Université de Lille 1 Sciences et Technologies
Ecole doctorale Sciences Pour l'Ingénieur

Thèse de doctorat

pour obtenir le diplôme de

Docteur de l'Université de Lille 1
Spécialité: Informatique

défendue par Bayrem Tounsi

Contributions à la chaîne logistique e-commerce: Intégration dans l'e-fulfillment et tarification de services de livraison.

Villeneuve d'Ascq, 19 Décembre 2016

Membres du jury :

Directrice de thèse :	Luce Brotcorne	Chargé de recherche, INRIA, France.
Co-directeur de thèse :	Yezekael Hayel	Maître de conférence, Université d'Avignon, France.
Rapporteurs :	Yves Crama	Professeur, Université de Liège, Belgique.
	Patrick Maillé	Maître de conférence, Telecom Bretagne, France.
Examinateurs :	Olivier Peton	Maître assistant, Ecole des Mines de Nantes, France.
	Dominique Feillet	Professeur, École des Mines de St Etienne, France.
	Clarisse Dhaenens	Professeur, Université de Lille 1, France.
	Frédéric Semet	Professeur, Ecole Centrale Lille, France.

Acknowledgments

First, I express my gratitude to my two supervisors Dr. Luce Brotcorne and Dr. Yezekael Hayel. All along the thesis, they provided me support and encouragements. The work with them was stimulating and instructive, in a very good atmosphere.

I would like to thank Pr. Yves Crama and Dr. Patrick Maillé for being part of my committee and reviewing the present manuscript. Their corrections and suggestions helped me to improve my manuscrit. I thank Dr. Dominique Feillet, Dr. Olivier Péton, Pr. Frederic Semet and Pr. Clarisse Dhaenens for being part of my committee.

On the RESPET project, I had the opportunity to work with other researchers. I thank Pr. Frederic Semet, Dr. Dominique Quadri and Dr. Diego Cattaruzza. I thank Pascal Olivier for his cooperation and for organizing a visit to DHL warehouse.

During my thesis I was pleased to meet and know numerous people. I thank my friends and colleagues in INOCS and DOLPHIN teams, and in LIA. I wish to all of them success. I would like to express my warm thanks to Diego, Maria and Maxime for their help and support during my defense's rehearsal.

I finally thank Inria personnel, particularly Julie and Aurore, for their kindness and their availability every time I had a request.

Contents

List of Figures

List of Tables

Title: Contributions in the e-commerce supply chain: Integration in e-fulfilment and delivery services pricing.

Abstract: All over the world, the growth of e-commerce has led to an increasing importance of the inherent supply chain. This thesis is dedicated to the study of two phases of the e-fulillment process. In the first part, we focus on picking and shipping operations conducted in the warehouse at the e-fulfilment process uphill. We propose a global model based on picking and shipping coordination, and tactical-operational integration. The solution method proposed is based on decomposition of the problem into three phases and it was shown to significantly outperform commercial solver. Then, we propose a second model based on a mechanism with rolling horizon that captures the information knowledge and decision making dynamics. We present deterministic and stochastic approaches incorporated in the rolling horizon procedure in order to find good policies under demand uncertainty.

In the second part of the thesis, we study a last mile delivery system offering two classical type of services: home delivery and pick up at relay station. We address a service pricing problem that takes into account the customers behaviour. Customers are sensitive to the tariff of a delivery service, but also to its quality. We propose a bi-level model where at the upper level, the provider control the services tariffs. At the lower level, users react by choosing their delivery service according to a utility function which includes the provider tariff and the perceived congestion. We model the customers reaction using a nested logit model and compute the resulting stochastic user equilibrium. Based on a local search that exploits a sensitivity analysis of the equilibrium choice probabilities, a new heuristic algorithm for the bi-level services pricing problem is proposed and compared to others existing approaches.

Keywords: E-commerce logistics, coordination, integration, matheuristic, rolling horizon, stochastic, bi-level, discrete choice models, heuristics.

Chapter 1
Introduction

In developed countries, e-commerce logistics represent the latest big driver of change in logistics and physical distribution networks. The logistic industry has been innovating and mutating in order to follow the continuous growth of e-commerce. In France, for example, on-line sales have risen by 20% in 2013 and by 11% in 2014[1].

E-commerce logistics has evolved substantially over the past 40 years or so, impacting the organisation of the supply chain, involved actors, interactions between them and processes. Since the beginning of the 21^{st} century, e-commerce began to rapidly expand with pure-play (internet only) retailers leading the way in developing e-fulfillment distribution networks [2].

In this thesis, we are interested in e-commerce business that generates physical flows, where the supply chain efficiency is a crucial success key. Along with the maturation of the on-line market and the generalized use of information and communication systems, the e-commerce supply chain has to meet new challenges related to all levels of infrastructure design, resources management and operations optimization.

This introduction describes first a major actor in e-commerce supply chain; the third party logistics. Then, general description of the e-commerce supply chain is given. Finally, the contributions of the thesis are depicted.

1.1 Third party logistics

In traditional commerce, customers buy and pick their goods in a shop. In e-commerce, customers order on-line and goods are delivered to them. What happens during the time a customer waits for his goods is crucial to his overall experience. The challenge of e-fulfillment is that many undesirable events can occur between the time an order is placed and when the customer receive the goods. Retailers improve continuously their business to offer good products at attractive prices, but they can miss the order, not have inventory to ship, enter a wrong shipping address. Not to mention that the package may never get there or the item can be damaged when it does arrive.

Some of these mistakes can be eliminated when using a service whose logistic is the core competence and that specializes in *the e-fulfillment defined as the steps*

[1] http://www.fevad.com/espace-presse/les-ventes-sur-internet-en-hausse-de-16-au-2eme-trimestre-2013

[2] http://cerasis.com/2014/04/30/e-commerce-logistics/

involved in receiving, processing and delivering orders to end customers. The development of e-commerce induced the emergence of third party logistics providers (3PL). A third-party logistics provider is a firm which provides outsourced logistics services for part, or all of its customers' supply chain management functions [3]. Services include warehousing and distribution, but also administration and customs procedures. The relation between a 3PL and its clients can vary from providing the basic logistics activities to a full integration of logistics function. For the clients, this organisation results in cost and time saving, low capital commitment, focusing on the core business and flexibility.

The global 3PL market reached 750 billion USD in 2014, and grew by 157 billion USD in the US while demand growth for 3PL services in the US (7.4%) outpaced the growth of the US economy in 2014 [4].

E-fulfillment services help retailers provide a better customer experience by offering faster shipping times, guarantee well-packed items, consistent on-time delivery, and an easy return process. Nowadays, the fulfillment is a competitive advantage. Retailers with a reliable and efficient fulfillment process gain a competitive advantage in the market.

The 3PL, like our project partner[5], offers a global support dedicated to e-commerce shop, including the conception of logistic sites and equipments, the modelling and the optimization of process, the management of demand variation and peaks, etc.

1.2 E-commerce supply chain

In commerce, supply chain management (SCM), the management of the flow of goods and services, involves the movement and storage of raw materials, of work-in-process inventory, and of finished goods from point of origin to point of consumption. Interconnected or interlinked networks, channels and node businesses combine in the provision of products and services required by end customers in a supply chain. Supply-chain management has been defined as the design, planning, execution, control, and monitoring of supply chain activities with the objective of creating net value, building a competitive infrastructure, leveraging worldwide logistics, synchronizing supply with demand and measuring performance globally [6].

All along the different phases thousands of decisions and more are taken related to planning and operations management. Classically in SCM, three levels of decisions are defined [Anthony 1965]:

-*Strategic level*: includes long term decisions that establish the bases of the

[3]https://en.wikipedia.org/wiki/Thirdparty_logistics

[4]http://www.3plogistics.com/big-deal-2014-3pl-results-and-2015-estimates/

[5]http://www.dhl.fr/fr/logistique/solutions_de_chane_logistique/ce_que_nous_faisons/e-fulfillment.html

[6]https://en.wikipedia.org/wiki/Supply_chain_management

development of the supply chain, like for example the design of the supply chain structure.

-*Tactical level*: includes mean term decisions that deal with the organisation of regular activities, like for example the quantities of flows and resources.

-*Operational level*: includes short term decisions that gives detailed instructions for immediate execution, like for example jobs scheduling.

The issues of SCM are different as we are in a *Business to Business* (B2B) case or a *Business to Consumer* (B2C) one, regarding problem characteristics and service requirements. The durations of decisions vary from an application to the other. They depend on many parameters like data acquisition frequency or resources' characteristics. In e-commerce for example, operational decisions are set every hour since orders are made continuously and need to be processed. While tactical decisions like the number of temporary workers to use can be set at daily base.

The e-commerce supply chain (ESC) is a complex system that involves a wide and multi-modal transportation network and manages several physical flows. Several studies highlighted the strengths needed for efficient ESC : **reliability**, **flexibility** and **profitability**.

1.2.1 E-fulfillment network

The design of the distribution network is the core of big investment and the distribution strategy. The facilities location have to find trade-off between costs and delivery efficiency.

Figure 4.1 depicts an example of an e-fulfillment network. Traditional e-commerce activity is based on a warehouse that stores all products and cover a wide geographical area (a country or even more). We detail in chapter 2 the structure and the operations of the warehouse.

The transportation network is structured with delivery centres which handle the final delivery. These centres are the starting point of the last mile delivery segment that ends at the customer's home or at a designated collection point.

Shoppers online want to choose the location of the delivery. Last studies shows that 66% of french e-shoppers prefer the delivery at a relay point [7]. The relay point is usually a shop that makes a partnership with the retailers to offer the possibility to receive and deposit delivered goods so that end customer can pick them later.

1.2.2 Transportation in e-commerce

In e-commerce, the main physical flow is the fulfilment flow, or direct, flow. It insures the replenishment and the delivery of online ordered goods. The direct flow is divided in two physical flows : upstream and downstream [Agatz 2008].

[7] http://www.fevad.com

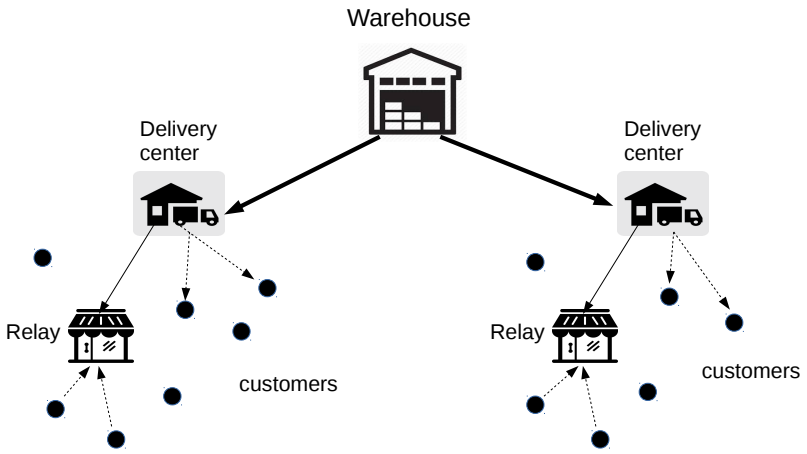

Figure 1.1: The e-fulfillment network

The upstream flow replenishes the warehouse by products coming from providers to be stocked in the warehouse. These operations are usually done long time before products are up for sale online. We do not consider this flow in this thesis, and even if it is part of the fulfilment, we will use fulfilment to refer to the downstream flow.

As in any supply chain, transportation is more than crucial in e-fulfillment. Once shipped from the warehouse, parcels make a long journey to reach the end customer. The journey is made of consecutive segments and combines different transportation modes. In e-commerce transportation includes also the management of the returned parcels.

1.2.2.1 Long-haul transportation

The downstream flow is initiated every time an order is received by the warehouse. The order is picked and the formed parcels are shipped from the warehouse to be delivered to the customer.

The warehouse is designed to cover customers scattered in a wide geographical area. Parcels shipped to the same region are consolidated in pallets and shipped according to a transportation mode. The first segment is the transportation from the warehouse to a regional delivery centre. Transportation is usually performed by *Less-Then-Truckload* carrier. The long haul freight transportation generally analyses multicommodity transportation systems at a regional, national or global level [Crainic 2003]. It combines several modes including rail, truck, ship, etc. The performance of such system is evaluated regarding different, and sometimes conflicting criteria like cost, safety and pollution. It rises strategic design decisions related to the network and services design involving big investments. It also includes planning and management issues related to resources allocation and journeys' scheduling.

1.2.2.2 Last-mile delivery

The second e-fulfilment segment is the transportation of parcels from the regional centre to end users. At the regional delivery center, parcels are shipped directly to end users home or dropped in relay stations, generally shops. The second segment is then the last segment in direct contact with the end user. Generally, distances done in this phase are small compared to the previous phase. Thus we speak about *Last-Mile delivery* problem. For both delivery options, trips need to be planned to serve a set of destinations. But the delivery problem is different as the option is home delivery or at a relay station. For home delivery a trip serves a big number of customers, scattered and with little delivered parcels (often only one). Inversely, a trip for delivery to a relay station includes a limited number of depots with many delivered parcels. Moreover, home delivery must be done during a time window defined by either the delivery company or the customer. While the delivery at a relay station is free of such constraint. Thus, the delivery resources and the planning of the trip are specific to each option.

Home delivery can be addressed as a variant of the *vehicle routing problem* with a fleet of vehicles and customer time windows like in [Cattaruzza 2015a]. Vehicles can perform more than one trip. The planning deals with the clustering of customers into trips and the routing of each trip. Regarding drivers, trips must respect the maximum legal driving time per day and the legal time breaks.

Delivery at relay station is close to city's freight distribution systems or city logistics. City logistics rely on consolidation to provide the best trade-off between operations efficiency and resources utilization [Crainic 2009]. At the delivery center parcels are sorted by relay station, then vehicles load are composed to serve a set of stations.

1.2.2.3 Returns

For online customers, the ability to easily return purchased items is an important part of the buying decision. For online sellers, the ability to effectively handle these returns is critical to customer satisfaction. Customers should know how and where to send returns, while the e-fulfilment service should set how returns will be handled once they arrive back at the warehouse.

Customers can return items in different ways. They can post them back to the warehouse and then get the shipping cost refunded. Or they can drop it at a relay station, and the e-fulfillment company handles the transportation of returns to the warehouse. The management of the return flow in e-commerce leads naturally to additional transportation issues that are close to what is known as *Reverse Logistics*. Advanced systems combine the delivery with returns by planning a pick up and drop trips.

1.3 Contributions

Our contributions aim to give a general understanding of e-commerce supply chain and also to design efficient tools for e-fulfilment management. We focus particularly on two issues:

- The coordination and integration of decisions in e-fulfilment management.
- The design of last-mile delivery services under customer choice.

More precisely, our **first contribution**, presented in chapter 2, deals with warehouse management, by investigating the benefits of two capital innovations. The first is the coordination between order picking and shipping schedules. These two phases are classically done separately and sequentially, which leads in high activity to strained situations. The proposed model is an approach for enhancing the supply chain **flexibility** by considering alternative actions like order postponement or shipping mode change.

The second innovation of the model is the integration of tactical and operational decision over a multi-period planning horizon. In fact the optimization of processed quantities enables smoothing peaks and better resources usage, but ignoring the operational constraints results in infeasible solution. The tactical-operational integration is a safe approach for a better resource usage and consequently global cost reduction and **profitability**.

The proposed model is computationally challenging. Several costs of different nature are taken into account including labour and trailers. Also there are penalties related to orders postponement or mode changes. The model looks for a planning through both tactical and operational decisions. It incorporates a large set of constraints related to order picking and to shipping schedules. The classical solution methods fail to provide an optimal solution. We then propose an advanced heuristic method capable of providing good solution at reasonable computation time.

The **second contribution**, presented in chapter 3, focus on a main difficulty of e-fulfillment management related to the uncertainty of demand. In practice, a planning decision aims to take decisions related to future activities based on data. Although forecast models are usually used for a-priori data estimation, there are always errors compared to a-posteriori revealed data. Those errors can be more or less important and deteriorate the optimality of the planning or lead to constraint violation. Such situations are of course undesirable since they affect seriously the **reliability** of the supply chain.

We introduce a second model with rolling horizon that reflects better the information acquisition process. The model also enables us to include stochastic information of future periods in different fashions. We evaluate the performance of stochastic methods and deterministic policies with classical lower and upper bounds.

The **third contribution**, presented in chapter 4, studies last-mile delivery system with two family of services: home delivery and pick up at relay station. In this part of the e-commerce supply chain several decision actors interact. There are two interactions. The first one is between the delivery company and the customers. The second interaction is between the customers themselves through congestion effect.

The originality of our contribution is twofold. The first is the formulation of company-customers interaction as a bi-level model. At the upper level, the provider controls services' tariffs. At the lower level, customers react by choosing their delivery service according to a utility function. The second originality is the integration of customers behaviour. In addition to tariffs, the utility function includes a congestion measure depending on the service. Due to correlation between services of the same family we use a nested Logit model and compute the resulting stochastic user equilibrium. A sensitivity analysis of the SUE is then conducted, it gives explicit expression of the derivatives of customers' decisions with respect to services' tariffs. Based on a local search that exploits the derivatives information, a new heuristic algorithm for the bi-level services design problem is developed and compared to others existing approaches.

The e-commerce supply will continue to attract research and we expect more innovation in e-commerce logistics regarding demand forecast, delivery and real time order tracking.

Chapter 2

Integration in E-fulfilment optimization

The works presented in this chapter were published in the conference Odysseus 2015 [Cattaruzza 2015b].

Contents

2.1 Introduction

An e-commerce business is based on a warehouse that stores all products and covers a wide geographical area (a country or even more). The warehouse is the heart of e-commerce activities where upstream supply chain phases, replenishment and storage, and downstream supply chain phases, orders picking and shipping, are closely conducted [Agatz 2008] [Gu 2006].

In this chapter we focus on downstream process in order to investigate the potential advantages on cost reduction of jointly scheduling orders picking and shipments. Main costs of e-fulfillment are related to resources required for performing orders picking, workers, and for shipping, trucks.

In a standard organisation, where picking and shipping are not jointly determined, activity peaks results in bottleneck situations. To handle such situations temporary workers are usually hired to enhance the permanent team. Such policy is however costly. Thus we define a new global approach based on the integration of resources determination and operation planning. The latter includes the coordination between picking and shipping.

In addition, different delivery options, called modes, are offered to customers. Each mode is associated with different trucks that guarantee different delivery times (packages are delivered within 3-5 business days for the normal mode, and within 24 hours for the quick mode).

The coordination enlarges the set of possible actions that improve the process flexibility and resources productivity. First, orders can be postponed and prepared during a day later than the one of arrival. Second, an order is allowed to be assigned to a mode different from its default one. Our approach looks for a global efficient planning that takes into account the impact of postponements and mode changes on the process of future demand.

In section 2.2 we describe the e-fulfilment process, and we give a state of the art of related works in section 2.3. A mathematical formulation of the integrated e-fulfilment problem is introduced in section 2.4. The section 2.5 is dedicated to decomposition based matheuristic proposed for solving the problem. In section 2.6 we describe the numerical experimentations and comment the results. Finally, we conclude the chapter in section 2.7.

2.2 Warehouse description

The warehouse is at the heart of an e-commerce business. This platform requires a flawless organisation including replenishment, inventory management, order picking and shipping. Warehouse management systems are continuously elaborated and optimized in order to improve efficiency and profitability and meet the online sales growth. We can distinguish three principal zones in a classical warehouse: storage zone, packing workshops and shipping docks (see figure 2.1).

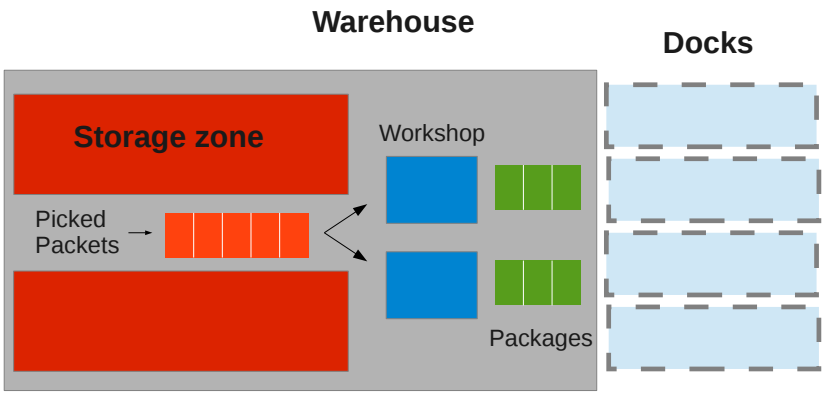

Figure 2.1: The warehouse structure

We now describe the main phases of e-fulfilment that are carried out in the warehouse.

2.2.1 Replenishment

The replenishment of the warehouse is the first phase of e-fulfillment. Goods are received, counted, inspected, labelled, indexed in the warehouse management software and stored. Replenishment is optimized based on inventory management theory to find the good trade-off between reactivity and profitability. The problem of managing the inventory can be addressed for example using multi-echelon theory [yu Kevin 2005], [Alptekinoglu 2005]. It aims to determine the base-stock levels and the warehouse replenishment (from providers) strategies in a way that minimizes transportation, inventory handling and backordering costs.

E-commerce enables shoppers to make orders at any time and from anywhere and thus it helps increase the sales. As a consequence, the inventory management is more challenging. After replenishment, the warehouse is ready to receive and process orders. In this thesis we do not study e-commerce replenishment issues, we rather focus on the downstream flow that deals with the process of orders.

2.2.2 Order picking

When a customer orders a set of items, the latter are picked off the shelves by one of the workers and transported to an order packing workshop. There, a worker makes a quality control, selects the packaging materials, scans items, adds protection and bill, seals the box, and then moves it to one of the shipping dock.

Different order picking methods can be employed in a warehouse, for example single-order picking, batching and sort-while-pick, batching and sort-after-pick, single-order picking with zoning, and batching with zoning [Gu 2006]. Many e-commerce retailers outsource their order fulfilment, as they simply don't have the

in-house experience, expertise or technology to process orders efficiently.

