carrier A being able to compete in the market at a cost of C_{a2}, lower than carrier B's cost of C_{b2}.

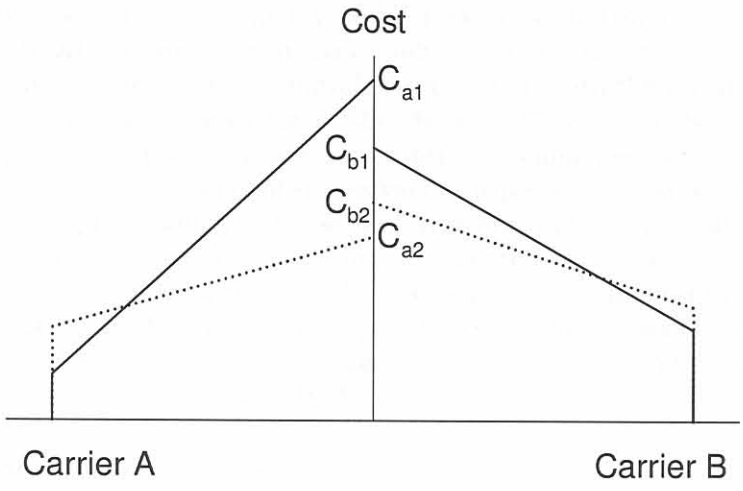

Figure 2.2: Effect of transport advances on comparative advantage

Economies of scale influences the cost per unit transported, and is illustrated in figure 2.3. This can be achieved should if a carrier is able to

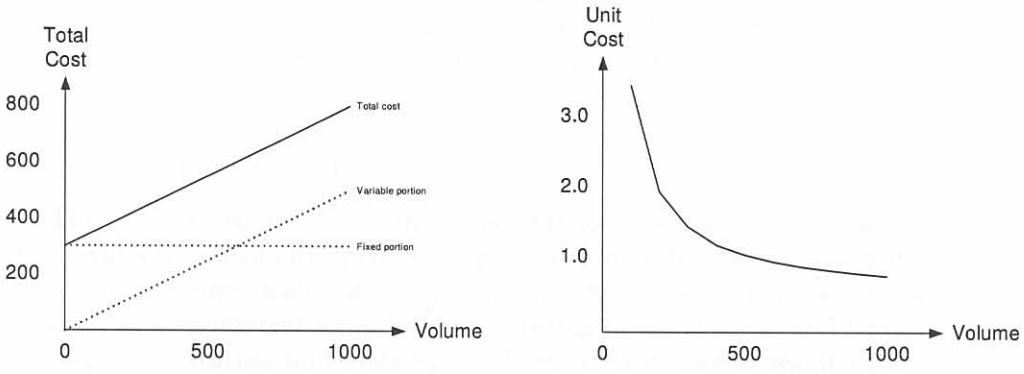

Figure 2.3: Fixed and variable costs

consolidate loads on a vehicle, or acquire new business for that vehicle.

Although transport services is generally initiated by the private sector, the concept of *public-private-partnerships* have gained prominence, and could benefit both sectors. It is important to preserve transport through joint initiatives. The term *preserve* is used to emphasize the perishable nature of the service. Once a vehicle, such as a delivery van, or a bus, have departed with a load factor less than 1, the opportunity to sell the *empty*

seats for that trip is immediately, and permanently, destroyed.
The South African government has introduced commercialization and privatization policies to share the control over transport with the private sector. Examples include the creation of the *Airports Company of South Africa* and the concession of rail services to private companies such as *Metro-Rail* [36, 37]. Government subsidizes various aspects of transport activities in both the public and private sector. Bus and rail subsidies amount to R3,737 million in 2002/03, which include maintenance, as well as the creation of additional infrastructure [15]. Government is implementing a *Taxi Re-capitalization Programme* in the private sector whereby economic potential is unlocked in a taxi industry that provides 65% of the 2.5 billion annual passenger trips in urban areas [8, 12, 15].

Political aspects of transport

The relationship between government and transport is bidirectional. As a *user* of transport, the government communicates its decisions to all areas. This happens indirectly when provincial and local government officials travel between parliament and their representative offices. Government also uses transport to move its defence force to defend the county's borders. Transport is used more strategically to achieve developmental goals in the country, and to achieve political goals in the form of incentives and investments to attract industrial parties to settle in specific areas. As a transport *influencer*, government provides funds for infrastructure, and supports innovations through pilot and demonstration projects.

Social aspects of transport

The main social impact of improved transport is the reduced friction that results from distance: people are given the opportunity to participate, socially and economically, in a larger geographical area. This results in various cultures interacting and communicating different ideas and frames of reference. Mutual understanding is fostered that should result in the elimination of suspicion between races and/or cultures. The improvement of transport, specifically air travel, has widened the scope of opportunities between nations to the extent that the world is often referred to as a *global village.*

Developments in the automotive industry, and the accessibility to cars in South Africa, have contributed to increased use of this mode of transport. In the beginning of 2002, there were some 6.9 million registered vehicles in South Africa, more than 3.98 million of which were cars [15]. Although the private car provides spatial freedom to travel easily, and give status in terms of prosperity and adulthood, it does produce some significant social

problems:

- Residential development occur at a lower density, and further away from the working place.
- Specialized urban activities, such as shopping and entertainment, are concentrated in a way that is supporting the use of private cars, as opposed to infrastructure supporting pedestrian access.
- A reduction in the number of public transport trips. This results in inefficient public transport use, decreased service levels, and eventually more people reverting to private car use.