2.2.3 Packages shipping

Once an order has been picked and packed, it is moved to a shipping dock and loaded in the corresponding truck. The shipping truck depends on the delivery option (quick/ normal). In addition, some goods, like fragile or big ones, are transported by a specific service. Also for a given mode, different levels of service are generally offered, since some customers are ready to pay more if they can receive their goods earlier. The 3PL is responsible for efficient journey's design for each order under tariff and delay commitment. The 3PL manages several transportation contracts with different partners, each of them having its own conditions regarding tariffs and volumes. This particular environment, that we call multi-mode shipping, impacts the warehouse structure, order picking process and also the use of human resources.

2.2.4 Picking and shipping coordination

Traditionally order picking and shipping are done separately. Orders are processed in the order of their reception, while a zone close to the docks is dedicated to packages before shipping. This situation may lead sometimes to bottlenecks in high activity period. A problem in an order's process, or related to a shipping truck impacts the overall activity.

This leads us to suggest the coordination of the two phases. Coordination can be defined as the simultaneous consideration of two or more processes or functions, that are usually sequential and interdependent, in order to improve global efficiency.

Performing jointly order picking and shipping leads in some situations to improve resources use regarding the demand configuration. The use of resources at the picking phase (workers) can be planned with respect to the available shipping resources (trucks) and their departure schedule. In this way, a situation where picked orders can not be shipped because of missing truck's capacity can be avoided.

When the decision of picking an order is taken, simultaneously the shipping is scheduled. On the ground, the loading of trucks is jointly done with the order picking. This implies that the two phases are more connected and inter-dependant. For example, the picking of an order cannot start unless a truck of the assigned mode is at docks.

The coordination opens a field to improve flexibility and resource use. Order processing is no more a simple first in first out process without taking into account global needs and resources. Coordination enables consolidation and helps avoiding extreme and difficult situations. For example an order can be postponed if it is costly at the moment, an order can be switched to a different shipping mode from the one selected by customer. Some of these decisions may not be costless, they also can cause customer disappointment, but they lead to improve the overall performance.

2.2.5 Resources optimization

As in every business profitability comes with optimization of resources. The resource levels should lead to the best trade-off between customers satisfaction and costs. In some application, when resources are scarce, an order can only be partially satisfied. This can be acceptable by the customer. In e-commerce, an unmet order results in the loss of customer, and a negative image of the retailer. On the other hand, an unused resource or a resource not used at its best productivity is a wasted profit.

In e-fulfilment the demand is known a very short time before the start of processing and it fluctuates from day to day. This aspect highlights the specificity of e-fulfilment with respect to production problems or transportation problems where demand is known in advance.

Resources required in e-fulfillment are mainly *workers* and *trucks*. Workers are in charge of the picking and shipping operations. They are of two types: permanent and temporary. The number of permanent is known in advance. Temporary workers can be hired to work at least for one shift that represent a fixed number of hours. Temporary workers have a higher cost than permanent ones, and they are used to handle activity peaks. The average productivity of a worker is computed as the average amount of order fulfilled by a worker in an hour. Picking requires more time than packing and loading, and the number of picked products is a linear function of the number of workers. We thus assume that an increase in demand would increase the total number of workers linearly.

In the problem studied in this section, the number of trucks is determined for each shipping mode. Traditionally the number of trucks per day is set in advance in a contract over a given term. It is crucial to determine appropriately the number of trucks. Undersized truck capacity would induce unmet demand. Since the shipping capacity bounds the quantity of packages picked, undersized capacity can also result in an underuse of workers. On the other side oversized capacity generates high costs. In extreme cases, it is possible to increase *ad hoc* the number of trucks at a prohibitive additional cost.

After describing the main operations in picking and shipping phases, and highlighting difficulties and challenges, we introduce in the following section our approach that provides a global framework for e-fulfilment optimization.

2.2.6 The integrated picking and shipping problem

We highlighted earlier that the success of an e-fulfilment system depends on its flexibility and profitability. Our approach aims to improve these criteria by coupling two features: the integration of tactical and operational decisions and the coordination between picking and shipping phases.

The integration considers simultaneously tactical and operational decisions. The first are related to the determination of the number of workers and the number of trucks. The second include orders processing and trucks moves (docking and

undocking). We use later IPSP to refer to the integrated picking and shipping problem. The problem is defined over several days to allow postponement of orders process.

The tactical-operational integration takes into account picking and shipping process in the resource allocation. Indeed considering only the global demand to compute the required resources may lead to the violation of some constraints and consequently to an infeasible planning. For example the manager needs to know the number of workers at each shift and not just the workforce corresponding to the global demand. Also all orders assigned to a mode have to be processed before a fixed departure time slots. In addition, docks are at a limited number. So it is necessary to control the movement of trucks and to know docks state (busy or free).

Looking at operations process, orders contain one or more items and they generate one or more packages. The process of one order can be divided and partially made during different time. Thus order process is addressed by a lot-sizing model. More precisely we consider a planning horizon of several periods where a period represents a working day. A period is composed of T time slots, a slot representing a duration of one hour in practice. For workers management, a period is also divided into disjoint and consecutive shifts. During each shift, a given number of permanent and temporary workers are in charge of processing orders. Each period is associated with a given number of orders. The process of an order is to determine the exact number of packages that are prepared at each slot, for every order of every period. A demand with a volume greater than one can be prepared during different and not necessarily consecutive slots.

As new packages are made they are immediately loaded in trucks. The packages of one order can be made during different slots and loaded in different trucks, but they have to be assigned to the same delivery mode. The coordination between picking and shipping implies that when packages are made, a truck of the corresponding mode is present at the docks to be charged. The warehouse is equipped with a limited number of docks that bounds the number of trucks that can be simultaneously charged. Each truck contains only packages of orders associated with a specific delivery mode. Then, if different modes are involved at a given slot, at least one truck for each mode has to be present at the docks. Moreover, at each slot the number of packages that are processed takes into account the total workers productivity the total trucks capacity.

The coordination between picking and shipping in IPSP enables mode change. Indeed, it is possible to deliver an order with different mode than the one chosen by the customer. In that case a penalty needs to be paid. This penalty represents the extra cost needed to deliver the order in the case the change is from the normal mode to the express mode. It quantifies the dissatisfaction of the customer for a delivery delay in the opposite case. In addition, as the problem addressed is multi-period problem, the possibility to postpone the process of an order to a future period is also allowed. As a counterpart, a penalty is paid when the postponement causes a

delay. When an order is postponed to a future period, all its packages are made during that period. Mode change and order postponement occur when they improve resources use. The two strategies can also be combined, when an order associated with the normal mode can be delivered with the express mode of the day after.

In summary, IPSP looks for an operational plan that minimizes the total cost of orders picking and shipping over a number of periods computed as the sum of workers (fixed and temporary) cost, trucks cost and the penalties generated by mode changes and order postponements. In addition, considering that the number of docks is a critical resource, the solution of IPSP minimizes their total use computed as the sum of occupied docks at each slot.

2.3 State of the art

In this section a revue of literature related to IPSP is presented. We focus on lot-sizing problems, coordination problems and integration problems.

2.3.1 Lot-sizing

The joint picking and shipping process studied in this chapter is close to capacitated multi-item lot sizing problem classically used to model and solve production, manufacturing and inventory problems [Karimi 2003] [Gicquel 2008] [Lee 2013]. In its general form the lot-sizing problem aims to determine the quantities of one product or of a set of products to meet the demand at each period of an interval of time, having that no shortage is allowed. The solution finds the optimal trade-off between production costs, set-up costs and inventory costs. Producing large quantities reduces setup cost but in the same time increases inventory costs. The problem can be uncapacitated or capacitated when a constraint on the total quantities produced at each period have to remain under a bounding level. The capacitated lot-sizing problem (CLSP) is known to be NP-hard [Karimi 2003].

As in lot-sizing, the solution of IPSP determines periodically produced quantities. However, the difference is about the decisions frequency which is higher in IPSP. Another common point between CLSP and IPSP is the existence of flow conservation constraints. In CLSP, these constraints insure transition between production and inventory, while in IPSP, the balance is between production and loading.

IPSP differs from CLSP in that several capacities are involved and they are variable. In [Helber 2013] a variant of lot-sizing is considered where in addition to an available capacity, an overtime is added when needed.

When demand shortage is allowed, it is possible to delay the process of a part of the demand. This usually adds a backlog cost in the objective function and the resulting problem is more difficult to solve. An other feature that affect the modelling and the complexity of CLSP is the setup cost structure. For example,

a carry-over setup cost is only applied if the product was not produced in the last period. In some application the setup cost depends on the production sequence.

Lot-sizing problems involves cost function with different structures. The initial CLSP considers only costs with linear functions. While the costs in [Lee 2013] include a piecewise linear function and a piecewise constant function. The complexity of the problem and the solution approach depends on the structure of the considered costs.

The IPSP consider three main costs: the cost of workers, the cost of trucks and the penalties. The cost of workers and the cost of trucks are piecewise constant functions of the number of processed orders. Each function has a particular structure depending on the resource characteristics (unit cost, capacity, productivity). Such resource cost are classically used in CLSP. The cost of penalties function is a linear function of the postponed orders and it is similar to backlogging cost.

2.3.2 Functional coordination

Coordination is a key function in modern management. A big company for example is composed by a set of departments, and each one is in charge of one phase of the final output. Phases are sequential and interdependent. In IPSP for example, for a package to be loaded for shipping, the corresponding order has to be picked before. Thus it is necessary that the planning of one department activity takes into account some informations from dependant departments to insure a satisfying fluidity that is beneficial for the global process. Typically a person or a group, a coordinator, is in charge of insuring the planning of activities and the communication between departments. Coordination is as well addressed as an optimization problem where two or more planning, corresponding to different process phases, are jointly determined taking into account inter-dependencies and optimizing a global objective.

The terms of synchronization or integration are also used to refer to coordination. Although in this thesis we make a distinction between coordination and integration. Our understanding for the latter is detailed in the next section 2.3.3.

In [Chen 2010] author survey works on production and distribution coordination. This problem can be identified in many applications where production and distribution are consecutive in a short time: e-commerce, perishable products, computer, etc. In [Z.L. Chen 2005], the production planning, modeled as a scheduling problem, is coordinated with a distribution problem, modeled as a vehicle routing problem.

In [Baptiste 2008] production is coordinated with shipment schedules: production lines prepare batches that are subsequently charged into trucks and shipped to final destinations. The problem consists in planning the production in order to ship full truck and minimize the delivery costs.

In [Wang 2005] the problem is to coordinate production scheduling of jobs with the selection of delivery mode. The objective is to minimize delivery cost under

job delivery dates. the production stage is modeled as a scheduling problem on one machine. The delivery stage consider that jobs can be immediately shipped after production, while in other application, shipping can only starts at fixed dates.

In the problems addressed in [Cetinkaya 2005] [Lee 2003] inventory replenishment is coordinated with outbound shipments schedule in a multi-period problem. The optimal solution specifies how often, and in what quantities, the stock should be replenished at the warehouse, and how often an outbound shipment should be released so that transportation scale economies are realized and customer requirements are satisfied at a minimal cost. Transport costs are the cost of used cargo and penalties due to early delivery and late delivery. IPSP differs from the cited works in that coordinated phases, order picking and shipping, are modelled with higher level of detail, at the operational level.

2.3.3 Decisions Integration

The second main feature of the proposed approach in this chapter is decision integration. Decisions in e-fulfilment, as it is the case in many planning problems, are of three levels. Classically, first, strategic decisions are determined and implemented for a long term. Then, with a higher frequency, mean term tactical decisions are taken. Strategic and tactical decisions define the environment in which operational decisions which execute concretely the process are determined on a daily or hourly base. Decision integration is the joint determination of decisions of different levels. Decisions integration is different from functional coordination in that the first is vertical while the second is horizontal.

An integrated problem includes for example strategic and tactical decisions or tactical and operational decisions. Since IPSP belongs to the second case, we limit the literature revue here to tactical-operational integration. Tactical decisions in IPSP are related to the resources: workers and trucks. They include the number of truck for every mode and the number of permanent and temporary workers for every shift. While the operational decisions include at every slot processed orders and truck moves in the docks.

Tactical-operational integration in planning problems attracts increasing interest as an approach to match resources with process requirements. It can be motivated by demand uncertainty which makes resources design a complex issue. Indeed, one traditional strategy of coping with demand's fluctuation is to build up inventory during periods with low demand and meeting the demand in excess of production capacity from inventory [Atamturk 2001]. For various reasons (inventory costs, product value, ...) it is excluded for some companies to carry an inventory. Moreover in modern labour market, it is possible to use interim workers to enhance long term contracted workers during activity peaks making possible the practice of dynamic capacity adjustments.

In such environment, companies look for efficient planning that integrates capac-

ity allocation, subcontracting, production and inventory. This leads to integration of tactical and operational decisions. Pac et al. [Pac 2009, Alp 2006] studied inventory management with dynamic continuous capacity adjustments for handling the fluctuations more effectively. In [Pac 2009] authors address a dynamic approach for integrated problem involving tactical decisions (permanent workforce size and contracted number of workers) and operational decisions (temporary workers and produced quantities). A particular situation where a delay needs to be respected before temporary workers are available is studied in [Mincsovics 2009].

Pinker and Larson [Pinker 2001] develop a model for flexible workforce management in environments with uncertainty in the demand and in the supply of labor. The idea is to determine first regular workers and contingent workers levels over all the planning horizon. Jointly at each period, the allocation of temporary workers and overtime are decided to handle the stochastic demand under possible regular workers absenteeism.

An application of joint workforce planning and operations management in mail treatment can be found in [Judice 2004]. The process of mails through a sequence of units is determined using short slots of time, while the staff planning is determined using shifts composed by a number of slots.

2.4 Problem definition, notation and model

In this section, the IPSP is mathematically formulated. We consider a planning horizon of H periods, a period representing a period in practice, indexed in $\mathcal{H} = \{0, \ldots, H-1\}$. For each period $h \in \mathcal{H}$ we need to process a number of orders D_h (indexed in $\mathcal{D}_h = \{0, \ldots, D_h - 1\}$). Orders revealed on period h need to be prepared in one of the following $\bar{H}$ periods, indexed in $\bar{\mathcal{H}} = \{0, \ldots, \bar{H} - 1\}$. $\bar{h} = 0$ indicates that orders are not postponed. There are V available delivery modes, indexed in $\mathcal{V} = \{0, \ldots, V-1\}$.

Each order $d \in \mathcal{D}_h$ of period h is characterized by its

- volume vol_{hd}, the number of packages it is composed;
- mode v_{hd};
- time slot at which the order becomes known (release date of the order) r_{hd};
- penalty $p^{\bar{h}}_{v_{hd}v}$ for processing the order at period $h + \bar{h}$ and assigning it to mode v. $\bar{h} = 1$ corresponds to a postponement while $v_{hd} \neq v$ corresponds to a mode change ($p^0_{v_{hd}v_{hd}} = 0$).

We assume that the penalty for postponing a package from a period h to a period $h + \bar{h}$ or changing its mode is identical for all orders.

Each delivery mode v is characterized by its

- departure slot t_v. No truck associated with mode v will be available after t_v.

All the trucks have the same capacity Q and the same cost c^{truck}. Moreover, $N_{\max}$ is the number of available docks at the warehouse, thus at most $N_{\max}$ trucks can be simultaneously loaded.

Each period h is divided into the same number of shifts S, and each shift into the same number of slots $\bar{T}$. It follows that each period is divided into $T = S\bar{T}$ slots. Each shift s of period h is characterized by its

- starting slot $start_{hs}$;
- ending slot end_{hs};
- cost for a permanent worker c_{hs}^{per};
- cost for a temporary worker c_{hs}^{temp};
- number of packages a permanent worker can prepare $prod_{hs}^{per}$;
- number of packages a temporary worker can prepare $prod_{hs}^{temp}$;
- maximum number of permanent workers $e_{hs}^{\max}$.

Trucks are managed according to the following *truck movement policy*. Each truck is assigned to one and only one mode and will distribute only packages assigned to that mode. Trucks can be made available at the docks at any slot. When a truck gets full during a slot, it is undocked by the end of that slot, and the dock it has occupied becomes free for use at the beginning of the next slot. Thus if a truck associated with a delivery mode gets full during a slot, and in the same time packages continue to be processed and assigned to that mode during the same slot, a new empty truck (or more) is (are) docked, and at that slot more than one dock are used by the same mode. If the truck is not fully loaded at the end of a slot, it remains docked for the next slot. Non-full trucks for mode v are undocked in two cases: at slot t_v or if no more packages for mode v will be assigned to the corresponding mode during the following slots of the period. This policy achieves the least docks occupation having that an undocked truck will not be docked again.

Over the planning horizon, IPSP aims to determine the number of workers and trucks, an order process planning that consists in identifying the exact slot during which each package of each order is processed, and a truck management planning (i.e., when to dock and undock truck) in order to minimize the sum of the workers cost, trucks cost, penalties, and the docks occupation.

We introduce now the variables of the model. For each period $h \in \mathcal{H}$, for each $\bar{h} \in \bar{\mathcal{H}}$, for each $d \in \mathcal{D}_h$, for each mode $v \in \mathcal{V}$ and for each shift $s \in \mathcal{S}$ we have:

- The tactical variables:

– z_{hs}^{per} the number of permanent workers working on shift s of period h;
– z_{hs}^{temp} the number of temporary workers working on shift s of period h;

- and the operational variables:

– $x_{hd}^{\bar{h}v}$ equals 1 if the order d of period h is prepared in period $h + \bar{h}$ and affected to mode v, 0 otherwise;

– y_h^{vt} equals 1 if the number of empty trucks for mode v during period h at a slot $\bar{t} \geq t$ is not null, 0 otherwise;

– $f_{hd}^{\bar{h}vt}$ indicates the number of packages of order d prepared in slot t of period $h + \bar{h}$ assigned to mode v;

– w_h^{vt} is the number of docked trucks for mode v at period h in slot t;

– u_h^{vt} is the number of empty trucks for mode v that are docked at period h in slot t;

– k_h^{vt} is the residual capacity of trucks at period h in slot t for mode v.

Before analysing deeply the model, it can be noted from variables x and f that orders are processed individually. The model offers a highly precise tracking of orders process information, and it is possible to return for each order the exact slot of its process. Such precise tracking is required in e-commerce for the management of the whole delivery journey of the order and also for the management of customer relationship.

The objective function (2.1) is to minimize the cost of processing all the orders. This cost is given by the sum of four terms computed over the planning horizon. The first term is the sum of all penalties due each time the process of an order is postponed to a future period, or each time the delivery mode of an order is changed. The second term in the objective function is the total labour cost computed as the sum of all workers costs, while the third is the cost of the used trucks. The fourth term is a measure of the docks occupation, and it is incremented each time one of the dock is occupied by a truck during one time-slot. We express this term in dock-slot as it is the case when measuring an amount of work using man-hour or man-day units.

Constraints (2.2) ensure that all the packages that compose an order are prepared. Constraints (2.3) and Constraints (2.4) impose that each order is assigned to only one mode and processed entirely during the same period. Constraints (2.5) forbid to prepare orders before their release date. Constraints (2.6) is the packages flow conservation: processed packages are loaded in an already docked truck with residual capacity or in an empty truck. These constraints are formulated differently for the first slot of each period. The reader can note the update of the trucks residual capacity at every slot. Constraints (2.7) (resp. Constraints (2.8)) force variables y_h^{vt} to be one (resp. zero) if (resp. if no) additional trucks for the mode v will be used

during the slot t or the slots after t of period h.

$$(IPSP)\quad \min \sum_{h\in\mathcal{H}}\sum_{d\in\mathcal{D}_h}\sum_{\bar{h}\in\bar{\mathcal{H}}}\sum_{v\in\mathcal{V}} p^{\bar{h}}_{v_{hd}v}\, vol_{hd}\, x^{\bar{h}v}_{hd} + \sum_{h\in\mathcal{H}}\sum_{s\in\mathcal{S}}(c^{per}_{hs} z^{per}_{hs} + c^{temp}_{hs} z^{temp}_{hs}) +$$

$$+ c^{truck}\sum_{h\in\mathcal{H}}\sum_{v\in\mathcal{V}}\sum_{t\in\mathcal{T}} u^{vt}_h + \sum_{h\in\mathcal{H}}\sum_{v\in\mathcal{V}}\sum_{t\in\mathcal{T}} w^{vt}_h \tag{2.1}$$

$$\text{s.t.} \sum_{\bar{h}\in\bar{\mathcal{H}}}\sum_{v\in\mathcal{V}}\sum_{t\in\mathcal{T}} f^{\bar{h}vt}_{hd} = vol_{hd} \quad \forall h\in\mathcal{H},\ \forall d\in\mathcal{D}_h \tag{2.2}$$

$$\sum_{t\in\mathcal{T}} f^{\bar{h}vt}_{hd} \le vol_{hd}\, x^{\bar{h}v}_{hd} \quad \forall h\in\mathcal{H},\ \forall d\in\mathcal{D}_{\langle},\ \forall \bar{h}\in\bar{\mathcal{H}},\ \forall v\in\mathcal{V} \tag{2.3}$$

$$\sum_{\bar{h}\in\bar{\mathcal{H}}}\sum_{v\in\mathcal{V}} x^{\bar{h}v}_{hd} = 1 \quad \forall h\in\mathcal{H},\ \forall d\in\mathcal{D}_{\langle} \tag{2.4}$$

$$f^{0vt}_{hd} = 0 \quad \forall h\in\mathcal{H},\ \forall d\in\mathcal{D}_{\langle},\ \forall v\in\mathcal{V},\ 0\le t< r_{hd} \tag{2.5}$$

$$\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\ge 0}}\sum_{d\in\mathcal{D}_{h-\bar{h}}} f^{\bar{h}v0}_{(h-\bar{h})d} + k^{v0}_h = Qu^{v0}_h \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V}$$

$$\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\ge 0}}\sum_{d\in\mathcal{D}_{h-\bar{h}}} f^{\bar{h}vt}_{(h-\bar{h})d} + k^{vt}_h = k^{v(t-1)}_h + Qu^{vt}_h \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V},\ 0<t\le t_v \tag{2.6}$$

$$\sum_{t=\bar{t}}^{t_v} u^{vt}_h \le t_v N_{\max}\, y^{v\bar{t}}_h \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V},\ 0\le\bar{t}\le t_v \tag{2.7}$$

$$y^{v\bar{t}}_h \le \sum_{t=\bar{t}}^{t_v} u^{vt}_h \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V},\ 0\le\bar{t}\le t_v \tag{2.8}$$

$$Qu^{v0}_h \le Qw^{v0}_h \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V}$$

$$Qu^{vt}_h + k^{v(t-1)}_h \le Qw^{vt}_h + Q(1-y^{vt}_h) \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V},\ 0<t\le t_v \tag{2.9}$$

$$Qu^{vt}_h + k^{v(t-1)}_h - k^{v(t_v-1)}_h \le Qw^{vt}_h \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V},\ 0<t\le t_v \tag{2.10}$$

$$\sum_{v\in\mathcal{V}} w^{vt}_h \le N_{\max} \quad \forall h\in\mathcal{H},\ \forall t\in\mathcal{T} \tag{2.11}$$

$$\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\ge 0}}\sum_{d\in\mathcal{D}_{h-\bar{h}}}\sum_{v\in\mathcal{V}} f^{\bar{h}vt}_{(h-\bar{h})d} \le prod^{per}_{hs} z^{per}_{hs} + prod^{temp}_{hs} z^{temp}_{hs}$$

$$\forall h\in\mathcal{H},\ \forall s\in\mathcal{S}, start_{hs}\le t\le end_{hs} \tag{2.12}$$

$$z^{per}_{hs} \le e^{\max}_{hs} \quad \forall h\in\mathcal{H},\ \forall s\in\mathcal{S} \tag{2.13}$$

$$z^{temp}_{hs} \le z^{per}_{hs} \quad \forall h\in\mathcal{H},\ \forall s\in\mathcal{S} \tag{2.14}$$

$$x^{\bar{h}v}_{hd}\in\{0,1\}\ \forall h\in\mathcal{H},\ \forall d\in\mathcal{D}_h,\ \forall\bar{h}\in\bar{\mathcal{H}},\ \forall v\in\mathcal{V} \tag{2.15}$$

$$y^{vt}_h\in\{0,1\}\ \forall h\in\mathcal{H},\ \forall v\in\mathcal{V},\ \forall t\in\mathcal{T} \tag{2.16}$$

$$z^{per}_{hs}, z^{temp}_{hs}\in\mathbb{N}\ \forall h\in\mathcal{H},\ \forall s\in\mathcal{S} \tag{2.17}$$

$$f^{\bar{h}vt}_{hd}\in\mathbb{N}\ \forall h\in\mathcal{H},\ \forall d\in\mathcal{D}_h,\ \forall\bar{h}\in\bar{\mathcal{H}},\ \forall v\in\mathcal{V},\ \forall t\in\mathcal{T} \tag{2.18}$$

$$w^{vt}_h, k^{vt}_h, u^{vt}_h\in\mathbb{N}\ \forall h\in\mathcal{H},\ \forall v\in\mathcal{V},\ \forall t\in\mathcal{T} \tag{2.19}$$