The higher number of vehicles also results in a demand for more roads: infrastructure that, in itself, create pedestrian barriers, force inhabitants to relocate, change land use patterns, and reduce residential (social) quality. The effects of freeways are even more profound. De Boer [16] introduces transport sociology and states that traffic, and the desires of planners to provide additional infrastructure, should be tamed. The emphasis should be shifted from the *expansion* of infrastructure, to the *management* of the available transport infrastructure. When a transport system is in place, its utilization and physical condition should be improved. Additional infrastructure creates potential competition, resulting in suboptimal utilization of all infrastructure.

Environmental aspects of transport

Transportation is a major source of air pollution, emitting tons of carbon monoxide (CO), hydrocarbons (HO), and nitrogen oxides (NO_x) into the air [35]. The environmental effects of a transport system is not limited to a negative change in the chemical composition of the atmosphere, but also impacts the social environment. Noise pollution created by traffic results in sleep and speech interference, annoyance, and impairment of hearing after exposure over long periods of time [39].

2.2.2 A new approach to City Logistics Modelling

It is clear from the previous section that stakeholders do not participate as individual entities, but rather as complex networks. De Bruijn and tenHeuvelhof [17] identifies four important characteristics of stakeholder networks:

- *Interdependence.* Actors cannot achieve their goals without cooperation, as they are dependent on the resources of others actors, such as funding, information, and statutory powers.

- *Pluriformity.* Corporate actors do not behave as individuals, but as coalitions, since their constituents may have diverging and competing interests.

- *Self-containment.* Corporate actors are inclined to close themselves off from their environment, developing their own frame of reference and norms, making it harder to induce the cooperation.

- *Instability.* Positions and relations in policy networks are continually undergoing changes.

Villa [63] confirms autonomous decision making in networks. These characteristics impede the ability of stakeholders to make decisions that is optimal for the sustainability of the system as a whole. It does open a window of opportunity to address the dynamics of the relationships according to each stakeholder's perception. The Thomas-theorem states that *if man define situations to be real, they are real in their consequences* [60]. It implies that the operations research practitioner should not pursue a comprehensive model of a specific situation, but rather a set of models reflecting the diversity in actor perceptions. The soundness of the set of Thomas-models is not determined by the degree of correspondence to reality, but by the acuteness with which it mirrors the assumptions that actors make about their reality.
It may appear as if the Thomas-models are in conflict with figure 1.4 and the discussion in paragraph 1.4 that argues that the model assumptions should be challenged to ensure that a real representation of the problem is modelled. It is the opinion of the author that the Thomas-model approach emphasizes the importance of the operations research practitioners' skillful definition of the target audience (object) before engaging in the modelling task.

Van Duin *et al* [62] address the perspectives of individual actors and their strategic behavior through *perception based modelling.* There are various analysis and modelling techniques to capture decision-making processes. Taniguchi *et al* [57] discuss a computerized support tool, called *Dynamic Actor Network Analysis* (DANA), and proceed to introduce a new approach to city logistics modelling.

Dynamic actor network analysis

The first step towards modelling the stakeholder network is to capture the perceptions of multiple actors. The DANA tool is set up as an open database into which several perceptions of actors can be submitted. Perceptions are modelled in terms of assumptions. *Factual assumptions* represent how an actor perceives the current state of his environment. *Causal assumptions* represent changes that will occur in the perception of the actor, and

uses *if-then* statements. *Teleological assumptions* represent the actor's view on his desirability, or purposefulness in the network. An important feature of the tool is a query generator. Queries include questions like: *which actors have conflicting goals on a specific factor*, or *which actors have different definitions for a factor?* These conflicts are brought to the modeler's attention during the design of a logistics model.

Performance measures

Performance measures, or performance indicators, are the normative values of the perception-based factors that represent how an actor perceives the current state of his environment. It is used to compare actual performance against a pre-defined norm. An actor creates the norm by attaching a quantified value, based on his perceptions, to a factor influencing his environment.

Logistics modelling

The last part of the approach is directed towards the calculation of the value part of the performance indicators. Dedicated models are developed to measure the impact of logistics concepts. The models are based on the important factors identified during the dynamic actor network analysis, as interpreted from the operations research practitioner's perspective. The challenge is to develop a model at such a level that it is both comprehensive and easy to understand for the actors involved, yet sufficiently detailed to be validated in practice.

2.3 Vehicle routing

There are significant features of truck operations in urban areas affecting mobility in cities:

- Pickup/delivery trucks visit a number of customers on a single trip. Optimized route schedules decreases operational expenses
- Several pickup/delivery trucks are usually operated as a group by a shipper/carrier. It is possible to reduce the number of vehicles used, as improved load factors increase fleet efficiency.
- Each customer specifies a time window to be visited by the pickup/delivery vehicle, complicating route schedules for carriers

Vehicle routing and scheduling procedures address the mobility in cities [57]. *Vehicle travel time/distance*, *total fleet cost*, and *customer satisfaction* are factors that are valued by both carriers and shippers. Vehicle routing and scheduling problems involve an optimization process of assigning customers

to trucks and determining the visiting order of customers on truck routes.

According to Van Breedam [61] and Laporte [27], the general *Vehicle Routing Problem* (VRP) can be defined as the problem of finding a set of routes for a fleet of vehicles, which have to serve a number of customers (also referred to as stops or nodes) by offloading their goods. The vehicles depart from, and return to, a single depot. Vehicles must complete their individual routes within a maximum total route time. Although the basic VRP has been described mathematically in section 1.2.3, it will now formally be defined.

Let:

N	be the total number of customers
q_i	be the known demand for node i, where $i = \{1, 2, \ldots, N\}$
s_i	be the service time at node i, where $i = \{1, 2, \ldots, N\}$
d_{ij}	be the distance between nodes i and j, where $i, j = \{1, 2, \ldots, N\}$
c_{ij}	be the cost incurred on the arc between nodes i and j, where $i, j = \{1, 2, \ldots, N\}$
t_{ij}	be the travel time between nodes i and j, where $i, j = \{1, 2, \ldots, N\}$
K	be the total number of vehicles available
p	be the capacity of each vehicle in the homogeneous fleet

The principle decision variable, x_{ijk}, is defined as

$$x_{ijk} = \begin{cases} 1 & \text{if vehicle } k \text{ travels from node } i \text{ to node } j, \text{ where} \\ & i, j = \{1, 2, \ldots, N | i \neq j\}, k = \{1, 2, \ldots, K\} \\ 0 & \text{otherwise} \end{cases}$$

$$\min z = \sum_{i=0}^{N} \sum_{j=0, j \neq i}^{N} \sum_{k=1}^{K} c_{ij} x_{ijk} \tag{2.1}$$

subject to

$$\sum_{j=1}^{N} x_{0jk} = \sum_{j=1}^{N} x_{j0k} = 1 \qquad \forall k = \{1, 2, \ldots, K\} \tag{2.2}$$

$$\sum_{j=1}^{N} \sum_{k=1}^{K} x_{0jk} \leq K \tag{2.3}$$

$$\sum_{i=1;i\neq j}^{N} \sum_{k=1}^{K} x_{ijk} = 1 \qquad \forall j \in \{1, 2, \ldots, N\} \tag{2.4}$$

$$\sum_{j=1;j\neq i}^{N} \sum_{k=1}^{K} x_{ijk} = 1 \qquad \forall i \in \{1, 2, \ldots, N\} \tag{2.5}$$

$$\sum_{i=1}^{N} q_i \sum_{j=0;j\neq i}^{N} x_{ijk} \leq p \qquad \forall k \in \{1, 2, \ldots, K\} \tag{2.6}$$

$$x_{ijk} \in \{0, 1\} \tag{2.7}$$

The objective of the problem is to minimize the total travel cost incurred during the process of servicing the N customers. The element c_{ij} in (2.1) can be replaced with either t_{ij} to minimize travel time, or with d_{ij} to minimize travel distance. Constraint (2.2) ensures that all routes start and end at the depot, node 0, while (2.3) ensures that the maximum number of vehicles/routes are not exceeded. Constraints (2.4) and (2.5) limit the number of visits to each node to 1, while (2.6) ensures that the cumulative demand on any route does not exceed the vehicle capacity.

2.3.1 The vehicle routing problem and its variants

The basic VRP assumes that nodes can be visited anytime during the route, that all vehicles are homogeneous in terms of cost and capacity, and that each vehicle can only service one route during a scheduling period. There exist numerous extensions to the VRP. These arise when additional side constraints are added to adapt the basic VRP to many real-life business scenarios. The three variants considered in this dissertation are described in the following paragraphs.

Time windows

A *time window* can be described as a window of opportunity for deliveries. It is an extension of the VRP that has been researched extensively. Examples include the work of Ibaraki *et al.* [25], Taillard [50], Taillard *et al.* [51], and Tan *et al.* [54]. A time window is the period of time during which deliveries can be made to a specific customer i, and has three main characteristics:

- the earliest allowed arrival time, e_i (also referred to as the *opening time*),
- the latest allowed arrival time, l_i (also referred to as the *closing time*),
- whether the time window is considered *soft* or *hard.*

Consider the example, illustrated in figure 2.4, where customer i requests delivery between 07:30 and 17:00.

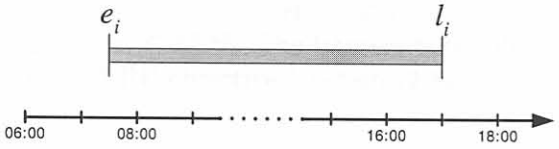

Figure 2.4: Double sided hard time window

To distinguish between the actual and the specified times of arrival, the variable a_i denotes the actual time of arrival at node i. Should the actual arrival time at node i, a_i, be earlier than the earliest allowed arrival at the node, e_i, then the vehicle will incur a waiting time, w_i, which can be calculated as $w_i = \max\{0, e_i - a_i\}$. The introduction of time windows to the basic VRP sees the introduction of three new constraints.

$$a_0 = w_0 = s_0 = 0 \tag{2.8}$$

$$\sum_{k=1}^{K} \sum_{i=0;i\neq j}^{N} x_{ijk}(a_i + w_i + s_i + t_{ij}) \leq a_j \qquad \forall j = \{1, 2, \ldots, N\} \tag{2.9}$$

$$e_i \leq (a_i + w_i) \leq l_i \qquad \forall i = \{1, 2, \ldots, N\} \tag{2.10}$$

Constraint (2.8) assumes that vehicles are ready and loaded by the time the depot opens, which is indicated as time 0 (zero). Constraint (2.9) calculates the actual arrival time, while (2.10) ensures that each customer i is serviced within its time window.
When both an earliest and latest allowed arrival is stipulated, the time window is referred to as *double sided.* If no arrivals are allowed outside of the given parameters, the time window is said to be *hard*, as is the case in figure 2.4. When delivery is allowed outside the specified time window, the time window is said to be *soft*, and the lateness is penalized at a cost of α_i. Customer i may specify a maximum lateness, L_i^{max}. The example illustrated in figure 2.5 sees customer i specifying a time window between 07:30 and 15:30. The customer will, however, allow late deliveries until 17:00. A hard time window is therefor a special type of soft time window where $L_i^{max} = 0$.