Constraints (2.9) and (2.10) combined enable to apply the truck movement policy explained earlier in this section. They update variables w_h^{vt} that represent the exact number of docks occupied by the trucks associated to each mode at each slot. For constraints (2.9) a particular formulation that corresponds to the first slot of each period is given first. These formulations are different from the general form because they do not implicate trucks residual capacities. In constraint (2.9), a new truck docked at a given slot (variables u_h^{vt}) imply naturally as much occupied docks. While a truck docked earlier remains on dock only if more trucks associated with the same mode are expected to be used in the upcoming slots ($y_h^{vt} = 1$). Constrains (2.10) complete the truck movement policy by handling the particular case, not handled by constraint (2.9), where at a give slot, a truck docked earlier should remain on dock because a quantity of packages, inferior than the current residual capacity, is expected in the upcoming slots without the need for an additional truck. Note that it is possible to formulate constraints (2.9) and (2.10) differently in a more compact way, by expressing variables y_h^{vt} in terms of expected upcoming packages instead of expected upcoming trucks. The proposed formulation was preferred because it presents a good separability, in the sense that constraints (2.7) and (2.8) implicate variables associated with the same period only.

Constraints (2.11) impose a limit on the number of docks that are available. Constraints (2.12) impose that the number of packages to be prepared in each slot should not exceed the production capacity of the workers. Constraints (2.13) impose a limit on the number of permanent workers. Constraints (2.14) ensure that there are not more temporary workers than permanent workers. Otherwise, we assume that permanent workers should be on duty. Constraints (2.15)–(2.19) define the integrality or binary requirements.

Based on the proposed formulation, we give in the following the complexity property of IPSP.

Proposition 1 *The Packaging and Shipping Problem (IPSP) is $\mathcal{NP}$-hard.*

We prove the $\mathcal{NP}$-hardness of (IPSP) by reduction from the knapsack problem (KP). Given a knapsack with volume B and a set $\mathcal{N}$ of N items, indexed from 1 to N, each with a volume b_i and a value c_i, the KP consists in selecting a subset $\bar{\mathcal{N}}$ of $\mathcal{N}$ under the budget constraint which imposes that the total volume is less than or equal to B, such that $\sum_{i\in\bar{\mathcal{N}}} c_i$ is maximised. It can be formulated as follows:

$$\text{(KP)} \quad \max \quad \sum_{i=1}^{N} c_i x_i \tag{2.20}$$

$$s.t. \quad \sum_{i=1}^{N} b_i x_i \leq B \tag{2.21}$$

$$x_i \in \{0,1\}, \forall i \in \mathcal{N} \tag{2.22}$$

where binary variable x_i is equal to 1 if the i-th item has been selected, and zero otherwise. The objective function (2.20) is to maximize the value of the selected

items. Constraint (2.21) is the budget constraint, while Constraints (2.22) impose variables to be binary.

For each instance of the KP we construct the following IPSP instance. For each item $i \in \mathcal{N}$, we construct an order d, such that $d \in \mathcal{D}_0$ (i.e., it is an order associated with the first period of the horizon), $vol_{0d} = b_i$, $r_{0d} = 0$ and it is assigned to a unique mode v. The horizon is made of two periods, i.e., $\mathcal{H} = \{0, 1\}$. Orders received the period 0, can be processed during period 1, i.e., $\bar{\mathcal{H}} = \{0, 1\}$. Each period is made by one shift indexed with zero, i.e., $\mathcal{S} = \{0\}$. Shifts are made by only one slot, then $\mathcal{T} = \{0\}$. This implies only one slot forms the period. No order is received during period 1. For sake of simplicity, in this section we omit the index related to the period and the shift as well as the mode index.

At most one permanent worker is available for each shift, namely, $e_0^{\max} = e_1^{\max} = 1$ with a null cost. Productivity is $\sum_{i \in \mathcal{N}} vol_i$ of the permanent worker working during period 0 (this worker can process all orders arrived in period 0), and B for the permanent worker of period 1. On the other side, temporary workers have a null productivity and their cost is fixed to a strict positive constant, i.e., 1. Due to construction, mode change is not possible (only one mode is available). Postponing order d to period 1 generates a penalty

$$p_d^{\bar{h}} = \begin{cases} \tilde{c}_d = -\frac{c_d}{vol_d} & \text{if } \bar{h} = 1, \\ 0 & \text{if } \bar{h} = 0. \end{cases} \tag{2.23}$$

The cost of a truck is set to zero, i.e., $c^{truck} = 0$. truck capacity is set to $\sum_{i \in \mathcal{N}} vol_i$: a truck can contain all the orders received in period 0. It is supposed that only one dock is available, $N_{\max}$ is set to 1. Other time parameters like the shift starting period, are trivially fixed. This transformation of a KP instance into a IPSP instance is polynomial in time and takes $O(|\mathcal{N}|)$ operations.

For the obtained instance, the model (IPSP) is reduced to (2.24)–(2.42). Note that variables related to the truck management are not present in the objective function. Then, we can suppose variables w, u and y fixed to 1 and get rid of them. Note that this assumption verifies Constraints (2.29)–(2.33).

$$\min \sum_{d\in\mathcal{D}_0}\sum_{\bar{h}\in\bar{\mathcal{H}}} p_d^{\bar{h}}\, vol_d\, x_d^{\bar{h}} + \sum_{h\in\mathcal{H}}(c_h^{per} z_h^{per} + c_h^{temp} z_h^{temp}) \tag{2.24}$$

$$\text{s.t.} \sum_{\bar{h}\in\bar{\mathcal{H}}} f_d^{\bar{h}} = vol_d \quad \forall d \in \mathcal{D}_0 \tag{2.25}$$

$$\sum_{\bar{h}\in\bar{\mathcal{H}}} x_d^{\bar{h}} = 1 \quad \forall d \in D_0 \tag{2.26}$$

$$f_d^{\bar{h}} \leq vol_d\, x_d^{\bar{h}} \quad \forall d \in D_0,\ \forall \bar{h} \in \bar{\mathcal{H}} \tag{2.27}$$

$$\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}=0}}\sum_{d\in\mathcal{D}_0} f_d^{\bar{h}} + k_h = Qu_h \quad \forall h \in \mathcal{H} \tag{2.28}$$

$$u_h \leq N_{\max}\, y_h \quad \forall h \in \mathcal{H} \tag{2.29}$$

$$y_h \leq u_h \quad \forall h \in \mathcal{H} \tag{2.30}$$

$$Qu_h \leq Qw_h + Q(1-y_h) \quad \forall h \in \mathcal{H} \tag{2.31}$$

$$Qu_h \leq Qw_h \quad \forall h \in \mathcal{H} \tag{2.32}$$

$$w_h \leq N_{\max} \quad \forall h \in \mathcal{H} \tag{2.33}$$

$$\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}=0}}\sum_{d\in\mathcal{D}_0} f_d^{\bar{h}} \leq prod_h^{per} z_h^{per} + prod_h^{temp} z_h^{temp} \quad \forall h \in \mathcal{H} \tag{2.34}$$

$$z_h^{per} \leq 1 \quad \forall h \in \mathcal{H} \tag{2.35}$$

$$z_h^{temp} \leq z_h^{per} \quad \forall h \in \mathcal{H} \tag{2.36}$$

$$x_d^{\bar{h}} \in \{0,1\}\ \forall d \in \mathcal{D}_0,\ \forall \bar{h} \in \bar{\mathcal{H}} \tag{2.37}$$

$$y_h \in \{0,1\}\ \forall h \in \mathcal{H} \tag{2.38}$$

$$z_h^{per}, z_h^{temp} \in \mathbb{N}\ \forall h \in \mathcal{H} \tag{2.39}$$

$$f_d^{\bar{h}} \in \mathbb{N}\ \forall d \in \mathcal{D}_0,\ \forall \bar{h} \in \bar{\mathcal{H}} \tag{2.40}$$

$$w_h, k_h, u_h \in \mathbb{N}\ \forall h \in \mathcal{H} \tag{2.41}$$

Moreover, since periods are constituted by only one period, variables k_h become useless and Constraints (2.28) can be replaced by

$$\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}=0}}\sum_{d\in\mathcal{D}_0} f_d^{\bar{h}} \leq Q \quad \forall h \in \mathcal{H} \tag{2.42}$$

Due to construction, Constraints (2.42) are trivially satisfied (a truck can contain the full orders received in period 0). Finally, due to construction, we are sure that in the optimal solution only one fix worker works each period ($z_0^{per} = z_1^{per} = 1$), while no temporary workers will be hired ($z_0^{temp} = z_1^{temp} = 0$). Constraints (2.35)–(2.36) are trivially verified. The model reduces to

$$\min \sum_{d\in\mathcal{D}_0}\sum_{\bar{h}\in\bar{\mathcal{H}}} p_d^{\bar{h}}\, vol_d\, x_d^{\bar{h}} \tag{2.43}$$

$$\text{s.t.} \sum_{\bar{h}\in\bar{\mathcal{H}}} f_d^{\bar{h}} = vol_d \quad \forall d\in\mathcal{D}_0 \tag{2.44}$$

$$\sum_{\bar{h}\in\bar{\mathcal{H}}} x_d^{\bar{h}} = 1 \quad \forall d\in D_0 \tag{2.45}$$

$$f_d^{\bar{h}} \le vol_d\, x_d^{\bar{h}} \quad \forall d\in D_0,\ \forall \bar{h}\in\bar{\mathcal{H}} \tag{2.46}$$

$$\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}=0}}\sum_{d\in\mathcal{D}_0} f_d^{\bar{h}} \le prod_h^{per} \quad \forall h\in\mathcal{H} \tag{2.47}$$

$$x_d^{\bar{h}} \in \{0,1\}\ \forall d\in\mathcal{D}_0,\ \forall \bar{h}\in\bar{\mathcal{H}} \tag{2.48}$$

$$f_d^{\bar{h}} \in \mathbb{N}\ \forall d\in\mathcal{D}_0,\ \forall \bar{h}\in\bar{\mathcal{H}} \tag{2.49}$$

Replacing $p_d^{\bar{h}}$ with the definition given in Equation (2.23), the objective function (2.43) is

$$\sum_{d\in\mathcal{D}_0}\sum_{\bar{h}\in\bar{\mathcal{H}}} p_d^{\bar{h}}\, vol_d\, x_d^{\bar{h}} = \sum_{d\in\mathcal{D}_0}(p_d^0\, vol_d\, x_d^0 + p_d^1\, vol_d\, x_d^1) = \sum_{d\in\mathcal{D}_0} -\frac{c_d}{vol_d}\, vol_d\, x_d^1 = \sum_{d\in\mathcal{D}_0} -c_d\, x_d^1$$

and Constraints (2.47) decompose into

$$\sum_{d\in\mathcal{D}_0} f_d^0 \le prod_0^{per} = \sum_{d\in\mathcal{D}_0} f_d^0 \le \sum_{d\in\mathcal{D}_0} vol_d \tag{2.50}$$

$$\sum_{d\in\mathcal{D}_0} f_d^1 \le prod_1^{per} = \sum_{d\in\mathcal{D}_0} f_d^1 \le B \tag{2.51}$$

Constraint (2.50) is always verified and can be removed. The model becomes

$$\min \sum_{d\in\mathcal{D}_0} -c_d\, x_d^1 \tag{2.52}$$

$$\text{s.t.} \sum_{\bar{h}\in\bar{\mathcal{H}}} f_d^{\bar{h}} = vol_d \quad \forall d\in\mathcal{D}_0 \tag{2.53}$$

$$\sum_{\bar{h}\in\bar{\mathcal{H}}} x_d^{\bar{h}} = 1 \quad \forall d\in D_0 \tag{2.54}$$

$$f_d^{\bar{h}} \le vol_d\, x_d^{\bar{h}} \quad \forall d\in D_0,\ \forall \bar{h}\in\bar{\mathcal{H}} \tag{2.55}$$

$$\sum_{d\in\mathcal{D}_0} f_d^1 \le B \tag{2.56}$$

$$x_d^{\bar{h}} \in \{0,1\}\ \forall d\in\mathcal{D}_0,\ \forall \bar{h}\in\bar{\mathcal{H}} \tag{2.57}$$

$$f_d^{\bar{h}} \in \mathbb{N}\ \forall d\in\mathcal{D}_0,\ \forall \bar{h}\in\bar{\mathcal{H}} \tag{2.58}$$

Note that Constraints (2.55) are never strict, and inequalities can be changed to

$$f_d^{\bar{h}} = vol_d\, x_d^{\bar{h}} \quad \forall d\in D_0,\ \forall \bar{h}\in\bar{\mathcal{H}}$$

(perfect relation between variables x and f is due to the fact that periods are made by only one period). Then the model is equivalent to

$$\text{(IPSP-KP)} \quad \max \sum_{d\in\mathcal{D}_0} c_d\, x_d^1 \tag{2.59}$$

$$\text{s.t.} \sum_{\bar{h}\in\bar{\mathcal{H}}} x_d^{\bar{h}} = 1 \quad \forall d \in D_0 \tag{2.60}$$

$$\sum_{d\in\mathcal{D}_0} vol_d x_d^1 \leq B \tag{2.61}$$

$$x_d^{\bar{h}} \in \{0,1\}\ \forall d \in \mathcal{D}_0,\ \forall \bar{h} \in \bar{\mathcal{H}} \tag{2.62}$$

It is trivial to see that problems defined by models (IPSP-KP) and (KP) have the same optimal solution, and this concludes the proof. ■

2.5 A three-phase matheuristic

To solve the IPSP we propose an algorithm based on a three-phase matheuristic Following the classification of matheuristics proposed by Ball [Ball 2011], our procedure falls into the *decomposition approach* category: sub-problems are sequentially solved in order to identify a feasible solution for the original problem. Our three-phase approach sequentially solves three sub-problems of the IPSP, in a way that the solution of each sub-problem (or part of it) is the input for the next phase. The solution of the third sub-problem is in turn a solution for the IPSP.

In our three-phase approach, the first phase solves a relaxation of the IPSP model presented in Section 2.4. It determines the workers needed to process all the orders. In other words, for all $h \in \mathcal{H}$ and $s \in \mathcal{S}$ we fix the values of variables z_{hs}^{per} and z_{hs}^{temp}. This phase leads to take the tactical decisions under aggregated operational constraints. The second phase determines the complete orders process planning and sets the mode changes and the postponements. Specifically, it determines, for each $h \in \mathcal{H}$, $d \in \mathcal{D}_h$, $\bar{h} \in \bar{\mathcal{H}}$, $v \in \mathcal{V}$, the values of variables $x_{hd}^{\bar{h}v}$. This phase focuses at the operational level based on decisions taken on the previous phase of the method. The solution provided by the second phase does not consider docks occupation minimization, but provides a feasible solution for IPSP. Therefore, the algorithm could be stopped after this phase.

If the algorithm is continued, the output of the second phase is used as input for the last phase which considers the truck movement policy and minimizes the dock occupation. This phase refines operational decisions to optimize the dock occupation. The phases are detailed in the next sections. An outline of the three-phase procedure is given in Algorithm 1.

For each of the three phases of the algorithm, we propose a speed-up technique to cut computation times. In particular, we compute two valid lower-bounds on the objective function for the models solved in phase I and phase II. Each of these lower

bounds is used to define a stopping criterion. Before starting the third phase of the algorithm, we implement a orders *aggregation* procedure which groups all the orders with the same characteristics. Indeed, the mode changes and the postponements are determined in phase II, and the aggregation procedure does not reduce the set of feasible solutions as it will be further explained in Section 2.5.4.3.

Algorithm 1 Three-phase algorithm

```
1: Phase I
2: Compute lower-bound for model IPSP I LBI (Section 2.5.4.2)
3: while Time limit not reached do
4:    Solve model IPSP I (Section 2.5.1)
5:    for all Feasible solutions found do
6:       if Solution optimality proved or solution cost equal to LBI then
7:          Go to Step 8
8:    end
9: end
10: Fix workers based on solution of model IPSP I
11: Phase II
12: Compute lower-bound for model IPSP II LBII (Section 2.5.4.1)
13: while Time limit not reached do
14:    Solve model IPSP II (Section 2.5.2)
15:    for all Feasible solutions found do
16:       if Solution optimality proved or solution cost equal to LBII then
17:          Go to Step 16
18:    end
19: end
20: Fix reassignments and postponements based on solution of model IPSP II
21: Phase III
22: Aggregate orders (Section 2.5.4.3)
23: while Time limit not reached do
24:    Solve model IPSP III
25: end
26: Disaggregate orders (Algorithm 2)
```

This decomposition approach is based on the distinction of the different decisions regarding their nature. In the first phase, the tactical decisions, namely, the workers needed for production, are determined. The second and the third phases concentrate on the operational decisions. We first determine a complete feasible planning and then we re-optimize the production planning in order to minimize the dock occupation.

Sections 2.5.1–2.5.3 present the three phases of the algorithm. Section 2.5.4 presents the speed-up techniques.

2.5.1 Phase I - Production capacity

This phase determines the number of workers that are required during each shift of each period of the planning horizon. To this aim, we solve a relaxation of the model (IPSP) presented in Section 2.4. The relaxation does not consider the truck management issues, i.e., constraints (2.6)–(2.11) are not taken into account. The relaxation of the model (IPSP) is based on the following proposition.

Proposition 2 *The following model*

$$(IPSP\ I) \quad \min \sum_{h\in\mathcal{H}}\sum_{d\in\mathcal{D}_h}\sum_{\bar{h}\in\bar{\mathcal{H}}}\sum_{v\in\mathcal{V}} p^{\bar{h}}_{v_{hd}v}\, vol_{hd}\, x^{\bar{h}v}_{hd} + \sum_{h\in\mathcal{H}}\sum_{s\in\mathcal{S}}(c^{per}_{hs} z^{per}_{hs} + c^{temp}_{hs} z^{temp}_{hs}) \tag{2.63}$$

$$s.t.\ (2.2)\text{–}(2.5)$$

$$\sum_{v\in\mathcal{V}}\Big(\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}}\sum_{d\in\mathcal{D}_{h-\bar{h}}} f^{\bar{h}vt}_{(h-\bar{h})d} + k^{vt}_h\Big) \leq QN_{\max} + QV \quad \forall h\in\mathcal{H},\ 0\leq t\leq t_v \tag{2.64}$$

$$(2.12)\text{–}(2.18)$$

is a valid relaxation for model (IPSP).

Note the change in the objective function of (IPSP I) compared to (2.1), and the substitution of constraints (2.6)–(2.11) by the constraints (2.64).

By summing constraints (2.6) over all modes in $\mathcal{V}$, we obtain (using Constraints (2.9))

$$\sum_{v\in\mathcal{V}}\Big(\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}}\sum_{d\in\mathcal{D}_{h-\bar{h}}} f^{\bar{h}vt}_{(h-\bar{h})d}\Big) \leq Q\sum_{v\in\mathcal{V}}(w^{vt}_h + (1-y^{vt}_h)) - \sum_{v\in\mathcal{V}} k^{vt}_h$$

From Constraint (2.11) it follows

$$\sum_{v\in\mathcal{V}}\Big(\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}}\sum_{d\in\mathcal{D}_{h-\bar{h}}} f^{\bar{h}vt}_{(h-\bar{h})d} + k^{vt}_h\Big) \leq QN_{\max} + Q\Big(\sum_{v\in\mathcal{V}}(1-y^{vt}_h)\Big) - \sum_{v\in\mathcal{V}} k^{vt}_h$$

From Constraint (2.19) on the variables, it follows that the term $\sum_{v\in\mathcal{V}} k^{vt}_h$ is positive, then

$$\sum_{v\in\mathcal{V}}\Big(\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}}\sum_{d\in\mathcal{D}_{h-\bar{h}}} f^{\bar{h}vt}_{(h-\bar{h})d} + k^{vt}_h\Big) \leq QN_{\max} + Q\Big(\sum_{v\in\mathcal{V}}(1-y^{vt}_h)\Big)$$

Finally, the term $\sum_{v\in\mathcal{V}}(1-y^{vt}_h) \in \{0,\ldots,V\}$ equals V when all the variables y^{vt}_h equal 0, i.e., when the process is ended. Then, we obtain:

$$\sum_{v\in\mathcal{V}}\Big(\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}}\sum_{d\in\mathcal{D}_{h-\bar{h}}} f^{\bar{h}vt}_{(h-\bar{h})d} + k^{vt}_h\Big) \leq QN_{\max} + QV$$

■

(IPSP I) is solved with a commercial solver, and the solution is used to determine the number of workers assigned to each shift of each period.