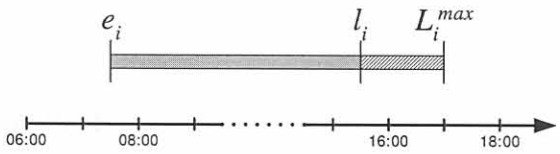

Figure 2.5: Soft time window

Should a vehicle arrive after the latest allowed arrival time, l_i, but prior to the maximum lateness, L_i^{max}, the lateness at node i, L_i, can be calculated as $L_i = \max\{0, a_i - l_i\} | a_i \leq L_i^{max}$. The lateness is penalized by introducing a penalty term to the VRP objective function (2.1).

$$\min \sum_{i=0}^{N} \sum_{j=0, j\neq i}^{N} \sum_{k=1}^{K} c_{ij} x_{ijk} + \sum_{i=1}^{N} \alpha_i \times \max\{0, L_i\} \tag{2.11}$$

The time window for the depot, node 0, can be specified. The case illustrated in figure 2.6 sees the depot specifying operating hours (time window) from 06:00 to 18:00, while the first customer on the route, customer 1, specifies a time window between 07:00 and 09:00, and the last customer, customer n, requests delivery between 15:00 and 17:00.

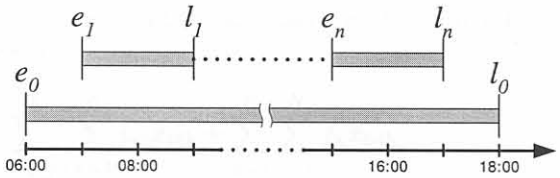

Figure 2.6: Time window for the depot, node 0

Should a customer specify multiple time windows, an indexing symbol, a, is introduced as superscript to the earliest and latest allowed arrival times, respectively, where $a \in \{1, 2, \ldots, A\}$ in which A indicates the maximum number of time windows allowed for each customer. Consider the example where customer n requests delivery either between 06:30 and 09:00, or between 16:00 and 17:30. The case is illustrated in figure 2.7. This example

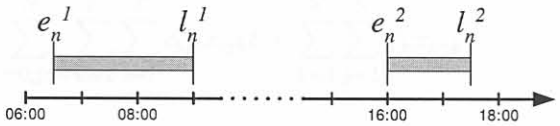

Figure 2.7: Multiple time windows

is typical of residents requesting home shopping deliveries outside business

hours.

Heterogeneous fleet

Gendreau *et al.* [20] propose a solution methodology for cases where the fleet is heterogeneous, that is, where the fleet is composed of vehicles with different capacities and costs. Their objective is to determine what the optimal fleet composition should be, and is referred to as either a *Heterogeneous Fleet Vehicle Routing Problem* (HVRP), or a *Fleet Size and Mix Vehicle Routing Problem* (FSMVRP). Taillard [50] formulates the *Vehicle Routing Problem with a Heterogeneous fleet of vehicles* (VRPHE) where the number of vehicles of type t in the fleet is limited; the objective being to optimize the utilization of the given fleet. Salhi and Rand [45] incorporates vehicle routing into the vehicle composition problem, and refer to it as the *Vehicle Fleet Mix problem* (VFM).

The implication of a heterogeneous fleet on the standard VRP is that T type of vehicles are introduced, with $t \in \{1, 2, \ldots, T\}$. The vehicle capacity parameter p is changed. The new parameter, p_t, represents the capacity of vehicles of type t, resulting in each vehicle k having a unique capacity, p_k. The use of one vehicle of type t implies a fixed cost f_t. A unique fixed cost, f_k, is introduced to each vehicle k, based on its vehicle type. The objective function changes to

$$\min \sum_{i=0}^{N} \sum_{j=0, j\neq i}^{N} \sum_{k=1}^{K} c_{ij} x_{ijk} + \sum_{k=1}^{K} \sum_{j=1}^{N} f_k x_{0jk} \tag{2.12}$$

while (2.6) changes to indicate the new capacity parameter

$$\sum_{i=1}^{N} q_i \sum_{j=0; j\neq i}^{N} x_{ijk} \leq p_k \quad \forall k = \{1, 2, \ldots, K\} \tag{2.13}$$

Taillard [50] introduces a variable c_{ijt} to represent the cost of travelling between nodes i and j, using a vehicle of type t. It is possible to introduce the variable portion of the vehicle cost into the objective function (2.12). The introduction will lead to

$$\min \sum_{i=0}^{N} \sum_{j=0, j\neq i}^{N} \sum_{k=1}^{K} \sum_{t=1}^{T} c_{ijt} x_{ijk} \xi + \sum_{k=1}^{K} \sum_{j=1}^{N} f_k x_{0jk} \tag{2.14}$$

where

$$\xi = \begin{cases} 1 & \text{if vehicle } k \text{ is of type } t, \text{ where } k = \{1, 2, \ldots, K\}, \\ & t = \{1, 2, \ldots, T\} \\ 0 & \text{otherwise} \end{cases}$$