2.5.2 Phase II - Reassignment and postponement

Based on the decisions obtained in phase I, the second phase of the algorithm determines the assignment of each order to a period and to a delivery mode (variables $x_{hd}^{\bar{h}v}$). The productivity capacity during each shift is known from Phase I, i.e., variables z_{hs}^{per} and z_{hs}^{temp} are now fixed. Moreover, we do not minimize the platform occupation, i.e., the term $\sum_{h\in\mathcal{H}}\sum_{v\in\mathcal{V}}\sum_{t\in\mathcal{T}} w_h^{vt}$ is removed from the objective function. The model solved in this phase is the following (the variables z_{hs}^{per} and z_{hs}^{temp} are replaced by the parameters ζ_{hs}^{per} and ζ_{hs}^{temp})

$$\text{(IPSP II)} \quad \min \sum_{h\in\mathcal{H}}\sum_{d\in\mathcal{D}_h}\sum_{\bar{h}\in\bar{\mathcal{H}}}\sum_{v\in\mathcal{V}} p_{v_{hd}v}^{\bar{h}}\, vol_{hd}\, x_{hd}^{\bar{h}v} + c^{truck}\sum_{h\in\mathcal{H}}\sum_{v\in\mathcal{V}}\sum_{t\in\mathcal{T}} u_h^{vt} \tag{2.65}$$

$$\text{s.t. } (2.2)\text{–}(2.11)$$

$$\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}}\sum_{d\in\mathcal{D}_{h-\bar{h}}}\sum_{v\in\mathcal{V}} f_{(h-\bar{h})d}^{\bar{h}vt} \leq prod_{hs}^{per}\zeta_{hs}^{per} + prod_{hs}^{temp}\zeta_{hs}^{temp}$$

$$\forall h\in\mathcal{H},\ \forall s\in\mathcal{S}, \forall t = start_{hs},\ldots,end_{hs} \tag{2.66}$$

$$(2.15)\text{–}(2.16)$$

$$(2.18)\text{–}(2.19)$$

Note that the solution that is obtained after phase II is a feasible solution for the (IPSP) model: by construction, it satisfies constraints (2.2)–(2.19). This solution can be used as an initial feasible solution in the last phase.

2.5.3 Phase III - Dock management

In the last phase, the platform occupation is optimized, i.e., we look to minimize the number of slots during which vehicles are present at the docks. The workers to hire, i.e., the values of variables z_{hs}^{per} and z_{hs}^{temp} and the possible reassignment or postponement of order preparation, i.e., the values of variables $x_{hd}^{\bar{h}v}$, are fixed and are parameters of the model (indicated with $\chi_{hd}^{\bar{h}v}$). The mathematical model solved

in this phase is the following:

$$\text{(IPSP III)} \quad \min \sum_{h\in\mathcal{H}}\sum_{v\in\mathcal{V}}\sum_{t\in\mathcal{T}} w_h^{vt} \tag{2.67}$$

$$\text{s.t.} \sum_{t\in\mathcal{T}} f_{hd}^{\bar{h}vt} \le vol_{hd}\,\chi_{hd}^{\bar{h}v} \quad \forall h\in\mathcal{H},\ \forall d\in D_h,\ \forall \bar{h}\in\bar{\mathcal{H}},\ \forall v\in\mathcal{V} \tag{2.68}$$

$$\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\ge 0}}\sum_{d\in\mathcal{D}_{h-\bar{h}}}\sum_{v\in\mathcal{V}} f_{(h-\bar{h})d}^{\bar{h}vt} \le prod_{hs}^{per}\zeta_{hs}^{per} + prod_{hs}^{temp}\zeta_{hs}^{temp} \quad \forall h\in\mathcal{H},\ \forall s\in\mathcal{S}, \forall t = start_{hs},\dots,end_{hs} \tag{2.69}$$

(2.2), (2.5)–(2.11)

(2.16), (2.18)–(2.19)

Variables $x_{hd}^{\bar{h}v}$, z_{hs}^{per} and z_{hs}^{temp} are initially fixed to the values obtained in phase II. The solution provided by Phase III is the final solution obtained for the IPSP.

2.5.4 Speed-up techniques

To speed-up the algorithm, we compute two valid lower-bounds on the objective function values of models (IPSP I) and (IPSP II). The lower-bounds are given by the solution of two specific arc-flows problems. Since the computation of these lower-bounds is based on the same idea, we detail only the computation of the lower bound for (IPSP II).

2.5.4.1 Lower-bound for (IPSP II)

We first recall that in phase II, the objective function is given by the sum of the penalties due to the mode changes and the postponements plus the cost of used trucks. The lower-bound is obtained by solving the following relaxation of (IPSP II).

$$\text{(RIPSP II)} \quad \min \sum_{h\in\mathcal{H}}\sum_{v\in\mathcal{V}}\sum_{\bar{h}\in\mathcal{H}}\sum_{\bar{v}\in\mathcal{V}} p_{v\bar{v}}^{\bar{h}}\xi_{hv}^{\bar{h}\bar{v}} + c^{truck}\sum_{h\in\mathcal{H}}\sum_{v\in\mathcal{V}}\zeta_h^v \tag{2.70}$$

$$D_{hv} + \sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\ge 0}}\sum_{\bar{v}\in\mathcal{V}}\xi_{hv}^{\bar{h}\bar{v}} - \xi_h^v - \sum_{\bar{h}\in\bar{\mathcal{H}}}\sum_{\bar{v}\in\mathcal{V}}\xi_{hv}^{\bar{h}\bar{v}} = 0 \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V} \tag{2.71}$$

$$\xi_h^v \le Q\zeta_v^h \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V} \tag{2.72}$$

$$\xi_{hv}^{\bar{h}\bar{v}} \in \mathbb{N}, \quad \forall h,\bar{h}\in\mathcal{H},\ \forall v,\bar{v}\in\mathcal{V} \tag{2.73}$$

$$\xi_h^v, \zeta_v^h \in \mathbb{N}, \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V} \tag{2.74}$$

where

- $\xi_{hv}^{\bar{h}\bar{v}}$ represents the total volume of orders for period h and mode v treated on period $\bar{h}$ by mode $\bar{v}$
- ξ_h^v represents the total volume of orders for period h and mode v that is not postponed or reassigned
- ζ_h^v represents the total number of needed vehicles for mode v in period h

and D_{hv} is the total number of packages which should be prepared on period h and delivered by mode v.

Proposition 3 *Model (RIPSP II) is a relaxation of model (IPSP II).*

From Equations (2.4), multiplying both terms by vol_{hd}, we obtain

$$\sum_{\bar{h}\in\bar{\mathcal{H}}}\sum_{v\in\mathcal{V}} vol_{hd}x_{hd}^{\bar{h}v} = vol_{hd} \quad \forall h\in\mathcal{H},\ \forall d\in D_h$$

and summing up on the demands $d\in\mathcal{D}_h$ we obtain

$$\sum_{d\in\mathcal{D}_h}\sum_{\bar{h}\in\bar{\mathcal{H}}}\sum_{v\in\mathcal{V}} vol_{hd}x_{hd}^{\bar{h}v} = \sum_{d\in\mathcal{D}_h} vol_{hd} = \sum_{v\in\mathcal{V}}\sum_{\substack{d\in\mathcal{D}_h\\ v_{hd}=v}} vol_{hd} \quad \forall h\in\mathcal{H}$$

$$\sum_{v\in\mathcal{V}}\sum_{\substack{d\in\mathcal{D}_h\\ v_{hd}=v}}\sum_{\bar{h}\in\bar{\mathcal{H}}}\sum_{\bar{v}\in\mathcal{V}} vol_{hd}x_{hd}^{\bar{h}\bar{v}} = \sum_{d\in\mathcal{D}_h} vol_{hd} = \sum_{v\in\mathcal{V}}\sum_{\substack{d\in\mathcal{D}_h\\ v_{hd}=v}} vol_{hd} \quad \forall h\in\mathcal{H}$$

and, for $v\in\mathcal{V}, h\in\mathcal{H}$ let us define $\xi_{v\bar{v}}^{h\bar{h}} = \sum_{\substack{d\in\mathcal{D}_h\\ v_{hd}=v}} vol_{hd}x_{hd}^{\bar{h}\bar{v}}$ as the total demand of day h assigned to mode v that is delivered on period $\bar{h}$ by mode $\bar{v}$. We then have

$$\sum_{v\in\mathcal{V}}\sum_{\bar{h}\in\bar{\mathcal{H}}}\sum_{\bar{v}\in\mathcal{V}}\xi_{v\bar{v}}^{h\bar{h}} = \sum_{d\in\mathcal{D}_h} vol_{hd} = \sum_{v\in\mathcal{V}}\sum_{\substack{d\in\mathcal{D}_h\\ v_{hd}=v}} vol_{hd} = \sum_{v\in\mathcal{V}} D_v^h = D^h \quad \forall h\in\mathcal{H} \tag{2.75}$$

where D_v^h and D^h are respectively the total demand of day h originally associated with mode v and the total demand of day h. Note that since, for each $h\in\mathcal{H}$, for each $d\in\mathcal{D}_h$, for each $v\in\mathcal{V}$ and for each $\bar{h}\in\bar{\mathcal{H}}$ there exist only one variable $x_{hd}^{\bar{h}v}$ that is equal to one, we can write Equations (2.75) as

$$\sum_{\bar{h}\in\bar{\mathcal{H}}}\sum_{\bar{v}\in\mathcal{V}}\xi_{v\bar{v}}^{h\bar{h}} = \sum_{\substack{d\in\mathcal{D}_h\\ v_{hd}=v}} vol_{hd} = D_v^h \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V}$$

$$D_v^h - \sum_{\bar{h}\in\bar{\mathcal{H}}}\sum_{\bar{v}\in\mathcal{V}}\xi_{v\bar{v}}^{h\bar{h}} = 0 \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V} \tag{2.76}$$

From Equations (2.6), summing on $t \in \mathcal{T}$ we obtain

$$\sum_{t\in\mathcal{T}} \sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}} \sum_{d\in\mathcal{D}_{h-\bar{h}}} f^{\bar{h}vt}_{(h-\bar{h})d} + \sum_{t\in\mathcal{T}} k^{vt}_h = \sum_{\substack{t\in\mathcal{T}\\ t>0}} k^{v(t-1)}_h + Q\sum_{t\in\mathcal{T}} u^{vt}_h \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V}$$

$$\sum_{t\in\mathcal{T}} \sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}} \sum_{d\in\mathcal{D}_{h-\bar{h}}} f^{\bar{h}vt}_{(h-\bar{h})d} + k^{vT}_h = Q\sum_{t\in\mathcal{T}} u^{vt}_h \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V}$$

$$\sum_{t\in\mathcal{T}} \sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}} \sum_{d\in\mathcal{D}_{h-\bar{h}}} f^{\bar{h}vt}_{(h-\bar{h})d} \leq Q\sum_{t\in\mathcal{T}} u^{vt}_h \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V}$$

Let us now define $\xi^h_v = \sum_{t\in\mathcal{T}} \sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}} \sum_{d\in\mathcal{D}_{h-\bar{h}}} f^{\bar{h}vt}_{(h-\bar{h})d}$ and $\zeta^h_v = \sum_{t\in\mathcal{T}} u^{vt}_h$. Then we obtain

$$\xi^h_v \leq Q\zeta^h_v \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V} \tag{2.77}$$

that are Constraints (2.72). ξ^h_v represents the total volume of packages that need to be prepared in day h and delivered by mode v after postponing and re-affecting operations. ζ^h_v represents the number of vehicle needed to transport the ξ^h_v packages.

From the definition of ξ^h_v we have

$$\xi^h_v = \sum_{t\in\mathcal{T}} \sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}} \sum_{d\in\mathcal{D}_{h-\bar{h}}} f^{\bar{h}vt}_{(h-\bar{h})d}, \quad \forall h\in\mathcal{H}, \forall v\in\mathcal{V}$$

$$\xi^h_v = \sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}} \sum_{\bar{v}\in\mathcal{V}} \sum_{\substack{d\in\mathcal{D}_{h-\bar{h}}\\ \bar{v}=v_{hd}}} \sum_{t\in\mathcal{T}} f^{\bar{h}vt}_{(h-\bar{h})d}, \quad \forall h\in\mathcal{H}, \forall v\in\mathcal{V}$$

$$\xi^h_v = \sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}} \sum_{\bar{v}\in\mathcal{V}} \xi^{\bar{h}h}_{\bar{v}v}, \quad \forall h\in\mathcal{H}, \forall v\in\mathcal{V}$$

$$\sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}} \sum_{\bar{v}\in\mathcal{V}} \xi^{\bar{h}h}_{\bar{v}v} - \xi^h_v = 0, \quad \forall h\in\mathcal{H}, \forall v\in\mathcal{V} \tag{2.78}$$

where we have defined $\xi^{\bar{h}h}_{\bar{v}v} = \sum_{\substack{d\in\mathcal{D}_{h-\bar{h}}\\ \bar{v}=v_{hd}}} \sum_{t\in\mathcal{T}} f^{\bar{h}vt}_{(h-\bar{h})d}$, that represent all the packages originally assigned to day $\bar{h}-h$ and mode $\bar{v}$ that are prepared on day h (i.e., are postponed by $\bar{h}$) and mode v.

Summing Equations (2.76) and (2.78) we obtain:

$$D^h_v + \sum_{\substack{\bar{h}\in\bar{\mathcal{H}}\\ h-\bar{h}\geq 0}} \sum_{\bar{v}\in\mathcal{V}} \xi^{\bar{h}h}_{\bar{v}v} - \xi^h_v - \sum_{\bar{h}\in\bar{\mathcal{H}}} \sum_{\bar{v}\in\mathcal{V}} \xi^{h\bar{h}}_{v\bar{v}} = 0 \quad \forall h\in\mathcal{H},\ \forall v\in\mathcal{V} \tag{2.79}$$

that are Constraints (2.71). All the other constraints in the model (PSP II) are relaxed. ■

From Proposition 3 it follows that the value of the optimal solution of (RIPSP II) is a lower-bound for (IPSP II).

The model (RIPSP II) is a special case of the multi-commodity capacitated network design problem where only one commodity has to be routed on the network, and capacities have to be respected or installed in order to satisfy the demand (see, for example, Gendron et al. [Gendron 1999]). In particular, (RIPSP II) is equivalent to

$$\text{(AF-RIPSP II)} \quad \min \sum_{v\in\mathcal{V}}\sum_{\bar{v}\in\mathcal{V}}\sum_{h\in\mathcal{H}}\sum_{\bar{h}\in\mathcal{H}} p^{\bar{h}}_{v\bar{v}}\xi^{\bar{h}\bar{v}}_{hv} + c^{truck}\sum_{v\in\mathcal{V}}\sum_{h\in\mathcal{H}}\zeta^{v}_{h} \tag{2.80}$$

$$\mathbf{A}\boldsymbol{\xi} = \mathbf{b} \tag{2.81}$$

$$\mathbf{0} \leq \boldsymbol{\xi} \leq c(\boldsymbol{\zeta}) \tag{2.82}$$

$$\boldsymbol{\zeta}, \boldsymbol{\xi} \in \mathbb{N} \tag{2.83}$$

Note that model (AF-RIPSP II) defines an arc-flow problem on an oriented graph $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ where $\mathcal{N} = \{s, t\} \cup \mathcal{N}^h_v$, and $\mathcal{N}^h_v$ contains a node n^h_v for each pair (v, h), $v \in \mathcal{V}$, $h \in \mathcal{H}$ and

$$\mathcal{A} = \{(s, i)|i \in \mathcal{N}^h_v\} \cup \{(i, t)|i \in \mathcal{N}^h_v\} \cup \{(i, j)|i, j \in \mathcal{N}^h_v, i \neq j\}$$

$\mathbf{A}$ is the adjacency matrix of graph $\mathcal{G}$. Vectors $\mathbf{b}$ and c are as follows:

$$b_i = \begin{cases} -\sum_{v\in\mathcal{V}}\sum_{h\in\mathcal{H}} D^h_v & \text{if } i = s \\ \sum_{v\in\mathcal{V}}\sum_{h\in\mathcal{H}} D^h_v & \text{if } i = t \\ 0 & \text{otherwise} \end{cases}$$

and,

$$c_a = \begin{cases} D^h_v & \text{if } a = (s, n^h_v) \\ Q\zeta^h_v & \text{if } a = (n^h_v, t) \\ \sum_{v\in\mathcal{V}}\sum_{h\in\mathcal{H}} D^h_v & \text{if } a = (n^h_v, n^{h_1}_{v_1}), i \neq s, t \end{cases}$$

Model (RIPSP II) is solved with a commercial solver and the value of the optimal solution gives the lower-bound for the phase. With respect to our testbed, the size of instances remains small, and optimal solutions are obtained almost instantaneously. Each time a feasible solution for model (IPSP II) is identified and its value is equal to the lower-bound, the solution of model (IPSP II) is stopped.

2.5.4.2 Lower-bound for phase I

The model (RIPSP II) determines orders that require postponement or mode change in order to minimize the number of vehicles, and it computes the resulting penalties. A valid lower-bound for phase I is obtained accordingly: an estimation of postponed orders over the horizon is computed to minimize the number of required workers. The problem can be formulated as arc-flow problem similar to (AF-RIPSP II). Due to similarities shared by both constructions we omit the details here.

2.5.4.3 Order aggregation for phase III

Phase II determines the quantities of orders assigned to each mode in each period. Based on these decisions, Phase III looks for a packages loading plan, in other words the quantities loaded at each slot and the required trucks movements, that minimizes the docks occupation. To speed up the solution of phase III model, we *aggregate* orders: we cluster orders that

- have the same release date,
- are assigned for process on the same period $h \in \mathcal{H}$,
- are assigned for delivery to the same mode $v \in \mathcal{V}$

as a unique order whose volume is the sum of individual order volumes. Orders in $\mathcal{D}_h$ postponed by $\bar{h} > 0$ are included in the volume of the unique order created for period $h + \bar{h}$ associated with a release date equal to zero. Since these orders were available $\bar{h}$ periods before, they are available at the beginning of period $h + \bar{h}$.

This aggregation can be performed since no postponement or mode change are allowed during this phase. Postponement or mode change must be done on the total volume of an order. Aggregation is not possible when postponement or reassignment are admissible, since the packages associated with each order have to be known. This information is be lost in case of aggregation.

Formally, for each $h \in \mathcal{H}$, for each $v \in \mathcal{V}$ and for each $t \in \mathcal{T}$ we define a unique order D^t_{hv} with a volume $vol_{D^t_{hv}}$ defined as follows:

$$vol_{D^t_{hv}} = \begin{cases} \sum_{\substack{d \in \mathcal{D}_h \\ r_{hd}=t}} vol_{hd} x^{0v}_{hd} + \sum_{\substack{d \in \mathcal{D}_{h-\bar{h}} \\ h-\bar{h} \geq 0,\ \bar{h}>0}} vol_{(h-\bar{h})d} x^{\bar{h}v}_{(h-\bar{h})d} & \text{if } t = 0 \\ \sum_{\substack{d \in \mathcal{D}_h \\ r_{hd}=t}} vol_{hd} x^{0v}_{hd} & \text{if } t > 0 \end{cases} \tag{2.84}$$

When $t = 0$, the volume D^t_{hv} corresponds to the sum of volumes of all orders released exactly at $t = 0$ on the period h and processed on the same period, plus the volume of all the orders released during period $h - \bar{h}$ and postponed by $\bar{h}$ periods. When an order is postponed to a given period h, it is known at the beginning of period h. When $t > 0$, the volume D^t_{hv} corresponds to the sum of volumes of all orders released exactly at $t > 0$ on the same period h.

Moreover, D^t_{hv} is characterized by its mode $v_{D^t_{hv}} = v$, and its release date $r_{D^t_{hv}} = t$.

Let $\bar{\mathcal{D}}_h$ denote the set of all aggregated orders. These orders are the input of the model solved by the commercial solver in phase III. Let us indicate with $\bar{f}^{\bar{h}vt}_{hd}$ the variables corresponding to orders in $\bar{\mathcal{D}}_h$. A solution of phase III determines an operational planning for orders $d \in \bar{\mathcal{D}}_h$. The solution of the problem in terms of variables $f^{\bar{h}vt}_{hd}$ can easily be obtained by applying a greedy algorithm using the values of variables $\bar{f}^{\bar{h}vt}_{hd}$.

Algorithm 2 gives an example of greedy procedure used to construct the process plan for the original orders (orders dis-aggregation).