Double scheduling

It is often not viable to assume that each vehicle will only complete a single route. *Double scheduling* is concerned with the case where a vehicle could complete deliveries on a scheduled route, return to the depot where its capacity is renewed, after which a second, or consecutive trip is executed with the renewed capacity. Taillard *et al.* [53] refer to this type of problem as the *Vehicle Routing Problem with Multiple use of vehicles* (VRPM). Butt and Ryan [11] consider the *Multiple Tour Maximum Collection Problem* (MTMCP) and assumes that the routes are constrained in such a way that all of the customers cannot be visted. Their approach aims to maximize the number of customers serviced. Brandão and Mercer [9] introduce the *Multi-Trip Vehicle Routing Problem* (MTVRP), and address the combination of multiple trips with time windows.

This dissertation considers a vehicle that starts and ends its tour at the depot. A *tour* consists of one or more *routes*, each starting and ending at the depot. The same vehicle can only be used for two or more routes if the routes do not overlap. As opposed to (2.8), multiple routes require a service time to be specified for the depot. Consider the example illustrated in figure 2.8. The depot has a time window from 06:00 to 18:00. A vehicle fills its

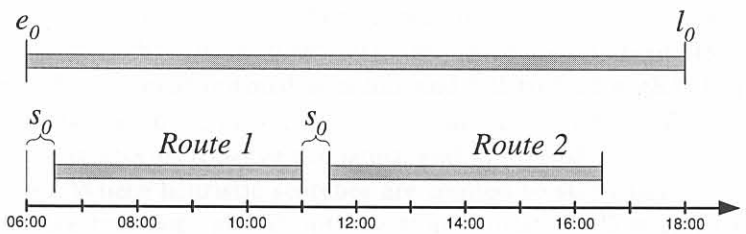

Figure 2.8: Double scheduling

capacity at the depot for a time period of $s_0 = 0.5$ hours. It leaves the depot at 06:30, services the first route, and returns to the depot at 11:00, where its capacity is renewed. A second route, of five hours, is serviced before the vehicle returns to the depot.

The mathematical implication of double scheduling on the basic VRP could not be established from literature. Taillard *et al.* [53] confirm that this type of problem has received very little attention in literature. This dissertation proposes a way to deal with multiple routes. The proposed solution involves a time verification process. If a vehicle arrives back at the depot at time a_m, and the service time is specified as s_0, then the vehicle is considered for an additional route on its current tour if, after the capacity has been renewed, the depot's time window is still open. The case is represented

in (2.15).

$$a_m + s_0 \leq l_0 \tag{2.15}$$

2.3.2 Computational complexity of the VRP

Problems that are considered hard to solve are those problems for which there are not polynomial solution algorithms, and are referred to as *nondeterministic* (NP) *class problems*. The problems that require an inordinate amount of computer processing time in the *NP* class are identified as *NP-hard* problems. Lenstra and Rinnooy, and Laporte, have classified vehicle routing and scheduling problems as *NP-hard* problems [27, 29].

When confronted with difficult problems to be solved (*NP-hard*), operations research (OR) practitioners have used approximation techniques to arrive at a good solution. The approximate rule-of-thumb techniques where optimality cannot be assured are classified as *heuristic techniques* in OR practice [65].

2.3.3 Solving the vehicle routing problem

Heuristics typically uses a greedy approach to obtain a good initial solution in efficient time and then incrementally improve the solution by neighborhood exchanges or local searches. As a result, heuristics tend to get trapped in a local optimal solution and fail to find a global optimum. Heuristics have evolved into *global optimization heuristics*. These are general master strategies to solve problems, and are based on intelligent search techniques. Where heuristic searches are limited to steps that will improve the objective function, global optimization heuristics allow steps that will temporarily decrease the objective function value, in an attempt to escape the local optimum and look for the global optimum, or at least a better local optimum. These global optimization heuristics are often called *metaheuristics* because the procedure used to generate new solution out of the current one, is embedded in a heuristic which determines the search strategy. The main drawback of metaheuristics is that they do not have definitive stopping criteria; the longer the computation time, the higher the probability of finding the global optimum [61].

Initial solution algorithms

Solomon divides VRP tour-building algorithms into either sequential or parallel methods [46]. Sequential procedures construct one route at a time until all customers are scheduled. Parallel procedures are characterized by the simultaneous construction of routes, while the number of parallel routes can either be limited to a predetermined number, or formed freely. Solomon

concludes that, from the five initial solution heuristics evaluated, the *sequential insertion heuristic* (SIH) proved to be very successful, both in terms of the quality of the solution, as well as the computational time required to find the solution.