Algorithm 2 Dis-aggregation algorithm

for all $h \in \mathcal{H}$ **do**
  for all $r \in \mathcal{T}$ **do**
    $d = 0$
    $vol = 0$
    while $d < |\mathcal{D}_h|$ **do**
      $t = 0$
      $h^* = \arg\max_{\bar{h} \in \bar{\mathcal{H}}} x_{hd}^{\bar{h}v}$
      $v^* = \arg\max_{v \in \mathcal{V}} x_{hd}^{\bar{h}v}$ {if $h^* = 0$, d is included in $D_{hv^*}^{r_d}$; if $h^* > 0$, d is included in $D_{(h+h^*)v^*}^{0}$}
      if $h^* = 0$ **then**
        $f_{hd}^{h^*v^*t} = \max\{\bar{f}_{hD_{hv^*}^r}^{0v^*t}, vol_{hd}\}$
        $\bar{f}_{hD_{hv^*}^r}^{0v^*t} = \bar{f}_{hD_{hv^*}^r}^{0v^*t} - f_{hd}^{h^*v^*t}$
      else
        $f_{hd}^{h^*v^*t} = \max\{\bar{f}_{(h+h^*)D_{(h+h^*)v^*}^0}^{0v^*t}, vol_{hd}\}$
        $\bar{f}_{(h+h^*)D_{(h+h^*)v^*}^0}^{0v^*t} = \bar{f}_{(h+h^*)D_{(h+h^*)v}^0}^{0v^*t} - f_{hd}^{h^*v^*t}$
      $vol = vol + f_{hd}^{h^*v^*t}$
      if $vol = vol_{hd}$ **then**
        $d = d + 1$
        $vol = 0$
      else
        $t = t + 1$

2.6 Computational results

This section discusses on the efficiency of the three-phase procedure we developed for the IPSP. We first describe the instances we created starting from data provided by an industrial partner (Section 2.6.1). Results on these instances are reported in Section 2.6.2.1. Sensitivity analysis of the three-phase algorithm with respect to slight modification of instances and with respect to different penalty profiles is reported, respectively in Sections 2.6.2.2 and 2.6.2.3. Section 2.6.2.4 shows the performance of the lower-bounds used in phases I and II introduced in Sections 2.5.4.1 and 2.5.4.2. Finally, Section 2.6.2.5 compares the three-phase algorithm with the solution of the (IPSP) using a commercial solver.

2.6.1 Instance generation

The (IPSP) being a new problem, we have to generate a set of instances to test the algorithm described in Section 2.5. The instances are based on real data provided by a logistic company in the e-commerce sector.

Each working day is identified by a *profile*, namely, a number of orders known during the day. We define three profiles, named *low*, *normal*, *high* respectively characterized by 1000, 3000, 5000 orders.

A list of common data is shared among the different instances. Specifically, we consider a horizon of three days, $\mathcal{H} = \{0, 1, 2\}$, while order process can be postponed by one day, i.e., $\bar{\mathcal{H}} = \{0, 1\}$. Orders are received only during days 0 and 1. The third day is only used in case the whole demand cannot be prepared during days 0 and 1. Each day includes two shifts made of eight time slot.

Two delivery modes are available, the *express* mode and the *normal* mode. Trucks associated to the express mode leave the warehouse earlier than the other trucks, that are supposed to leave at the end of the last shift. As an example, the express mode departure time is slot 12. This means packages can be loaded into vehicles until slot 11. Trucks have a capacity of 1300 packages and a fixed cost of 650 Euros for both modes.

Permanent workers productivity is set to 40 packages per time slot and cost 185 Euros. Temporary workers produce up to 30 packages per time slot and cost 210 Euros. Workers are hired for at least one shift. We impose the limit on the number of permanent workers to 15 for each day.

The number of available docks is set to 10. The penalties for a postponement or a mode change are as follows:

$$p_{hd}^{\bar{h}v} = \begin{cases} 0 & \text{if } \bar{h} = 0 \text{ and } v = v_{hd}, \\ 1 & \text{if } \bar{h} = 0 \text{ and } v \neq v_{hd} \text{ or } \bar{h} = 1 \text{ and } v = v_{hd}, \\ 2 & \text{if } \bar{h} = 1 \text{ and } v \neq v_{hd}, \\ \infty & \text{otherwise.} \end{cases} \tag{2.85}$$

We consider nine type of instances associated with all possible profile combinations for day 0 and day 1 chosen among low, normal and high profiles.

For each type of instance, five particular instances are generated randomly fixing the values of the release date, the number of packages that constitute an order as well as their delivery mode. Order volumes are uniformly drawn among values $\{1, 2, 3\}$ and release dates are drawn uniformly among the slots of the day. Modes are initially assigned to orders according to a uniform distribution.

2.6.2 Discussion

The algorithm is implemented in `C++` in Visual Studio environment. The models presented in Sections 2.5.1–2.5.3 are solved with Cplex 12.6. All tests are performed on an Intel® Core™ i7-4600U CPU 2.10 GHz. We allow a maximum of one hour of computation for each phase of the algorithm.

2.6.2.1 Results on the basic-instances

We run our three-phase algorithm on the nine basic-instances described in Section 2.6.1. Detailed results are reported in Table 2.1. Column *Instance* reports the name of the instance's type as a couple corresponding to the profiles of day 0 and 1. For each instance, we report results in four rows: the first three rows correspond to each of the algorithm phases. The forth row give total cost and time values.

Column *Phase* indicates the considered phase of the algorithm. Column *Cost* reports the cost of the objective function for each phase as well as the total solution cost. Note that the objective function at Phase I takes into account the penalties that occurs when setting the workforce. But after this phase, only the sum of workers cost is taken into account in the value of the final solution. Penalties are computed again in Phase II and then contribute effectively to the final solution cost. Columns *workers Per* and *workers Temp* report respectively the number of permanent and temporary workers. Note that these columns are blank in correspondence to phase II and phase III, as they are determined definitively during phase I. Column *Pen* reports the value of the sum of all postponement and mode change penalties. Column *Truck* indicates the number of trucks needed to deliver the orders. Phase III does not modify the values of these two columns and the corresponding slots are left blank. Column *Dock-slots* reports the docks occupation during operations. The solution of the Phase I model does not provide the values of *Truck* and *Dock-slots* columns (variables u_h^{vt} and w_h^{vt} are not present in model IPSP I). We compute these values at Phase I separately by an independent procedure for comparison purposes. Column *Gap* indicates the optimality gap observed when the respective model is not solved to optimality. Since we enhanced the solver with a specific lower bounds in Phase I and II, the reported gap is provided using either the value of the linear relaxation or the corresponding lower-bound (see Section 2.5.4). The gap is reported only when

it is strictly positive. When the slot is blank, an optimal solution (for that phase) is identified. Finally, column *Time* provides the CPU time in seconds.

For four type of instances and for each of the three phases, a (local) optimal solution is obtained. For other five types, phase II fails to reach the optimal solution within one hour of CPU time. However, the optimality gap is less than 2% for four of them. Phase II reveals to be the bottleneck of the procedure. From the total computation times, we observe that the algorithm can solve more easily instances with the same profile for both days.

For four types of instances, Phase I suggests to hire temporary workers even if the total availability of permanent workers for the first two days has not been used. We recall that 15 permanent workers are available for each shift, for a total of 60 workers for day-0 and day-1. The penalty scheme considered guide the optimization through solutions that favour temporary workers hiring, rather than order postponement.

For seven types of instances, phase II is able to reduce the number of used trucks that was first determined in Phase I. This is possibly due to the mode change strategy the company defines. Note that the increase in the penalties cost is always lower than the saving due to unused trucks. This result highlights the potential benefit of incorporating postponement and mode change in the process planning.

Finally, Phase III always reduces the number of dock-slots which is crucial to handle high activity peaks.

2.6.2.2 Algorithm behaviour analysis on instances of the Normal-Low type

We run our algorithm on five instances of the type Normal-Low. Since on E-commerce enterprise experiences often the same sequence of day profiles, but with different orders quantities. One of these usual sequence is the Normal-Low sequence that we selected to analyse the sensitiveness of the algorithm. Table 2.2 reports detailed results on the five runs. Column names correspond to those reported in Table 2.1. It can be seen that the results for different instances are equivalent. It is worthwhile to note that he solution time for phases I and III do not vary among instances significantly. The only significant variation is on the computation time for phase II of the fifth instance. We can conclude that our solution algorithm is not deeply impacted by the structure of the particular instance solved.

We made the same analysis for the other instance type solving each time five instances, and we ended up each time by the same conclusion for the Normal-Low type. Thus we omit to report detailed results on these cases.

2.6.2.3 Analysis of penalty schemes

In this section we compare the results obtained introducing different penalty schemes for the postponements and mode changes. In Table 2.3 we report results obtained when penalty values in Equation (2.85) are divided by 10.

Instance	Phase	Cost	workers		Pen	Truck	Dock-slots	Gap	Time
			Per	Temp					
Low-Low	I	2405 + 75	13	0	75	4	56		66
	II	2675			75	4	52		137
	III	45					45		1
		5125							204
Low-Normal	I	4810 + 75	26	0	75	8	60		117
	II	4665			115	7	59	1.61%	3602
	III	46					46		3
		9521							3722
Low-High	I	7290+75	36	3	75	11	63		94
	II	6735			235	10	58	1.11%	3602
	III	40					40		3
		14065							3699
Normal-Low	I	4810	26	0	0	8	60		13
	II	4960			410	7	48		269
	III	38					38		6
		9808							288
Normal-High	I	9671 + 10	50	2	10	15	76		235
	II	9550			450	14	54	7.23%	3610
	III	46					46		83
		19267							3928
Normal-Normal	I	7215	39	0	0	12	64		14
	II	7076			576	10	48		1549
	III	39					39		7
		14330							1570
High-Low	I	7290 + 2	36	3	2	10	62		31
	II	6501			1	10	57	0.02%	3626
	III	43					43		14
		13834							3671
High-Normal	I	9485+1	49	2	1	14	66		28
	II	8746			296	13	57	0.01%	3625
	III	45					45		19
		18276							3672
High-High	I	12150	60	5	0	17	69		30
	II	10560			160	16	61		525
	III	44					44		16
		22754							571

Table 2.1: Results on the basic-instances

Phase	Cost	workers		Pen	Truck	Dock	Gap	Time
		Per	Temp					
I	4810	26	0	0	8	60		13
II	4960			410	7	48		269
III	38					38		6
	9808							288
I	4810	26	0	0	8	59		19
II	4893			343	7	54		369
III	37					37		6
	9740							394
I	4810	26	0	0	8	60		13
II	4851			301	7	57		135
III	35					35		5
	9696							153
I	4810	26	0	0	8	59		13
II	4859			309	7	56		96
III	37					37		6
	9706							115
I	4625+114	25	0	114	8	60		13
II	4871			321	7	58		2684
III	37					37		5
	9533							2702

Table 2.2: Algorithm performance on 5 instances created from the same basic-instance Normal-Low

There are two main differences with the results reported in Table 2.1. The first is related to the number of temporary workers while the second concerns the higher optimality gap.

For the High-High instance, the algorithm a solution obtained postpones the order process to the third day (day 2) with a consequent use of 3 permanent workers. Note that 3 permanent workers guarantee a production (40 packages per slot times 8 slots times 3 equals to 960) equal to 4 temporary workers (30 times 8 times 4), but cost 555 instead of 840, leaving room for a large postponement that is favored by the low penalisation scheme adopted. In a rolling-horizon context, poor solution could considered if the profile of day 2 is high.

Moreover larger optimality gaps is obtained in Phase II. An explanation could be the following. Let us consider two orders d_1 and d_2 for the same day and with the same volume and the same release date. Let us suppose to have in hand the complete planning. Exchanging production of d_1 with d_2 would provide an equivalent planning. This leads to equivalent solutions which the solver needs to consider to prove optimality. When penalty is low, this symmetry is projected to the possibilities of postponement and mode change, making computation even harder.

Table 2.4 reports results where the penalty scheme proposed in Equation (2.85) is modified in order to change orders from the normal to the express mode for free (even if associated with postponement). On one side earlier deliveries increase the company's image. On the other side a joint postponement coupled with a change to a faster mode provide an on-time delivery. Similar observations as those formulated for Table 2.3 can be drawn. Inexpensive mode changes and postponements make disadvantageous to hire temporary workers and increase solution symmetry. The latter leads to significant optimality gaps that are reported in the Table.

2.6.2.4 Lower-bound effectiveness

In Table 2.5 we report the deviations of the lower-bounds defined in Sections 2.5.4.1–2.5.4.2 on the instances computing considered in Table 2.1. Columns *Instance* and *Phase* are self-explanatory. Column $Cplex_{gap}$ reports the gap value between upper- and lower-bounds provided by Cplex 12.6 when the solution of the corresponding phase is stopped. Column LB_{gap} indicates the gap value of the lower-bound computed solving the related arc-flow problem. When the gap is null, the cell is left blank.

A blank value (null gap value) in column LB_{gap} associated with a positive value in column $Cplex_{gap}$ certifies the effectiveness of the lower-bound that is used to stop the corresponding model solution. It can be seen that the lower-bound for the phase I (LB1) allows for an earlier stop of the computation 4 times, while the lower-bound for phase II (LB2) does it in 3 cases. When the optimal values are not reached, LB2 provides systematically a better optimality gap compared to the one given by $Cplex_{gap}$.

Instance	Phase	Cost	workers		Pen	Truck	Dock	Gap	Time
			Per	Temp					
Low-Low	I	2405 + 7.5	13	0	7.5	4	56		18
	II	2607.5			7.5	4	40		314
	III	35					35		1
		5047.5							333
Low-Normal	I	4810+7.5	26	0	7.5	8	60		3
	II	4659.2			109.2	7	44	2.26%	3602
	III	35					35		4
		9504.2							3609
Low-High	I	7030 +75.5	38	0	75.5	12	89		5
	II	7235.1			85.1	11	74	5.40%	3602
	III	66					66		7
		14331.1							3614
Normal-Low	I	4810.0	26	0	0	8	60		13
	II	4591.0			41.0	7	51		214
	III	38					38		6
		9439							233
Normal-High	I	9435 +49.1	51	0	49.1	16	92		16
	II	9193.0			93.0	14	73	7.64%	3610
	III	58					58		7
		18686.1							3633
Normal-Normal	I	7030+28.5	38	0	28.5	12	64		15
	II	6557.6			57.6	10	61		1063
	III	39					39		7
		13626.6							1085
High-Low	I	7030 +80.1	38	0	80.1	11	63		32
	II	7230.1			80.1	11	61	10.10%	3627
	III	44					44		20
		14304.1							3679
High-Normal	I	9460+48.1	50	1	48.1	14	66		38
	II	9183.0			83.0	14	60	7.66%	3628
	III	46					46		17
		18689.0							3683
High-High	I	11655+140.1	63	0	140.1	18	96		45
	II	11212.5			162.5	17	80	7.10%	3627
	III	66					66		15
		22933.5							3687

Table 2.3: Results on reduced penalties

Instance	Phase	Cost	workers		Pen	Truck	Dock	Gap	Time
			Per	Temp					
Low-Low	I	2405	13	0	0	6	68		266
	II	3457			207	5	44		75
	III	3291					41		1
		9153							342
Low-Normal	I	4810	26	0	0	9	72		3
	II	5200			0	8	68		662
	III	5244					44		2
		15254							667
Low-High	I	7030+3	38	0	3	11	74		3
	II	7150			0	11	69		1887
	III	7198					48		2
		21378							1892
Normal-Low	I	4810	26	0	0	8	60		28
	II	4893			343	7	50	7.01%	3609
	III	4587					37		7
		14290							3644
Normal-High	I	9435	51	0	0	14	73		16
	II	9100			0	14	66	7.14%	3609
	III	9152					52		5
		27687							3630
Normal-Normal	I	7030+54	38	0	54	12	75		399
	II	7150			0	11	57	9.09%	3610
	III	7195					45		7
		21375							4016
High-Low	I	7030	38	0	0	11	74		494
	II	7150			0	11	69	5.03%	3626
	III	7198					46		14
		21378							4134
High-Normal	I	9435	51	0	0	14	77		65
	II	9100			0	14	70	7.14%	3624
	III	9149					49		15
		27684							3704
High-High	I	11840	64	0	0	18	81		52
	II	11700			0	18	80	11.11%	3625
	III	11755					55		15
		35295							3692

Table 2.4: Results on free normal to express change

Instance	Phase	$Cplex_{gap}$	LB_{gap}
Low-Low	I	5.19%	
	II		2.80%
Low-Normal	I	4.36%	
	II	4.34%	1.61%
Low-High	I		3.53%
	II	3.78%	1.11%
Normal-Low	I	2.87%	
	II	12.85%	
Normal-Normal	I		2.54%
	II	14.46%	7.23%
Normal-High	I	2.80%	
	II	14.28%	
High-Low	I		1.39%
	II	2.31%	0.02%
High-Normal	I		0.54%
	II	7.60%	0.01%
High-high	I		2.75%
	II	4.85%	

Table 2.5: Lower-bound effectiveness

2.6.2.5 Comparaison with commercial solver

In this section we report on the comparison between our algorithm and the commercial solver Cplex 12.6. The result on the complexity of the IPSP suggests that only small size instances can be solved to optimality.

In Table 2.6, columns *Cplex* report the results obtained by the Cplex 12.6, while columns *Three-phase* report the results obtained by our algorithm. Columns *CPU* report the computational time in seconds. Columns *Cost* reports the value of the solution obtained. Finally, column *Gap* reports the gap between both solutions. Negative gaps correspond to better solution obtained by the three-phase algorithm. A time limit of 8 hours of computation is given to Cplex 12.6. We report the value of the solution found by the three-phase procedure and up to four solution values related to the Cplex solution. In particular, we report 1) the value of the first feasible solution found by Cplex 12.6; 2) the value of the solution Cplex 12.6 found after the CPU time required by the three-phase procedure; 3) the value of the first improved solution; 4) the value of the solution when CPU time limit is reached. In the last case, when the solution is optimal it is indicated by an asterisk. When Cplex 12.6 does not find a better solution than the three-phase algorithm only three values are reported. A dash indicates that Cplex 12.6 was not able to find a solution.

For Normal-Low type instances, the three-phase procedure takes 5 minutes to get a solution. Cplex 12.6 after 5 minutes has not got a feasible solution, it finds its

	Three-phase		Cplex		Gap
Instance	Cost	Cpu	Cost	Cpu	Cost
Low-Low	5125	3 mins	5243	1 min	-2.3%
			5102	2 mins	0.4%
			5102	3 mins	0.4%
			5099*	12 mins	0.5%
Low-Normal	9521	1 hour	10654	1 min	-11.9%
			9720	1 hour	-2.1%
			9720	8 hours	-2.1%
Low-High	14065	1 hour	15514	4 mins	-10.3%
			13910	5 mins	1.1%
			13896	1 hours	1.2%
			13896	8 hours	1.2%
Normal-Low	9808	5 mins	-	5 mins	-
			10642	8 mins	-8.5%
			9743	40 mins	0.7%
			9730	8 hours	0.8%
Normal-Normal	14330	25 mins	15211	25 mins	-6.1%
			14427	8 hours	-0.7%
Normal-High	19266	1 hour	19538	9 mins	-1.4%
			19235	45 mins	0.2%
			19235	1 hour	0.2%
			18527	8 hour	3.8%
High-Low	13834	1 hour	17362	22 mins	-1.5%
			13816	35 mins	0.1%
			13804	1 hour	0.1%
			13799	8 hours	0.1%
High-Normal	18277	1 hour	19766	13 mins	-8.1%
			18550	1 hour	-1.5%
			18265	2.5 hours	0.1%
			18262	8 hours	0.1%
High-High	22754	10 mins	-	10 mins	-
			26285	58 mins	-15.5%
			22994	8 hours	-1.1%

Table 2.6: Comparison with Cplex 12.6

first feasible solution after 8 minutes (the three-phase solution is 8.5% better). The first improving solution is found after 40 minutes.

For two types of instances, Cplex 12.6 has not found a feasible solution before the end of the time required by the three-phase solution (Normal-Low and High-High). Note that for three types, after 8 hours of computation the commercial solver is not able to improve the solution found by the three-phase procedure. On the other side, for four types Cplex 12.6 behaves better than our algorithm, even if for three types the two approches can be considered equivalent due to the small improvement provided by Cplex 12.6: at most 0.4%. In seven cases Cplex 12.6 has found a solution on the same time required by our algorithm. The latter provides better solution with an average decrease of the solution value of 1.3%. better.

2.7 Conclusions

In this paper we introduced the Packaging and Shipping Problem (IPSP) arising in E-commerce logistics. It consists in determining the number of employees required to precess a set of orders in a multi-day horizon setting. Furthermore, the problem asks to produce an operational planning to process the orders and load the packages into trucks for delivery that can be performed with different modes. We introduced two strategies in order to obtain overall solutions with a lower cost: *mode change* and *postponement*. The first strategy consists in changing the delivery chosen by the customer to another. The second consists in processing the order in a day later than to the one of arrival. These strategies generate penalties but they can lead to hire less employees or to use less trucks and consequently result into savings for the company.

We proposed a mathematical model for the IPSP and proved that the IPSP is NP-hard. It is then unlikely the IPSP can be efficiently solved to optimality in a reasonable time regardless the size of the instances (unless $\mathcal{P} = \mathcal{NP}$). We then proposed a three-phase matheuristic approach that allows to deal with large real-life instances. Our approach exploits the structure of IPSP by sequentially solving three sub-problems of IPSP to construct a feasible. We first take the tactical decisions, fixing the work force for each day, and consequently we determine the operational planning. Moreover, our approach is enhanced with speed up techniques based on lower-bounds for the sub-problems.

We created a set of instances for the IPSP based on data provided by our industrial partner. Instances with up to 5000 orders per day are then solved by the three-phase procedure we proposed. Results show the efficiency of the method which can provide high-quality solutions in a reasonable amount of time and performs significantly better than the commercial solver.

Future work could consider the stochastic nature of the problem. In this paper, we consider that all orders information are deterministic. In real life, total demand is only forecasted for the following days and, consequently, exposed to variations.

Since we have interaction between decision of consecutive periods, future demand uncertainty should be taken into account in the decision making process. We suggest to apply rolling horizon based procedure that fits well data acquisition and decision making dynamics in e-fulfillment, and to investigate appropriate stochastic optimization techniques.

Chapter 3

Dynamic optimization with rolling horizon

The works presented in this chapter were published in the conference Tristan 2016 [Tounsi 2016b].

Contents

3.1 Introduction

Chapter 2 was dedicated to the theoretical and computational study of the integrated picking and shipping problem. The proposed model suffers from two main drawbacks. First, the complexity of the problem is an obstacle for its scaling. It was shown that even for instances with a planning horizon of two periods, the computational effort is excessive. Second, we assumed that all the problem data are deterministic, including future orders, which is not true. Indeed the decision making process incorporates a part of uncertainty that has been ignored in IPSP.

In real life, orders process is conducted by continuously taking decisions of resources assignment and operations planning according mainly to orders arrivals. Over a horizon of several days, decisions are determined at regular frequency and decisions related to different periods are not independent.

At the beginning of each day, orders process plans and the level of required resources need to be determined. Decisions related to the current period take into account the orders received at that period. They also depends on decision taken during the last period and on the expected state of the system at the following period. For example, the decision of postponing an order to a future period change the e-fulfilment plan of that period. Massive orders postponing to a future period induces the use of additional resources during that period or a cascade of orders postponing. The process of repetitive data updates and decision making characterizes a *dynamic* problem.

On the other side, all data are not known at the moment of decision making, like in a classical deterministic problem. While part of the problem data is known all over the planning horizon, like truck capacities for example, data like the number of orders to process is updated at the beginning of each period. Some decision, like the process of a giving order at a given period, is taken when all data related to the period are fully known. While the decision of postponing the order to the following period is taken without knowing exactly the orders arriving at the beginning of that period. Thus the optimization problem at each period embeds uncertainty.

To treat these issues, we introduce in this chapter a dynamic model with rolling horizon for the *multi period picking and shipping problem with stochastic demand*, or MPSSD problem for short. The proposed model differs from IPSP in that it determines the daily tactical decisions: number of workers, number of trucks and the quantities of orders processed in the current period. Tactical optimization integrates operational constraints like the docks limitation to guarantee feasible solutions in practice. Focussing only on tactical decisions allows us to aggregate orders by delivery mode. Based on the decisions given by the model, IPSP can be then solved to determine the daily operational decisions (the assignment of each order to a slot for process). The combination of the two models leads thus to a decomposition of the global problem that helps to tackle big instances.

Numerous applications are modelled as multi-period problems with

rolling horizon, like for example vehicle routing problems [Cordeau 2015] [Albareda-Sambola 2014], trains scheduling [Meng 2011] or gaz delivery [Rakke 2011].

Procedure with rolling horizon are deterministic in [Cordeau 2015] and [Rakke 2011] and they are motivated by the impossibility to tackle the considered optimization problem over the desired horizon. In contrary, in [Meng 2011] [Pironet 2014] the authors include in their rolling horizon procedure stochastic data.

The organization of the chapter is as follows. The multi-period procedure with rolling horizon is described in section 3.2. We provide a formulation of MPSSD in section 3.3. In section 3.4, fully deterministic heuristics without use of information on future orders are proposed. The section 3.5 is dedicated to the stochastic optimization approaches selected for the studied application. Numerical experimentations and results are given in section 3.6. Finally, section 3.7 concludes the chapter.

3.2 Multi-period rolling horizon procedure

We describe in this section the rolling horizon mechanism and we give the detailed formulation of MPSSD problem solved at every period.

3.2.1 Rolling horizon mechanism

The solution of MPSSD is constructed gradually using a sequence of solutions obtained by considering iteratively the problem over a restricted horizon called the rolling horizon. More precisely, at the beginning of the procedure the rolling horizon is positioned at the first period of the planning horizon. At each iteration a solution of MPSSD is solved and the solution is used to determine definitively a subset of decisions. Then the rolling horizon is shifted by one period and the process is iterated until reaching the end of the planning horizon. A typical rolling horizon procedure is illustrated by Figure 3.1

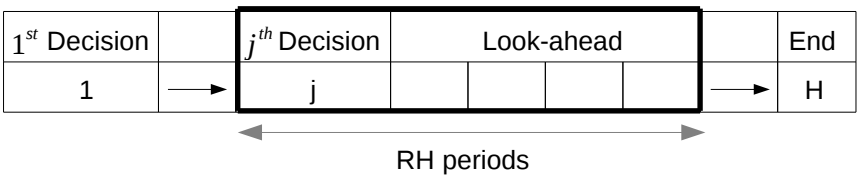

Figure 3.1: Rolling horizon procedure

At period j the rolling horizon contains the following parts:

- A *decision* part refers to the current period j because it is associated to a new solution of MPSSD. Decisions of the solution are not all transformed in actions,

mean definitely fixed in the final solution. In this chapter only decisions related to the current period j are definitively implemented in the final solution. Decisions related to the others periods of the rolling horizon are renewed at the next decision step. In other applications, during the current period, decisions related to a following period are taken. This creates a frozen part in the rolling horizon.

- A *look-ahead* part is composed of periods from period $j+1$ to $j+RH-1$. Data in this part can be either deterministic or stochastic. Since decisions of different periods are not independent, including the *look-ahead* part data in MPSSD helps in making the "best" decisions during the current period.

- There are periods starting from period $j+RH$ up to the last period H , i.e. periods in the remote part of the horizon. These periods are not taken into account for the optimization problem related to actions of the current period j.

The rolling horizon procedure is relevant when there are interactions between periods. In the e-fulfillment context, if the process of a quantity of orders requires additional worker or a truck, it can be advantageous to postpone it to the next period. But if it appears that the postponed quantity requires additional resources at the next decision step, the previous postponing decision should not have been taken.

At every iteration, decisions are made based an a solution of MPSSD that takes into account informations of RH periods. The obtained solution is analysed and some actions issued from this solution are performed. The costs or rewards of these actions are recorded into the final solution value. This concludes the decision phase for the decision period j. Then, data are updated when the horizon rolls from j to j + 1:

- data enters the rolling horizon, typically within the period j + RH

- data status is modified (e.g. stochastic data becoming deterministic or probability distribution modification)

- data are discarded from the rolling horizon as they are outdated, typically those from period j.

Finally, a new optimization process takes place in period j + 1 and the sequence of decisions and updates is repeated.

The presented rolling horizon structure in not unique. In some application, at each iteration the rolling horizon includes periods between the current one to the end of the planning horizon. Thus at every decision step the horizon length decreases. For example, the planning of a working day is optimized every hour for requests coming in along the day. In that case, the horizon is set initially to 8 hours, and it reduces by one hour at every decision step and it reaches one period at the final decision.

3.2.2 Policy Vs solution

In the rolling horizon procedure, the actions performed over the long term are decided based on a sequence of solutions of MPSSD over restricted horizons. This is an important feature of the multi-period process compared to a classical optimization problem solved globally where all decisions are jointly determined. Thus the decision making process does not look for a particular optimal solution, but rather for a policy leading to determining actions at decision periods. The value of the policy is computed as the cumulated cost or rewards of successive actions. This value is not sum of all objective function values over all rolling horizons, since at each decision period, the performed actions are a restricted portion of the decision variables of the solution over the rolling horizon.

Indeed, different algorithms may be available to compute a solution at every decision step. This solution might be issued from rules, exact methods, heuristics or meta-heuristics. Moreover, different approaches treat the stochastic data of the look-ahead part differently. The set of actions performed can vary according to the solution generated by these approaches. The aim is to find the algorithm that provides the best policy.

The choice of algorithm is far of being trivial for different reasons. First the algorithm that provides the best solution over the rolling horizon is not necessarily the one that provides the best policy over the planning horizon. As mentioned, the value of the policy is not the sum of all objective function of all generated solutions. Second the problem addressed at every decision step can be a stochastic problem, like the one considered in this chapter. In this case it is already hard to define the best stochastic approach for a single solution. Finally, in a practical context, the algorithm is not only evaluated over the quality of the solution, but also regarding computing time.

3.2.3 Start and end of horizon biases

In a multi-period optimization procedure every decision period is impacted by previous actions. This applies also for the first period. But it is difficult to include past decisions in the initial periods. Ignoring the initial conditions can induce a bias in the policy valuation. A remedy is to remove these initial periods of the whole horizon, and to consider them as a warm-up delay: we run the procedure up to certain period that is considered a representative of the system's steady state regarding past effects and information inside the rolling horizon.

In the other hand, the evaluation is necessarily stopped at a final period. As the procedure approaches the final period, decisions are taken based on a decreasing horizon. These decisions are also biased by the particular final conditions. To reduce this bias, The planning horizon H should be long enough to dissipate the biased extra cost or reward. In our simulations, solution values are considered without considering a number of periods at the end of the planning horizon.

3.2.4 Rolling horizon length

It is obvious that the length of the rolling horizon highly impacts the policy' quality. Since there are interactions between periods, considering jointly decisions of several successive periods improve the overall policy. In the same time when deciding at a given period, informations of periods far away in the horizon may not be useful. Thus the calibration of the rolling horizon length has to take into account the decision process of the considered application.

Moreover, as the rolling horizon increases, the size of MPSSD increases. Thus the rolling horizon length should be such a solution of MPSSD is obtained in an appropriate time, mostly when the planning horizon is big.

For those reasons the rolling horizon length in an important issue and should carefully calibrated by analysing different possibilities (see section 3.6).

3.2.5 Bounds with fully revealed information

The dynamism and uncertainty inherent to the multi-period problem make that the optimal solution cannot be reached. It is then important to define criteria for evaluating the quality of policies and assess. One classical way is to use bounds computed assuming that data is revealed [Pironet 2014]. In this chapter, we use:

1. The a-priori or myopic bound L_0 based on a solution of the multi-period problem only using the information from the deterministic part of the horizon. This value is associated to the myopic or a-priori policy.

2. The rolling horizon a-posteriori bound L_R obtained by the solution of the multi-period problem assuming fully revealed information over the rolling horizon RH. This value is associated to the rolling horizon a-posteriori policy.

3. The optimal a-posteriori bound L^* based on a single solution assuming the information fully revealed over the whole horizon H. In this case, the value of the policy is equal to the value of the solution over the entire horizon, an it is associated to the a-posteriori optimal policy. When it is impossible to obtain a solution over the entire horizon, L^* is obtained using the solution using the fully revealed information over the longest solvable rolling horizon.

These bounds using deterministic information are easier to compute than solving the real stochastic problem. The optimal a-posteriori one L^* outperforms naturally the a-priori solutions L_0, and we have that $L^* \leq L_0$ for a minimization problem (and conversely for a maximization one).

In the same sense, the value L_R is assumed to be in between the lower and the upper bound. Yet, this cannot be claimed as sometimes the myopic policy performs better than the rolling horizon a-posteriori one (see [Pironet 2014]). Therefore we

can state for a minimization problem that "usually":

$$L^* \leq L_R \leq L_0.$$

For maximization problem the equation is in reverse order. In the case where the multi period problem is stochastic, like in this chapter, each bound is estimated by the average value over the set of considered scenarios. We keep in this chapter the former notations to refer to these estimates.

3.3 Mathematical formulation

We provide in this section the mathematical formulation of MPSSD defined over the rolling horizon RH. MPSSD is solved for each period of the planning horizon H. At every solution over the rolling horizon, the decision related to the first period are implemented in the final solution. Some decision are the process of orders during the first period, while some decisions are the postponement of orders from the first period to the second period. These orders need to be saved and injected as input during the following decision step. More over, a postponed order can not be postponed a second time. Thus the process of these orders is then different from the process orders revealed at each period. $\mathcal{D}_v$ denotes the number of postponed orders assigned to mode v during the previous solution.

A period $j = 1, ..., RH$ represents a day in practice and it is associated with a set of new orders n_{jv} for each mode $v \in \mathcal{V}$, $\mathcal{V}$ being the set of modes.

All periods have the same time representation. A period is divided into T time slots, a slot represents an hour in practice. This fine time granulation is needed to take into account departure time of each mode t_v, which is assumed to be same all along the horizon, and to determine at each slot the state of the docks (occupied or free). More over, a period is divided into a number of consecutive and disjoint shifts for workers management. Figure 3.2 shows an example of period structure with 8 slots and 2 shifts.

We also consider that all along the planning horizon, and for all shipping modes, trucks are of the same capacity Q and the number of docks is N_{max}.

The set of working shifts is labelled $\mathcal{S}$. A shift $s \in \mathcal{S}$ starts at time slot h_s^d and finishes at time slot h_s^f. Besides the L permanent workers present at every shift, temporary workers can be hired at every shift for a cost cw_t per worker. The productivity of a worker is the quantity of picked orders per time slot. It is b_p for a permanent b_t for a temporary. The design of workers in MPSSD does not concern the permanent workers who are already determined. However the model takes into account their number and productivity to adjust the permanent workers number at every shift.

Orders are initially assigned to a delivery mode chosen by the customer. To enhance flexibility, an order may be assigned to a mode different from the one selected by the customer or postponed to a future period. Postponement and mode

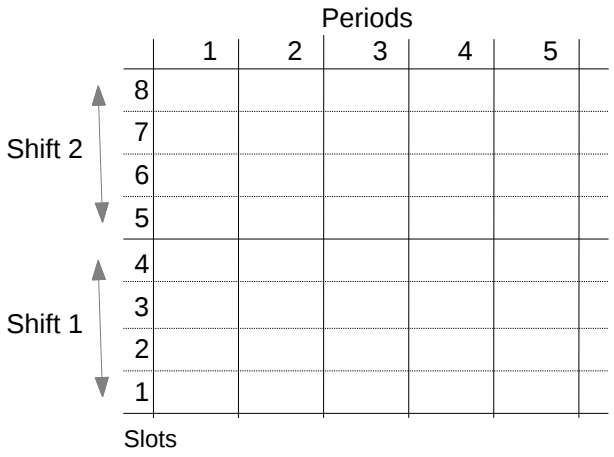

Figure 3.2: The time discretization

change are penalized according to a general penalty pattern $p^{\bar{j}}_{v\bar{v}}$, with $v, \bar{v} \in \mathcal{V}$ and $\bar{j} \in \{0, 1\}$. These actions can result in delivery delay or in an extra-cost for the company, however they can lead to reduce the global labour cost and transportation cost. A postponement happens when $\bar{j} = 1$ while a mode is changed when $v \neq \bar{v}$. In this chapter we allow postponement for only one period. The Figure 3.3 summarizes all possible mode assignments when the customer mode is mode 1 (left) and mode 2 (right). The green assignments respect the customer mode, the red ones generate a delay penalty and the blue ones a mode change penalty.

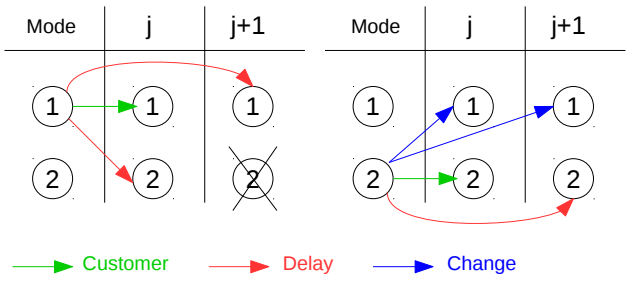

Figure 3.3: Penalty pattern

We define the following variables for all $j \in \mathcal{H}$, $v \in \mathcal{V}$, $\bar{j} = \{0, 1\}$, $\bar{v} \in \mathcal{V}$, $t = 1, ..., T$, and $s \in \mathcal{S}$

$f^{\bar{j}\bar{v}t}_{jv}$ Number of orders associated to mode v and period j assigned to period $j + \bar{j}$, mode $\bar{v}$, and time slot t;

$F^{\bar{v}t}_{v}$ Number of postponed orders associated to mode v assigned to mode $\bar{v}$ at time slot t;

u_{jv}^t Number of empty trucks for mode v of period j in time slot t ;
k_{jv}^t Residual capacity for mode v of period j in time slot t;
w_{jv}^t Number of used docks for mode v of period j in time slot t;
y_{jv}^t equals 1 if the number of empty truck for mode v during period h at slot $\bar{t} \geq t$ is not null, 0 otherwise;
z_{js} Temporary workers hired in shift s of period j.

The objective function (3.1) minimizes the sum of temporary workers cost, trucks cost and penalties. Constraints (3.2) (respectively (3.3) ensure that orders at each period of rolling horizon (the postponed orders during the previous optimization) are processed. Constraints (3.4) and (3.5) are the packages flow conservation equations for each mode at each time slot; packages are loaded in a residual capacity of a truck or in an empty truck. Note the particular formulation of this constraint for the first period (3.4) where the postponed orders are processed. Indeed these orders are not allowed to be postponed a second time. Moreover, constraints (3.4) and (3.5) have a particular formulation for each first slot of each period.

Constraints (3.6)–(3.9) implement the truck movement policy. We remind that the goal of the policy is to compute the occupied docks at each slot (variables w_{jv}^t) under particular rules explained in chapter 2.

Constraints (3.10) state that at a given time slot the total number of docked trucks does not exceed the number of docks. Constraints (3.11) ensure that at each time slot, the total quantity of picked orders is less than the total worker productivity. Note again the particular formulation related to the first period of the rolling horizon involving the postponed orders. After solution over RH, $\mathcal{D}_v$ is updated with the number of orders initially assigned to mode v postponed from the first period to the second period.

The model integrates tactical decisions: the number of workers, the number of trucks decisions and the quantities produced and loaded at each slot for each mode. This integration enables taking into account operational constraints like a feasible truck movements plan.

The construction of the final solution

The solution of the multi-period e-fulfilment problem is constructed progressively through a sequence of solutions. At each iteration decisions related to the current period are determined. The optimization problem (3.1)-(3.12) is solved over the rolling horizon where the current period is the first one. Using the resulting solution, the following decisions are implemented:
- The number of temporary workers for each shift and the number of trucks for each mode of the current period.
- The assignments of postponed orders to modes at each slot of the current period.
- The assignments of new orders to modes at each slot of the current period.
- The movements of trucks on docks at each slot of the current period.

$$\min \sum_{j=1}^{RH}\sum_{v,\bar{v}\in\mathcal{V}}\sum_{\bar{j}=0}^{1}\sum_{t=1}^{T} p_{v\bar{v}}^{\bar{j}} f_{jv}^{\bar{j}\bar{v}t} + \sum_{v,\bar{v}\in\mathcal{V}}\sum_{t=1}^{T} p_{v\bar{v}}^{1} F_v^{\bar{v}t} + \sum_{j=1}^{RH}\sum_{s\in\mathcal{S}} cw_t z_{js} + \sum_{j=1}^{RH}\sum_{v\in\mathcal{V}}\sum_{t=1}^{T} ctu_{jv}^{t} \tag{3.1}$$

$$\sum_{\bar{j}=0}^{1}\sum_{\bar{v}\in\mathcal{V}}\sum_{t=1}^{T} f_{jv}^{\bar{j}\bar{v}t} = n_{jv} \qquad j=1,...,RH, v\in\mathcal{V} \tag{3.2}$$

$$\sum_{\bar{v}\in\mathcal{V}}\sum_{t=1}^{T} F_v^{\bar{v}t} = \mathcal{D}_v \qquad v\in\mathcal{V} \tag{3.3}$$

$$\sum_{\bar{v}\in\mathcal{V}} f_{1v}^{0\bar{v}0} + \sum_{\bar{v}\in\mathcal{V}} F_v^{\bar{v}0} + k_{1v}^{0} = u_{1v}^{0}Q \qquad v\in\mathcal{V}$$

$$\sum_{\bar{v}\in\mathcal{V}} f_{1v}^{0\bar{v}t} + \sum_{\bar{v}\in\mathcal{V}} F_v^{\bar{v}t} + k_{1v}^{t} = k_{jv}^{t-1} + u_{1v}^{t}Q \qquad v\in\mathcal{V}, t=2,...,t_v \tag{3.4}$$

$$\sum_{\bar{j}=0}^{1}\sum_{\bar{v}\in\mathcal{V}} f_{(j-\bar{j})v}^{\bar{j}\bar{v}0} + k_{jv}^{0} = u_{jv}^{0}Q \qquad j=2,...,RH, v\in\mathcal{V}$$

$$\sum_{\bar{j}=0}^{1}\sum_{\bar{v}\in\mathcal{V}} f_{(j-\bar{j})v}^{\bar{j}\bar{v}t} + k_{jv}^{t} = k_{jv}^{t-1} + u_{jv}^{t}Q \quad j=2,...,RH, v\in\mathcal{V}, t=2,...,t_v \tag{3.5}$$

$$\sum_{\bar{t}=t}^{t_v} u_{jv}^{\bar{t}} \le N_{\max} t_v y_{jv}^{t} \quad j=1,...,RH, v\in\mathcal{V}, 1\le t\le t_v \tag{3.6}$$

$$y_{jv}^{t} \le \sum_{\bar{t}=t}^{t_v} u_{jv}^{\bar{t}} \quad j=1,...,RH, v\in\mathcal{V}, 1\le t\le t_v \tag{3.7}$$

$$Qu_{jv}^{1} \le Qw_{jv}^{1} \qquad j=1,...,RH, v\in\mathcal{V}$$

$$Qu_{jv}^{t} + k_{jv}^{t-1} \le Qw_{jv}^{t} + Q(1-y_{jv}^{t}) \quad j=1,...,RH, v\in\mathcal{V}, 1<t\le t_v \tag{3.8}$$

$$Qu_{jv}^{t} + k_{j}^{t-1} - k_{jv}^{t_v-1} \le Qw_{jv}^{t} \quad j=1,...,RH, v\in\mathcal{V}, 1<t\le t_v \tag{3.9}$$

$$\sum_{v:t_v\le t} w_{jv}^{t} \le N_{max} \qquad j=1,...,RH, t=1,...,T \tag{3.10}$$

$$\sum_{v\in\mathcal{V}}\sum_{\bar{v}\in\mathcal{V}} f_{1v}^{0\bar{v}t} + F_v^{\bar{v}t} \le b_p L + b_t z_{1s} \qquad s\in\mathcal{S}, t=h_s^d,...,h_s^f$$

$$\sum_{\bar{j}=0}^{1}\sum_{v\in\mathcal{V}}\sum_{\bar{v}\in\mathcal{V}} f_{(j-\bar{j})v}^{\bar{j}\bar{v}t} \le b_p L + b_t z_{js} \quad j=2,...,RH, s\in\mathcal{S}, t=h_s^d,...,h_s^f \tag{3.11}$$

$$f_{jv}^{\bar{j}\bar{v}t}, F_v^{\bar{v}t}, u_{jv}^{t}, k_{jv}^{t}, w_{jv}^{t}, z_{js} \in \mathbb{Z}^{+} \tag{3.12}$$

- The quantities of orders postponed to the following period.

Using the solution the cost of the process during the current period can be computed. The cost of the current period includes the used temporary workers and trucks and the penalty corresponding to postponements and mode changes. The cost of the final solution is then increased by the cost of the current period.

The final solution of MPSSD over the planning horizon depends on the solution method selected to obtain each solution of MPSSD over the rolling horizon. We present in the following two families of solution approaches. The first are deterministic, in the sense that at every decision period, they only use the revealed information. The second family includes stochastic approaches that, at every decision period, exploit the distribution of the uncertain demand of the future periods.