Initialization criteria refers to the process of finding the first customer to insert into a route. The most commonly used initialization criteria is the *farthest unrouted* customer, and the customer with the *earliest deadline*, or the earliest *latest allowed arrival*. The first customer inserted on a route is referred to as the *seed customer*. Once the seed customer has been identified and inserted, the SIH algorithm considers, for the unrouted nodes, the insertion place that minimizes a weighted average of the additional distance and time needed to include a customer in the current partially constructed route. This second step is referred to as the *insertion criteria*. The third step, the *selection criteria*, tries to maximize the benefit derived from inserting a customer in the current partial route rather than on a new direct route. Note that the terms *nodes* and *customers* are used interchangeably. The insertion and selection criteria can be simplified using the example illustrated in figure 2.9. The partially constructed route in the example consists

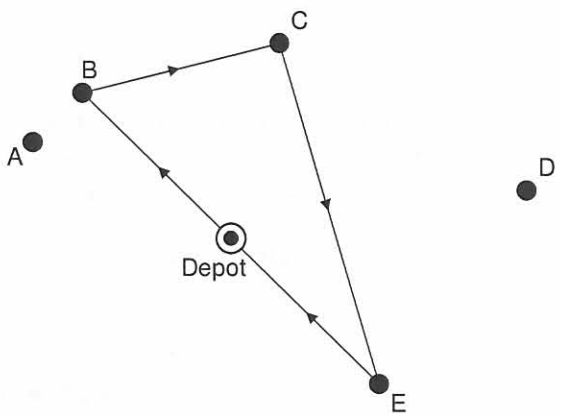

Figure 2.9: Sequential insertion of customers

of the depot and three routed nodes, namely B, C, and E. The route can be expressed as $Depot - B - C - E - Depot$. Nodes A and D are currently unrouted.

The *insertion criteria*, $c_1(i, u, j)$, calculates the best position and associated cost, between two adjacent nodes i and j on the partial route, to insert a customer u, and is calculated for each of the unrouted nodes. Consider node A in the example, there are currently four edges where the node can be inserted, namely $Depot - B$, $B - C$, $C - E$, or $E - Depot$, as illustrated in

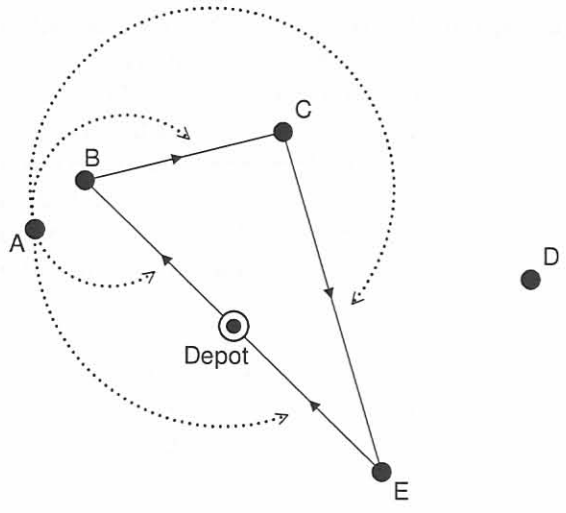

Figure 2.10: Selection criteria

figure 2.10. Dullaert *et al.* [19] extends Solomon's heuristic and determines $c_1(i, A, j)$ for the unrouted node A as

$$c_1(i, A, j) = \min_p [c_1(i_{p-1}, A, i_p)], p = \{1, 2, \ldots, m\} \tag{2.16}$$

in which m represents the routed nodes in the partially constructed route. If the expressions are generalized for all unrouted nodes u, the insertion criteria is calculated as

$$c_1(i, u, j) = \alpha_1 c_{11}(i, u, j) + \alpha_2 c_{12}(i, u, j) + \alpha_3 c_{13}(i, u, j) \tag{2.17}$$

with

$$c_{11}(i, u, j) = d_{iu} + d_{uj} - \mu d_{ij}, \mu \geq 0 \tag{2.18}$$
$$c_{12}(i, u, j) = a_j^{new} - a_j \tag{2.19}$$
$$c_{13}(i, u, j) = ACS, AOOS, or AROS \tag{2.20}$$

With the extension to Solomon's heuristic, the weighting factors α_i need not add up to 1. $c_{11}(i, u, j)$ denotes the additional distance, and $c_{12}(i, u, j)$ the additional time needed to serve customer u after customer i, but before customer j. The new actual arrival time at node j is denoted by b_j^{new} in (2.19). The vehicle savings criteria, denoted by $c_{13}(i, u, j)$, considers any one of three parallel approaches to vehicle cost, where Dullaert *et al.* [19] adapts the savings concepts first introduced by Golden *et al.* [22]. To elaborate on the concepts, let

$F(z)$	be the fixed cost of the smallest vehicle that can service a cumulative route demand of z
$F'(z)$	be the fixed cost of the largest vehicle whose capacity is less than or equal to z
$P(z)$	be the capacity of the smallest vehicle that can service a demand of z
Q	be the load of the vehicle currently servicing the route
$\overline{Q}$	be the maximum capacity of the vehicle currently servicing the route
Q^{new}	be the new load of the vehicle after the customer has been inserted into the route
$\overline{Q}^{new}$	be the (new) capacity of the vehicle after the customer has been inserted into the route

The *Adapted Combined Savings* (ACS) is defined as the difference between the fixed costs of the vehicles capable of transporting the load of the route after, and before, inserting customer u, and is calculated as

$$ACS = F(Q^{new}) - F(Q) \tag{2.21}$$

The *Adapted Optimistic Opportunity Savings* (AOOS) extends the ACS by subtracting the fixed cost of the vehicle that can service the unused capacity, and is calculated as

$$AOOS = [F(Q^{new}) - F(Q)] - F(\overline{Q}^{new} - Q^{new}) \tag{2.22}$$

The *Adapted Realistic Opportunity Savings* (AROS) takes the fixed cost of the largest vehicle smaller than or equal to the unused capacity, $F'(\overline{Q}^{new} - Q^{new})$, into account as an opportunity saving. It only does so if a larger vehicle is required to service the current route after a new customer has been inserted. AROS is calculated as

$$AROS = [F(Q^{new}) - F(Q)] - \delta(\omega)F'(\overline{Q}^{new} - Q^{new}) \tag{2.23}$$

where

$$\delta(\omega) = \begin{cases} 1 & \text{if } Q + q_u > \overline{Q} \\ 0 & \text{otherwise} \end{cases}$$

Any *one* of these savings criteria can be used as all three outperformed previous best published results for the initial solution [19]. Once the best position for each unrouted node has been determined, as illustrated in figure 2.11, the customer that is best according to the *selection criteria*, is selected. The procedure can be expressed mathematically as

$$c_2(i, u^\star, j) = \min_u [c_2(i, u, j)], u \text{ unrouted and feasible} \tag{2.24}$$

$$c_2(i, u, j) = \lambda(d_{ou} + t_{ou}) + s_u + F(q_u) - c_1(i, u, j), \lambda \geq 0 \tag{2.25}$$

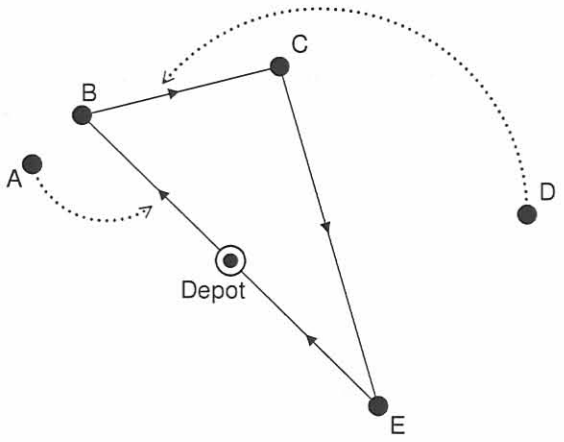

Figure 2.11: Best insertion position determined for each unrouted node

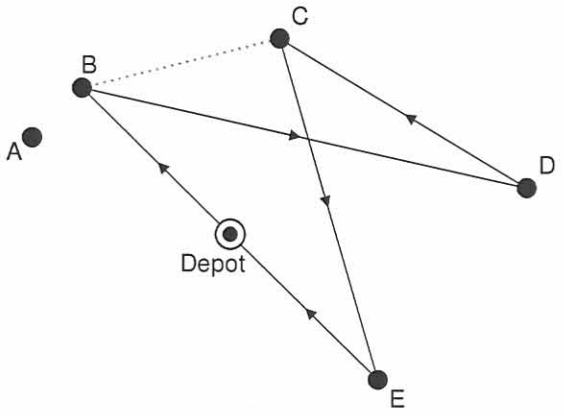

Figure 2.12: New route after inserting best customer

The best customer, $u^{\star}$, is then inserted into the partially created route between its specific nodes i and j. From figure 2.11, consider node D to be the best node. After inserting D into the current route, node A remains the only unrouted node, and the new route is illustrated in figure 2.12, and can be expressed as $Depot - B - D - C - E - Depot$. The insertion process is repeated until no remaining unrouted nodes have a feasible insertion place. A new route is then initialized and identified as the *current* route.

Van Breedam [61] introduces an initial solution parameter in his evaluation of improvement algorithms, and finds that, in most cases, a good initial solution gives significantly better final results.

Solution improvement algorithms

Initial solutions serve as input to improvement heuristics, and although these improvement heuristics are not the main focus of this dissertation, the three main metaheuristics is described briefly.

The evaluation of any heuristic is subject to the comparison of a number of criteria that relate to different aspects of the algorithm's performance. Such criteria include computational time, quality of solution, ease of implementation, and flexibility. Bräysy and Gendreau [10] state that flexibility is important criteria for algorithms that are designated to to be used in real-world problems, as an algorithm should be able to handle changes to the constraints and objective function. There is generally a trade-off between computational time and solution quality. This characteristic is a key feature of metaheuristics.

Tan *et al.* [54] compare the three popular meta heuristic methods with one another and concludes that there is no single heuristic that is generic enough to solve problems for all situations. Instead, they are inevitably problem specific. Each of the three popular metaheuristics has its advantages and disadvantages.

The *Simulated Annealing* (SA) methodology is similar to the annealing process of solids. When a metal is heated to high temperatures it is structurally weak and unstable. If the metal is allowed to cool slowly, it orders itself into a stable, structurally strong configuration. This process is called annealing. Kirkpatric *et al.* [26] first proposed the use of SA for optimization, while Osman [41] applied it to vehicle routing and scheduling problems. The states of the solids correspond to the feasible solution, and the temperature at each state corresponds to the improvement in objective function, with the minimum temperature being the optimal solution. SA involves a process in which the temperature is gradually reduced during the simulation. Often, the system is first heated and then cooled. The system is given the opportunity to surmount energy barriers in a search for conformations with temperatures lower than the local-minimum temperature found by energy minimization. At each step of the simulation algorithm, a new state of the system is constructed from the current state by giving a random displacement to a randomly selected particle. If the energy associated with this new state was lower than the energy of the current state, the displacement was accepted, that is, the new state becomes the current state. This basic step, called a metropolis step, can be repeated indefinitely. The whole procedure is called a metropolis loop. Some of the choices that need to be made with a SA strategy are

- the expression used to calculate the initial temperature,

- the cooling function used for the reduction of temperature,
- the conditions for thermal equilibrium – where the temperature can be lowered, and
- the stopping criteria – usually a number of consecutive degrading values.