3.4 Deterministic approaches

In the rolling horizon procedure proposed in this chapter, iteratively one period is considered and its decisions are determined (see figure 3.1). Since the final goal is to find the best solution over the planning horizon and decisions between periods are not independent, the decisions of a given period have to take into account the following periods. However, when deciding for a given period, if some informations related to the following periods are not available, it becomes difficult to take into account these periods. Deterministic approaches are used to determine each period decisions based only on the data revealed at that period, in this sense the deterministic policies are also called *myopic*. We present first myopic policies that are based either on a pessimistic or an optimistic assumption regarding the upcoming periods. Then we introduce a policy that operates order postponement to the next period based on their contribution to the resources cost of that period. Finally, we present a heuristic method that postpones a quantity of the orders of a given period based on the resources productivity level.

3.4.1 Pessimistic and optimistic policies

The pessimistic policy considers that the future is always not advantageous for orders postponement, in the sense that a postponement will not lead to reduce resources. As a consequence, the planner decides to fulfil all the orders of the current period during that period, and thus no postponement to the next period is operated.

Let L_0 refers to the value of the pessimistic policy solution. It is obtained by having the rolling horizon equal to one period and not allowing order postponement. The activity peaks are managed by operating mode change or by the increase of resources (temporary workers and trucks). The pessimistic solution is not expected to be effective, but it is easy to implement using reduced amount of data and easy to explain to practitioners. It also provides a benchmark against which other approaches can be evaluated.

A second policy is an optimistic policy that considers that the future is always advantageous for orders postponement, in the sense that postponement is automatically more advantageous than an increase of resource during the current period. This policy appears to be not efficient. The planner at a current period ignores totally the impact of massive postponement on upcoming periods and he ends up by high penalty cost and a high resources cost. Because of its bad performance on preliminary tests, the optimistic policy is not considered in the rest of the study.

Both the pessimistic and optimistic policies are based on an extreme assumption on the future when operating order postponement. We present now two deterministic approaches that look to determine, in a smarter way, what quantity of orders to postpone at each period.

3.4.2 Policy with linearised resource cost (PLRC)

In this section we introduce a deterministic policy, that we call the policy with linearised cost (PLRC), that can be used by the rolling horizon procedure at every decision period. The PLRC gives a solution of MPSSD over a horizon of one period.

When solving for a given period, it allows the postponement of orders to the following period. The decision to postpone an order takes into account the impact of its process on the resources cost during the following period.

More precisely, in addition to the penalty, the cost of postponing an order includes the contribution of the postponed order in the resources cost. The cost of this contribution is obtained by the linearised cost of resources. As resources are trucks and workers, the linearised cost includes two terms: $\bar{u}_t$ which is a truck cost divided by its capacity, and $\bar{u}_w$ which is the cost of a permanent worker divided by its productivity. Thus the PLRC defines the cost of postponing an order as follows:

$$\bar{p} = p^1_{v\bar{v}} + \bar{u}_w + \bar{u}_t$$

In conclusion, the solution of MPSSD using PLRC is obtained by setting the rolling horizon length to one period, allowing postponement with a cost of $\bar{p}$, instead of $p^{\bar{h}}_{v\bar{v}}$, per postponed order.

Obviously $\bar{u}_w + \bar{u}_t$ is different from the real resources cost which is not a linear function of processed orders. The contribution of a postponed order on the resources cost related to the next period depends on orders and resources configuration at that period. For example it can be null if the required resources are already available, or it can be high if additional resources are involved just to treat a postponed order. The suggested linearised cost coincides with the real contribution in resource cost if the resource is available and used at its maximum capacity.

Moreover, using PLRC, the solution of MPSSD combines the real resources cost, which is constant piecewise function, on the first period and a linearised cost on the second period. Thus the decision to add resource on the first period is triggered when the cumulated postponed orders reaches a certain threshold.

The linearised postponement cost $\bar{p}$ is a parameter representing a level of optimism of the policy at a given period regarding the following period. It also enables to take into account a weak information on the demand of the following period, when no exact information neither a good expected value are available. This is the case when just a trend, like a low activity or a high activity, is expected. In the first case only permanent workers will be used, while in the second temporary will be necessary. The linearised cost can then be adjusted depending on such forecasts.

In the next section, we present another deterministic approach based on a different policy for order postponement decision making.

3.4.3 Policy with resource productivity

The policy presented in this section is based, like the PLRC, on a rolling horizon of one period and it also allows orders postponement. The decision of postponing an order is taken in a different way, by considering the resource productivity at the current period. We call this approach the policy with resource productivity (PRP).

Roughly speaking, the PRP determines the number of each resource in two steps. First, a relaxed version of MPSSD is solved over the current period having that no postponement is allowed. Then one resource is considered and based on its productivity its final number is determined following branching rules defined in 3.4.4.1. This process is iterated until the determination of all resources number.

3.4.4 Resource productivity computation

We apply the PRP to only trucks as a first attempt. The resource productivity is computed by relaxing the integrity constraints on the number of trucks used by each mode. Then the productivity of trucks is a fractional value of the number of trucks in the obtained solution.

Due to interactions between modes, the PRP proceeds by considering separately all the modes. At every iteration, constraints on trucks number are relaxed for remaining modes. The solution of this relaxed MPSSD is used to determine the number of trucks for the considered mode following the branching rules defined in section 3.4.4.1. Once all modes are treated and all trucks numbers are fixed to integer values, an additional solution of MPSSD is carried to determine the final number of temporary workers needed at the current period and the quantities of postponed orders.

3.4.4.1 Branching

Iteratively, in the solution of the relaxed MPSSD the number of trucks corresponding to a subset of modes have real value. The number of trucks of one of these modes is considered to determine an integer value. Our approach is to round to the closest integer value: If the value is less than its integer value increased by 0.5 than the

truck number is fixed at the value of the integer value, else it is fixed at the integer value increased by 1.

A constraint imposing the considered number of trucks to its new integer value is added to the relaxed model, and this variable is definitely determined. The updated relaxed model is then solved for the next mode.

3.4.4.2 Feasibility correction

Fixing the number of trucks can make it impossible to find a feasible solution. This happens in the following situation. When deciding for period i, some orders, associated with the express mode for example, are postponed. At the following decision step, these orders have to be loaded in a truck associated with the express mode. However, the PRP can suggest to set the number of trucks of the express mode to 0. In this case, the orders associated with the express mode postponed at decision step i can not be processed, and no feasible solution is possible. Thus we define particular branching rules to avoid situations leading to infeasibility.

In the previous section, we presented different deterministic approaches that can be used as a decision policy in the rolling horizon procedure. At every decision period, these approaches use only the data revealed, and solve MPSSD over one period. Each of the PLRC and the PRP is based on specific mechanism to plan the resources and determine postponed orders quantities. We present in the next section another family of approaches that can also be incorporated in the rolling horizon procedure. These approaches belong to stochastic optimization and particularly to scenario-based techniques.

3.5 Scenario based approaches

Stochastic approaches address optimization problem where some of the data are uncertain or random. These problems occur in numerous application like in the *vehicle routing problem* where a provider has to determine a route or a set of routes to serve the demands of a set of clients with a capacitated truck at the least transport cost. In the stochastic version of the problem, some of the demands are not known in advance and are revealed at the arrival to the clients.

Modelling stochastic problems and dealing with data uncertainty differ from an application to another. The solution approaches depend on the available information related to the random data. These information are for example, the distribution of the random data, the values interval or a set of possible scenarios.

In our application, the uncertain parameters are the number of orders associated to each mode. As the rolling horizon advances, new information of the first period is revealed, while stochastic information related to the other periods are incorporated to MPSSD.

The scenario-based, also called sampling, techniques [Pironet 2014] for stochastic optimization problem are based on the insight that considering one scenario helps making decisions that take into account the upcoming periods. A scenario is a particular realization of the stochastic parameters.

The general framework for the studied stochastic approaches is the rolling horizon framework where the rolling horizon RH includes a number of look-ahead periods containing stochastic data. At every decision period, decisions are extracted from the solution of MPSSD over the rolling horizon associated with the selected scenario. We are interested in finding the technique that provides the policy that has a good performance over different possible scenarios.

The mono-scenario approximation has some advantages. It is easy to explain to practitioners and the data management is reduced. If there is an algorithm to solve the deterministic multi-period problem, this algorithm can be employed again for any other single scenario which is a deterministic outcome of the stochastic parameters. Nevertheless, the drawback is "how to find a representative scenario" if such a scenario exists. We present in the following classical scenarios that should be tested for any problem.

3.5.1 Expected value Solution

The expected scenario is build by replacing stochastic parameters by their expected solution. Usually, the expected scenario is supposed to be the most representative scenario. The expected value solution is the policy based on the expected scenario.

The expected value of the number of orders can be estimated based on old data. In e-fulfilment, the data of previous weeks can be analysed to determine the general shape of orders arrivals depending on each day of the week. Such study was conducted by [Boulanouar 2014] on data of three months and it shows that the distribution of orders number can be approximated by a Poisson distribution. We are then able to deduce the expected scenario over a week.

3.5.2 Random Value Solution

Even if we were able to build the expected scenario for orders arrival, when processing a new week, the data deviates from the expected scenario. We consider that the each random number of orders varies uniformly in an interval centred at the expected value. The random value solution consists in replacing each random parameter by a random value that corresponds to a possible realization.

3.5.3 Quantile Value Solution

Another interesting scenario is the quantile scenario where each random parameter r is replaced by particular value V that covers a fraction ξ of all possible scenarios. It is defined such as $P(r \leq V) = \xi$. When solving MPSSD using the quantile scenario,

the postponed quantities and resources plan are determined to satisfy the quantile value. So as long as the real orders number is less than the quantile there will be enough resources for the postponed quantity, and the postponement decision are likely to be good.

3.6 Numerical results

The multi-period formulation wit rolling horizon procedure was coded in C++ using concert technology and Cplex 12. Tests were performed on a 3.30GHz Intel Core processor with 3 GB RAM, under the Linux Ubuntu 12 operating system. A testbed of instances was constructed to assess the efficiency of the solution approaches.

3.6.1 Instances generation

The testbed was built on the basis of 2014 activity statistics provided by a major e-fulfillment company and through warehouse visits. The considered planning horizon is made of 8 periods to study a week planning. The last two periods are cooling periods: they are used to avoid end of horizon biases and then are not considered in the evaluation of the solution (see section 3.2).

3.6.1.1 Stochastic demand

First the expected scenario for order arrivals is built. Activity statistics show that orders arrivals fluctuate between three main levels: low, medium and high. Low level is associated with a nominal value of 500 orders, medium with 1000 and high with 1500. Based on the past data, the average distribution of demand levels is computed (see table 3.1).

	low	medium	high
Percentage	0.17%	0.69%	0.14%

Table 3.1: Average demand distribution

We obtain that the expected scenario of 8 periods includes one period of low demand, one period of high demand, and 6 periods of medium demand. For the e-fulfillment company, the management of the period with high demand and the period with low demand is challenging. Indeed, the first requires additional resources compared to other periods, while the second may lead to a low resource productivity. Consequently, for each of these periods, particular anticipatory strategies and recovery strategy are required the day before and the day after in order to manage demand fluctuation. For these reasons, in the considered expected scenario each of the period with high demand and the period with low demand is placed between two periods of medium demand (see in figure 3.4).

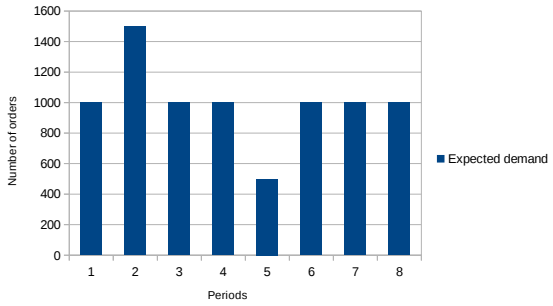

Figure 3.4: The expected demand scenario

As mentioned earlier, when real data are revealed, they deviate from the expected scenario. We suppose that the data vary uniformly in an interval centered on the expected value. We experiment different values for the standard deviation varying from 10% to 50% of the expected value.

3.6.1.2 Deterministic parameters

Each period is divided in 16 slots and 4 disjoint shifts. Shipping mode parameters are given in Table 3.2. The productivity of workers is 40 packages per slot for permanent workers and 40 packages for temporary workers. The latter cost 100 Euro per working shift.

Mode	Departure slot	Cost(euro)	Capacity
Express	8	430	1000
Standard	16	430	1000

Table 3.2: Delivery modes parameters

3.6.2 Algorithm Comparison

Table 3.6.2 reports the results of the solutions and bounds presented in this chapter. We provide results for the 5 levels of demand deviation with respect to the expected demand scenario The first line of the Table corresponds to the solution L_0, the myopic solution, that represents an upper bound on the solution of MPSSD. Two lines are associated with the others bounds and solutions.

The first line gives the total cost over the first 6 periods, while the second line gives the relative gap with respect to the myopic solution. This gap represents the improvement achieved by each solution. As the penalty value related to a postponement or mode change impacts the decisions taken, we compare the solutions for three different values $\{0, 0.5, 1\}$.

For each column and each method, the reported value corresponds to the average over 10 runs.

L_2 (respectively L_3) is the a-posteriori bound value assuming fully revealed information over a rolling horizon of 2 periods (respectively 3 periods). A-posteriori bounds are considered as lower bounds for the solution of MPSSD.

We report then the results of the solutions obtained by respectively PLRC and PRP. The four last lines of the table corresponds respectively to the mono-scenario approaches : mean scenario (L_{mean}), random scenario (L_{rand}), quantile of 75% (L_{quant+}) and quantile of 25% (L_{quant-}).

From Table 3.6.2 it can be observed that the multi period model improves the basic myopic solution. Taking into account the future period by allowing postponement results in cost reductions, for both deterministic policies or by stochastic techniques. The improvement achieved decreases when the penalties increase. In fact low penalties are an incentive to postpone orders, while high penalties reduce these opportunities. Thus the improvement is mainly due to orders postponement.

L_2 and L_3 provide almost the same results. Beyond one look-ahead period, the additional information on future demands is useless for the current decisions. We attract the attention of the reader to a specificity of multi-period problems with rolling horizon illustrated by the comparison of L_2 and L_3. Intuitively, when we compute a-posteriori bounds, we expect that the more periods are included in the rolling horizon, the better the solution is. However, the second column in Table 3.6.2 contradicts this intuition. It happens that the value for L_3 is worse than the value for L_2. The same column shows that L_3 is even worse than a L_{quant-}.

It is noteworthy the good performance of the PLRC. The policy reaches the best bound in one case and its average gap with respect to the best bound is 0.03%. In general the PLRC is close to the stochastic approaches while it requires less computational effort since it is based on solutions of MPSSD over one period. We also observe that the PLRC's performance is deteriorated when the penalties are high. For such values the postponement rules does not seem to be relevant.

Regarding the PRP, it appears to be very bad, even worse than L_0 in several cases. The postponement strategy of the PRP, based on resource productivity thresholds, is clearly not appropriate for our application.

On the other side, it is hard to distinguish the best mono-scenario technique. The solution L_{rand} outperforms the others stochastic approaches in 11 cases, and in 7 cases it reaches the best bounds. The quantile scenario L_{quant+} realizes the best performance in 10 cases, and in 4 cases it reaches the best bound.

Standard deviation	10%			20%			30%			40%			50%		
Penalty	0	0.5	1	0	0.5	1	0	0.5	1	0	0.5	1	0	0.5	1
L_0	7802	8067	8288	7782	8058	8305	7882	8174	8435	7942	8251	8522	8062	8403	8708
L_2	5160	6875	7241	5160	6968	7468	5250	7127	7638	5350	7229	7841	5410	7293	8028
	0.34	0.15	0.13	0.34	0.14	0.10	0.33	0.13	0.09	0.33	0.12	0.08	0.33	0.13	0.08
L_3	5160	6910	7223	5160	6953	7438	5250	7129	7626	5350	7208	7807	5410	7343	7982
	0.34	0.14	0.13	0.34	0.14	0.10	0.33	0.13	0.10	0.33	0.13	0.08	0.33	0.13	0.08
PLRC	5160	6951	7257	5263	7088	7622	5519	7240	8095	5466	7475	8245	5536	7886	8452
	0.34	0.14	0.12	0.32	0.12	0.08	0.30	0.11	0.04	0.31	0.09	0.03	0.31	0.06	0.03
PRP	6749	7788	10029	6736	7988	9846	6796	8189	10079	6674	8407	10228	6777	8593	10440
	0.13	0.03	-0.21	0.13	0.01	-0.19	0.14	0.00	-0.19	0.16	-0.02	-0.20	0.16	-0.02	-0.20
L_{mean}	5160	6911	7255	5160	7084	7630	5293	7324	7838	5393	7560	8040	5453	7698	8284
	0.34	0.14	0.12	0.34	0.12	0.08	0.33	0.10	0.07	0.32	0.08	0.06	0.32	0.08	0.05
L_{rand}	5160	6912	7239	5160	6912	7239	5293	7204	7730	5350	7340	7971	5324	7480	8204
	0.34	0.14	0.13	0.34	0.14	0.13	0.33	0.12	0.08	0.33	0.11	0.06	0.34	0.11	0.06
L_{quant+}	5160	6908	7255	5160	7023	7477	5293	7181	7722	5393	7249	7884	5453	7477	8182
	0.34	0.14	0.12	0.34	0.13	0.10	0.33	0.12	0.08	0.32	0.12	0.07	0.32	0.11	0.06
L_{quant-}	5160	6885	7244	5160	7028	7469	5293	7198	7752	5393	7397	7970	5367	7527	8293
	0.34	0.15	0.13	0.34	0.13	0.10	0.33	0.12	0.08	0.32	0.10	0.06	0.33	0.10	0.05

Table 3.3: Bounds, deterministic solutions and stochastic solutions.

3.6.3 Demand variability

The proposed policies often provide solution values equal to bounds. In the same time, it is hard to determine the best stochastic approach, since more than one achieve similar performance. Thus a deeper analysis on the impact of demand variability on the optimal solution structure is required. More precisely, when making decisions for a current period, we need a better understanding on how optimal decisions depend on the demand variability in the next period.

To better understand the specificity of MPSSD, we make an analogy with the simplified stochastic routing problem studied in [Birge 2011]. In this problem, a provider has to determine a route or a set of routes to serve a set of clients with one capacitated vehicle. The demands of the clients are assumed to be known in advance except one which is uncertain. For this problem, it is possible to divide the possible values of the uncertain demand into consecutive disjoint sub-intervals and then define a specific optimal solution related to each one of them.

When solving MPSSD over two periods, we are faced to a different situation. The solution includes the quantity of orders to process during the first period and the quantity of orders to postpone to the second period. The exact values of these quantities depend on the postponement penalty and on the possibility to process postponed orders during the second period without additional resource costs, in other words the residual capacity of the second period. Indeed, if some postponed orders lead to additional resources in the second period, it is better to not postpone these orders and add resources in the first period.

In turn, the residual capacity of the following period depends on its demand. For different levels of demand, we observe a similar level of the residual capacity of the second period. leading to the same decisions related for the first period. As consequence, as the demand of the second period vary, a solution can be observed in different range of values. This behaviour is related to the fact that resource levels are variable. Different levels of demand for the second period lead to the same set of decisions during the first period.

Therefore, it is challenging to compute the different threshold value that determine the distribution of solutions with respect to the future demand. This distribution depends on various parameters: demand levels, penalties and resource costs and capacities.

3.6.4 Sensitivity analysis

This section is dedicated to the analysis of the sensitivity of the linearised cost which is used by both the PLRC and the stochastic techniques. Figure 3.5 illustrates the variation of the solution values obtained by respectively the PLRC (diamond) and the quantile scenario (triangle) as an example of a stochastic approach.

It is noteworthy that the performance of the PLRC is very bad when the value of the linearised resource cost is low (near zero). We also note that all values above

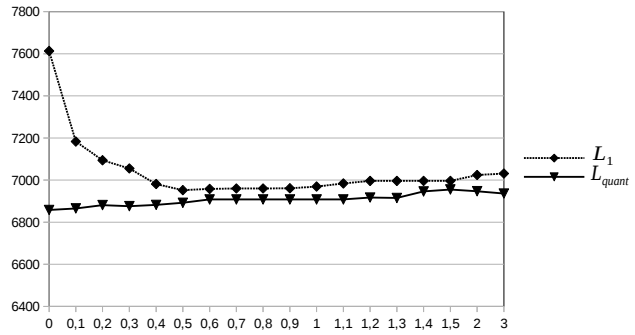

Figure 3.5: Results of sensitivity analysis

0.5 gives similar results. The stochastic technique seems not to be impacted by the linearised resource cost. This is explained by the fact that the MPSSD in this case is defined over 2 periods (and not only one like for the PLRC). This alleviates the impact of the linearised resource cost.

3.7 Conclusion

In this chapter we study a dynamic version of the integrated picking and shipping problem with stochastic demand (MPSSD). The MPSSD looks for a tactical resources design decisions under operational constraints. It is incorporated in a rolling horizon procedure to determine efficient solution methods over a set of periods. The structure of the rolling horizon is detailed and a mathematical formulation MPSSD is provided.

The main difficulty considered is the uncertainty of future demand. The rolling horizon framework enables us to use a number of *look-ahead* periods when deciding for a current period. Stochastic information can better guide decisions related to order postponement toward an efficient resources usage. Several stochastic approaches are proposed and compared.

We also investigated fully deterministic methods that can be implemented in the rolling horizon framework. Those methods are relevant when the decision maker have no reliable stochastic information.

Because of the complexity of the problem we provided bounds to evaluation the performances of the different methods. The results provide interesting insights on the dynamic management. The results highlight the benefit achieved by using the information of one look-ahead period compared to a myopic approach. It seems to be sufficient to limit the rolling horizon length to two periods. Including informations for more than one look-ahead period appears to be useless as shown by the gap between L_2 and L_3.

A first perspective is to investigate how to derive from the demand distribution

the distribution of the residual capacity, and how such distribution can be used in the design of an advanced stochastic approach. A second perspective is to conduct simulations on larger instances and investigate the limits of the rolling horizon procedure with respect to computation time.