SA produces good solutions much faster than other metaheuristics, in terms of computational time, although the solution quality of the final solution is inferior.

Reeves [44] describes the *Genetic Algorithm* (GA) as the intelligent exploitation of a random search that was first presented by Holland [23]. The name originates from the analogy between the representation of a complex structure by means of a vector of components, and the idea of the genetic structure of a chromosome. GA is described as follows:

> *A population of solutions are maintained and a reproductive process allows parent solutions to be selected from the population. Offspring solutions are produced which exhibit some of the characteristics of each parent. The fitness of each solution can be related to the objective function value, in this case the total distance travelled [4].*

Analogous to biological processes, offspring with good fitness levels are more likely to survive and reproduce. Selection of the fittest ensures that fitness levels throughout the population will improve with evolution. The result is a population of chromosomes, in this case vehicle routes, with high performance characteristics. In most applications the component vector, or chromosome, is represented as a string of bits (0 and 1). Tan *et al.* [54] argue the case of replacing the binary digits with integer digits. A chromosome can now easier represent the order in which customers are visited. Genetic operators are used to improve the quality of the population. The most common genetic operators used to manipulate theses chromosomes are crossover and mutation. *Crossover* is the exchange of sections of the parents' chromosomes, while *mutation* is the random modification of the chromosome.

Researchers have explored the adaptations of GA to form *hybrid-GA* [54, 59]. The highest potential for hybrid-GA seems to be when the basic principles of GA are combined with adaptive memory features introduced by Taillard *et al.* [52]. This flexibility of the hybrid-GA makes it a very popular heuristic and holds promising prospects for the application of GA. Thangiah [58] recognizes that GA does not perform well for problems in which customers are geographically clustered together.

The *Tabu Search* metaheuristic (TS), a memory based search strategy, deals with the problem of being trapped at a local optimum by temporarily forbidding moves that would return to a solution recently visited (cycling). The result is that the tabu search heuristic prevents short-term cycling, although solutions can be repeated over a longer period of time. This heuristic was proposed in its present form by Glover [21] and has been applied to many optimization problems besides vehicle routing. A *tabu list* is used to record these forbidden moves, which means that each iteration choose a non-tabu move. After each step, a collection of moves that includes any immediate return to the previous point is added to the tabu list [43]. These recently added moves are not allowed for a fixed number of iterations, but are eventually removed from the tabu list.

As with other metaheuristics, any particular iteration can either improve or degrade the objective function value. It is therefore important to continuously update the best feasible solution found so far, referred to as the *incumbent solution*. When the tabu search is complete, usually after a finite number of iterations have been completed, the incumbent solution is accepted as the final solution: an approximation of the optimum solution. It is important to design the criteria of the TS carefully. The decision of which moves to add or remove from the tabu list is imperative as it has a direct effect on both the accuracy and computational time of the search. If too few moves are disallowed, it may lead to cycling; too many disallowed moves restrict the search from finding superior solutions quickly.

The tabu list contains record of three elements: the list position, the original route and the string of stops moved. The list is implemented as a queue that operated on a first-in, first-out principle. The memory of the tabu list can be *recency* or *frequency* based. The recency based list, called short-term memory, contains the last x number of moves the algorithm has encountered and sets them as tabu, assuming that the tabu list size is x. The frequency based list, called long-term memory, complements the short-term memory by providing the additional information of how many times each tabu move have been attempted. Tan *et al.* [54] propose the use of a multi-functional list structure that serves both purposes of a recency and frequency-based list, and state that frequency-based memory provides better incentive as to the choice of the next move. Van Breedam [61] concludes that the use of long-term memory gives significantly worse solutions for the majority of problems. The impact of these contradictory opinions is that, should TS be used, the easier implementable option of short-term memory is preferred. A way of still retaining long-term memory is to increase the length of the tabu list.

Van Breedam [61] notes that unlike the SA heuristic, the performance of

the TS is highly dependent on the quality of the initial solution. The use of TS requires careful consideration of the mentioned characteristics of TS, as it should be chosen to best represent the problem environment of the project.

TS solutions are generally closest to optimal, but the computational time is about two to three times that of GA, and almost twenty times that of SA. The current computational power of modern computers makes the slower performance of TS less of a problem. Tan *et al.* [54] report that TS was able to solve 56 problem instances, each containing 100 customers, in an average of 1,500 seconds. Carriers furthermore require an operational scheduling system on a daily basis, with real-time scheduling as a future functionality.

2.4 Conclusion

The term *City Logistics* refer to the process of optimizing urban freight movement in a multi-actor environment. Changes occur in the environment, and stakeholders have conflicting perceptions with regards to how the changes will impact them, and their co-stakeholders. It is not contested that fleet optimization, in the form of vehicle routing and scheduling, will have advantages for all stakeholders. Literature indicates that, to address the specific needs and perceptions of logistic stakeholders, more realistic models are required to manage and predict factors influencing urban networks.
Realistic models require the introduction of additional constraints to the solution space and results in problems being harder to solve. The quality of the final solution is impacted by the quality of the initial solution generated through heuristic methods.

Chapter 3 introduces a new approach, *time window compatibility*, to improve the quality of the initial solution. The aim is to have a mechanism that will assist in identifying seed customers in an innovative way during the route building process, and also ease the computational burden.