Chapter 4

Last mile delivery services pricing with congestion

The works presented in this chapter were published in the revue Electronic Notes in Discrete Mathematics [Tounsi 2016a] and in the conference INOC 2015 [Tounsi 2015].

Contents

4.1 Introduction

The last segment of the supply chain is the delivery of packages from a local distribution center to the customer. The disproportionate expense of the last mile contributes to what's known as the "last-mile problem". This is because of the difficulty of reaching end users, especially in busy urban areas. This can translate to higher fuel costs due to the amount of time spent driving around making deliveries, since business-to-consumer deliveries often involve one package per stop, as compared to large volumes for business-to-business deliveries.

An efficient delivery system should offer various services and takes into account customers behaviour. In a highly competitive environment, customers are sensitive to the tariff of a delivery service, and also to its quality. The latter can be damaged by congestion experienced by customers when they are too many using one service.

In this chapter, we study a last mile delivery system, including the two most popular services: delivery at home and pick up at relay station. After making an order, each customer selects a delivery service by comparing the services utility functions. The utility includes using the service's tariff and a congestion measure. The latter, inspired from queuing theory, is specific to each service and it increases with the number of the service's users.

In economics and transport, discrete choice models describe and explain choices between two or more discrete alternatives. We use for present application the nested logit model to reflect better correlation between alternatives. At a given configuration (services tariffs and capacities), the system converges to a stable state, called an equilibrium, where no customer is willing to change unilaterally his/her choice.

Knowing the customers' behaviour, the system's manager is interested in determining the optimal system's parameters, like the tariffs, to optimize a given criteria which can be its revenue or the global customers' welfare. We address the system's manager problem with bi-level model, where at the upper level, he/she controls services' tariffs. At the lower level, users react by choosing their delivery service according to the utility function.

The chapter is organized as follows. In section 4.2 we introduce the last mile delivery problem and we describe the two main delivery services, home delivery and pick up. We provide in section 4.3 a state of the art on discrete choice models and bi-level problems. In section 4.4, services utility function are depicted along with the

customers' choice model. Section 4.5 is dedicated to the study and computation of the stochastic user equilibrium corresponding to the nested logit model. We present in section 4.6 a sensitivity analysis of the equilibrium that enables to compute the derivatives of equilibrium probabilities with respect to the services tariffs. Then, in section 4.7, we present the control problem for the leader which is a bi-level optimization problem formulated as a MPEC. We provide three solution techniques, and investigate the benefit of sensitivity analysis. We give experimental results in section 4.8. Finally section 4.9 concludes the chapter.

4.2 Last mile delivery

With the increased number of online sales, the "last-mile" deliveries become a crucial part of the supply chain. Enough unhappy consumers can have a very negative impact on online retailers and their parcel delivery company.

Bud Workmon, President of 3PD Inc, one of North America's largest and exclusive national providers of last-mile delivery and logistics services, highlights that 'Although last-mile logistics is only one small link in the supply chain, it's the only one that directly touches the customer - an important point to remember when considering your site options'.[1] More, in his article, Workmon claims that congestion, in terms of stem time, is one of the main concerns for the customers and should be for the last-mile delivery companies. It is natural to witness numerous and varied innovative delivery solutions looking to realize the trade-off between cost efficiency and customer requirement of timeliness and reliability. More and more retailers combine different delivery services and options offered to customers. We describe in the following the two main delivery services in e-commerce.

4.2.1 Delivery services

Delivery services can be classified into two main categories: Delivery at home (D) and Pick Up (P). In the first option, packages are delivered directly at home. This option is often convenient for the customer as there is no effort for picking the parcel. On the other side it has a non-negligible cost for the delivery company (vehicles, drivers, etc.), thus the customer generally has to pay for this service.

In addition to home delivery service, more and more on-line retailers (such as Amazon, Fnac, ..) offer customers to have their goods in a pick up location that can be a shop. This alternative way let customer choose a convenient time to collect their goods without having to be home for the delivery. This service is generally less charged by the delivery companies and even free of charge for the customer. The main drawback of this second option is the storage capacity of the pick up location, which is limited to small back office of standard shops, particularly in down-town cities and congested urban areas.

[1] LDW: Logistics, Distribution & Warehousing 2009, www.AreaDevelopment.com.

The latest trends in last mile delivery are parcel lockers. With parcel locker stations near shopping centers, couriers can place packages in lockers, and then provide customers with a unique key needed to access the locker.

4.2.2 Quality of service

After making the order and waiting for the delivery, the customer is finally ready to receive the goods. The customer is willing to receive the goods in a perfect state but also to be delivered as expected. If this experience turns bad, it will surely results in considerable disappointment. And the reason for a such sad ending can come from the delivery.

Home delivery can be inconvenient for the customer when loosing time waiting for the deliver-man. While letting the customer define delivery appointment would highly complicate the rationalization of deliveries tours for the companies. Dynamic Vehicle Routing Problem in which new orders arrive during operation [Pillac 2013] and Period Vehicle Routing Problem with Service Choice [Braekers 2016] are relevant for home delivery operations to determine both routes and service frequencies. But such planning approaches suffer from the repercussions of congestion due to unforeseen activity peaks. In this case additional delays result in missing delivery appointment. Defining the time windows menu is challenging for the success of home delivery. Moreover resources' use can be optimized through yield management, like for example, in [Asdemir 2008], where dynamic pricing influences customer's choice of a delivery window.

It may appear that the geographical location of the pick up is a main criteria for the customers. In reality, it is not that trivial since the customer plan usually the visit of the pick up location in a regular trip like the home-work trip or during shopping. Thus we do not include relay stations location in the attributes of customers choice. On the other hand, the congestion of the relay station implies an undesirable delay for the user because it impacts the rest of his (her) activities. The time spent at the location has thus to be as limited as possible.

We then notice for both types of service the importance of measuring and managing the congestion. The congestion measure depends on the type of service. In fact, the more customers choose the (D) service, the more trips have to be performed. As a consequence, the delivery time increases for a part of customers. While the number of customers that choose a (P) option, determines the probability of saturating its storage capacity.

4.3 State-of-the-art: methodologies

The following state of the art focuses on the two main features of the last mile delivery system design: discrete choice models and bi-level programming.

4.3.1 Discrete choice models

Discrete choice models (DCM) describe and predict choices between two or more discrete alternatives, such as entering or not entering the labor market, or choosing between modes of transport.

If we assume that customers have full information of each service's utility function and that they rationally make their decision, we obtain the deterministic user equilibrium, also called the Wardropian equilibrium. This assumption is too strong in our application. It is more accurate to consider that they can make errors in their choices for many reasons (lack of information, individual preference,...). Precisely it is assumed that the utility includes an error ε that can not be observed. In the multinomial Logit model (MNL), the error is modelled as a random variable that follows Gumbel distribution. The MNL, usually considered in traffic assignment problems, has many interesting properties like the *efficiency principle* [Erlander 1975], that assumes that the random terms ε are independent and identically distributed (IID) Gumbel variables. That leads to a closed-form expression of choice probabilities defining a stochastic user equilibrium (SUE).

But MNL does not take into account the correlations that may exist among different options (overlapping effects between routes in traffic context for example) and thus can give unrealistic choice probabilities. By using normally distributed random terms, the multinomial Probit model (MNP) does not have this drawback, because it considers the covariance between the random terms [Sheffi 1985]. However, MNP does not give a closed-form expression of choice probabilities, instead it requires computationally demanding Monte Carlo simulations, and thus can not be applied on large problems.

One family of extension overcomes the overlapping problem by modifying the utility function. The C-logit model for example [Zhou 2010] adds a communality factor. Nested Logit models (NLM) [Wen 2001] are based on a two-level structure and then can be used to model the possible correlation between options. The nested Logit models fits well the choice process considered in the delivery, we detail this model in section 4.4.

4.3.2 Bi-level programming

Bi-level models suit a wide variety of situations where a first actor, called the leader, integrates in his/her decision process the decision of a second part, called the follower. More precisely the leader solves an optimization problem that includes another optimization problem representing the decision of the follower. We talk also about the leader taking into account the response of the follower. It is the case for example when a private company decides the quantity of product to produce knowing that a competitor will react by adjusting its own production.

Bi-level problems (BLP) gathered considerable interest in economy and game theory and more recently in optimization and operations research (see [Colson 2005]

for a survey). Numerous situations can be modelled by BLP, we find applications in transportation network tolling [Brotcorne 2000], energy pricing [Brotcorne 2008] and telecommunication [Bouhtou 2007]. BLP are known to be challenging even in their simplest form where all objective functions and constraints are linear. Indeed linear BLP is $\mathcal{NP}$-hard, and even strongly $\mathcal{NP}$-hard [Brotcorne 2008].

In a Bi-level problem the leader knows how the follower makes a decision but he/she can not directly intervene in the latter's decision. Thus the leader considers the follower's reaction in his/her own decision problem. In some situations where followers impact each other, the leader can hardly predict their reaction. Consider the problem of designing an urban road network by deciding its link capacities, the aim being to reduce the global travel delay. Due to congestion effects, perverse effects can appear, like the Braess paradox, where increasing the capacity of a link (or building a new link) may result in a delay increase for every user of the network. In this case the reaction of follower is the result of an equilibrium based on the leader decision and the overall congestion. The leader decision problem is naturally formulated by a mathematical program with equilibrium constraint (MPEC) [Luo 1996].

Pieper [Pieper 2001] highlighted the intrinsic difficulty of MPEC by showing that the usual constraint qualification like Mangasarian-Fromovitz constraint qualification or linear independence constraint qualification does not hold at any feasible point. These assumptions, aside some others, are critical for the convergence of many algorithms used for standard nonlinear problems. The author applied piecewise sequential quadratic programming, penalty interior point algorithm and sequential quadratic programming. These methods are based on the formulation of the lower problem as a non linear complementarity constraint. Another approach uses smooth function [Facchinei 1999] and solves iteratively local approximation that gives a search direction for the upper level problem.

Considering a Wardropian (deterministic) equilibrium, Wang and Lo [Wang 2010] reduce the bi-level model to a single level one. They introduce a linearization scheme for the equilibrium constraint along with linear approximation of utility functions. Then, the resulting mixed integer linear problem can be solved using commercial software for an approximate global solution. Recently, Liu and Wang [Liu 2015] have proposed a global solution algorithm for a network design problem with stochastic user equilibrium. The algorithm is based on a linear relaxation of the initial non-convex problem.

At the heuristic front, sensitivity analysis (SA) has been shown to be a very useful tool to build efficient methods. (SA) consists generally in computing the derivatives of lower-level variables with respect to the upper-level variables. The derivative can then be incorporated in descent methods like in [Friesz 1990] where the gradient of upper-level objective function is computed using (SA), and used to update upper-level variables towards a local optimum. Based on the sensitivity analysis, Yang et al. in [Yang 1994a, Yang 1994b] build a linear approximation of the bilevel problem

that provides a descent direction for upper-level variables. In [Ying 2003] (SA) is implemented for Logit based (SUE) and used to find optimal road tolls and transit tax. Another approach based on (SA) can be find in [Meng 2001]. The major difficulty of these approaches lies in the fact that there is no guarantee that the solution obtained is a global optimum due to the inherent non-convexity of MPEC. We note the use of heuristics and metaheuristics with upper-level decisions being integer variables (decisions are for example lane layout, link/lane allocation, signal stages, etc) such as hill climbing, simulated annealing, tabu search [Cantarella 2005], scatter-search technique [Gallo 2010], genetic algorithm [Ceylan 2003]. The general principle consists in alternating upper level problem resolution (with lower level variables fixed) and equilibrium computing (for fixed upper level variables) until convergence.

4.4 Delivery services choice model

The delivery system is composed of a set $\mathcal{N}$of N services. All along this chapter, we consider the two classical service types: delivery at home (D) and pick up at relay station (P). Note that, the model addressed in this chapter and obtained results can directly be applied to a more general setting. In order to fit customers requirements, each service admits several options. For delivery at home service (D), options are distinguished by the time window of the delivery. Options of the pick up service (P) vary in the location areas: dense city center (usually with small capacity storage) or outside city site (higher storage capacity). For each service $n \in \mathcal{N}$ the set of options is denoted by $\mathcal{O}_n$. The delivery system is depicted on Figure 4.1. In a study of a similar delivery services including one (D) service and one (P) service [Hayel 2016], customers are interacting creating congestion which is measured using queuing theory. We use here the same service models.

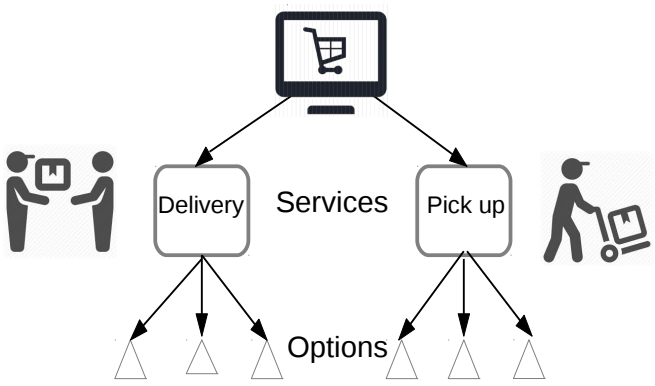

Figure 4.1: Last-mile delivery services system

We denote by λ the customers arrival rate per unit of time. Particularly, this

process is assumed to follow a Poisson process and the expected number of customers is equal to λ. Each customer selects a service n and an option j among options $\mathcal{O}_n$. All customers are identical and we denote by p_{nj} the probability that a new customer chooses option j of service n. This probability is equivalent to the fraction of customers choosing option $j \in \mathcal{O}_n$.

4.4.1 Utility functions

The customer decision is based on comparing option utility functions. The utility function of an option j of service n depends mainly on two attributes: the tariff t_{nj} of the option set by the provider, and a measure of the quality of service that reflects the satisfaction level of the customer. The measure of quality of service depends on the nature of the service, but it is always related to the congestion effect induced by customers decisions. These metrics are obtained by considering general queueing models. The congestion function of option j of service n is denoted by the function $f_n(p_{nj})$. Thus the general form of the utility c_{nj} of option $j \in \mathcal{O}_n$ is given by the following expression:

$$\forall n \in \mathcal{N}, \forall j \in \mathcal{O}_n, \quad c_{nj}(p_{nj}) = t_{nj} + \beta_n f_n(p_{nj}), \tag{4.1}$$

where β_n is a monetary conversion coefficient that is calibrated regarding the type of service (D) or (P).

4.4.1.1 Home delivery congestion function

A home delivery service is modeled as a M/D/1 queue where D is the time required by the transportation company to deliver one parcel. In other words, the transportation company can treat a maximum number of 1/D parcels per unit of time. We consider D is constant as it is related to the delivery capacity of the vehicles used by the transportation company. We could consider that this time depends also of some exogenous random conditions (traffic density, drivers, etc.) and then we should consider an M/G/1 queue. In order to keep the analysis simple and as clear as possible, we decide to keep the M/D/1 model with a First-In-First-Out discipline.

If a customer chooses the (D) services, the congestion is perceived in term of the average delivery delay. The arrival of demand for this service follows a Poisson process like the arrival of customers into a Markov queue. Customers are served one by one, and the service time of each customer is represented as a positive random variable S_{nj}. The average sojourn time (which corresponds to the average delivery delay) is given by the Pollaczek-Khinchin formula [Takács 1962] as the following non-linear function:

$$\forall n \in \mathcal{N}, \forall j \in \mathcal{O}_n, \quad f_n(p_{nj}) = m_1 + \frac{\lambda p_{nj} m_2}{2(1 - \lambda p_{nj} m_1)} \ .$$

with $m_1 = I\!E(S_{nj})$ and $m_2 = I\!E(S_{nj}^2)$. Both are decreasing functions of K_{nj}, the delivery capacity of option j of service n. This value may represent for option j in

service n the number of parcels that can be delivered per unit of time. A simple assumption could be that S_{nj} follows an exponential distribution with parameter K_{nj}, and then $m_1 = 1/K_{nj}$ and $m_2 = 2/K_{nj}^2$.

4.4.1.2 Relay station congestion function

A Pick up service is modeled as a M/M/K/K queue where K is the capacity of the relay, i.e. the number of packets that can be stored, waiting to be picked up by costumer. We assume that each packet occupies one storage unit (before being picked up) during a random duration which follows an exponential distribution with parameter μ. All these durations are independent and identically distributed. We do not consider the time it takes to deliver the packet to the relay station. We consider that the most important for a customer when choosing this option, is to be delivered in the chosen relay (which is usually close to his/her house or his/her office).

When using a service of type (P), parcels may be rejected if the relay station is full. Then the customer pays extra cost related to the dispatch the parcel to another relay station. I

The average congestion function for type (P) service is expressed by the blocking probability of a M/M/K_{nj}/K_{nj} queue where K_{nj} is the capacity of the relay station j. Then, the Erlang-B formula gives this blocking probability depending on the parameters of the system as:

$$\forall n \in \mathcal{N}, \forall j \in \mathcal{O}_n, \quad f_n(p_{nj}) = \frac{(\frac{\lambda p_{nj}}{\mu})^{K_{nj}}}{K_{nj}! \sum_{k=0}^{K_{nj}} (\frac{\lambda p_{nj}}{\mu})^k / k!} \ .$$

Each customer decides selfishly his/her best choice regarding a cost function that includes service's tariff and the congestion measure. While the tariff is known from customer when making the order, the congestion state can only be perceived during the delivery. We assume that the congestion state can be communicated to customer at the moment of the order. This information can be based on the current state of services and previsions.

4.4.2 Logit model

The logit model is classically used in discrete choices. It allows choices to not be full rational by including in utility function a random error term that captures the unobservable customers preferences. Utility functions have then the following form

$$\forall n \in \mathcal{N}, \forall j \in \mathcal{O}_n, \quad u_{nj}(p_{nj}) = c_{nj}(p_{n,j}) + \gamma_{n,j} \ . \tag{4.2}$$

The logit model assumes that all error terms $\gamma_{n,j}$ have a logistic distribution (Gumbel) and they are independent. It is then possible to obtain a closed form of the

probability p_{nj} of a choice to be chosen [Stevanovic 2006], as:

$$\forall n \in \mathcal{N}, \forall j \in \mathcal{O}_n, \quad p_{nj} = \frac{e^{-\theta c_{nj}}}{\sum_{m \in \mathcal{N}} \sum_{k \in \mathcal{O}_m} e^{-\theta c_{mk}}}. \tag{4.3}$$

The parameter θ determines the level of rationality of the decisions of the customers. When θ tends to infinity, we obtain the full rational case and the SUE collapses on the Wardrop equilibrium. As θ tends to 0, we have a complete irrational situation where options have the same probability to be chosen whatever are their costs (uniform distribution over the options).

The logit model has several good properties. We note that when one option of an alternative increases letting the others constant, the corresponding probability also increases while other probabilities decrease. Another feature of the logit model is that all probabilities are non negative. Thus if an alternative is supposed to not have any chance to be selected, it must be removed from the set.

4.4.3 Disadvantages of Logit

A main consequence of the logit model is the proportional substitution between alternatives. When the utility of an alternative varies, say it decreases, all other alternatives probabilities increase with the same proportion of their previous value. Indeed the ratio between two alternatives remains the same when another alternative is added, removed, or changed (in the sense its utility function is changed). The ratio between two alternatives only depends on their two utility functions. This property is called the independence from irrelevant alternatives (IIA).

The IIA is however wrong in some choice situations. This is illustrated by the blue bus/ red bus paradox. In this example, users chose their travel mode between two alternatives, car or bus. First, it is supposed that utility functions are equal and thus choice probabilities are: $P(\text{car}) = P(\text{bus}) = 1/2$.

Now, suppose that a second bus is added to alternatives set, and that it has the same service than the existing bus except that it has a different color (the first bus is red and the second is blue). Since utility function are equal (the bus color does not impact the choice), the new choice probabilities are : $P(\text{car}) = P(\text{red bus}) = P(\text{blue bus}) = 1/3$. And the IIA is verified.

This result is however not realistic because the two bus services are considered by users as one since only the color is different. The choice probabilities should by as follows. $P(\text{car}) = 1/2$ and $P(\text{red bus}) = P(\text{blue bus}) = 1/4$.

Thus for the IIA to be valid, the choice set has to contain strictly distinct alternatives. The blue bus/ red bus paradox shows that when some alternatives have similarities, the logit model fails to reflect truly the choice process. We present in the next section the nested logit model which is based on a hierarchical structure of alternatives that is present in our delivery system.

4.4.4 Nested Logit

The nested logit is a particular case of *Generalized Extreme Value* (GEV) models which take into account different types of correlation between alternatives. The nested logit is one of the most used and it is based on a two levels choice process. It particularly suits situations where alternatives can be partitioned in groups, called nests.

- Inside each nest, the IIA property is verified. The ratio between two alternatives probabilities does not depend on the existence or the attributes of others alternatives.
- For two alternatives from different nests, the ratio of probabilities depends on the others alternatives in the two nests, and IIA does no more hold.

In our setting, a nest represents a service $n \in \mathcal{N}$. The nested model we consider has non overlapping nests, and it is a particular case of the cross nested logit model proposed in [Bekhor 2003]. The choice process can be seen as divided in two steps. First, each customer determines the nest (D) or (P); second, the customer decides which option of the nest is chosen. To alleviate complicated notations, we refer to utility function by c_{nj} instead of $c_{nj}(p_{nj})$.

The GEV distribution of error terms gives the probability of option j in service n to be chosen by a customer as the product of the marginal probability $Pr(n)$ of the service n being chosen, and the conditional probability $Pr(j|n)$ of option j being chosen given that service n is chosen [Stevanovic 2006]:

$$\forall n \in \mathcal{N}, \forall j \in \mathcal{O}_n, \quad p_{nj} \;=\; Pr(n)Pr(j|n) \tag{4.4}$$

with

$$Pr(n) = \frac{\big(\sum\limits_{k\in\mathcal{O}_n} e^{-\theta c_{nk}/\phi_n}\big)^{\phi_n}}{\sum\limits_{m\in\mathcal{N}}\big(\sum\limits_{k\in\mathcal{O}_m} e^{-\theta c_{mk}/\phi_m}\big)^{\phi_m}} \qquad \text{and} \qquad Pr(j|n) = \frac{e^{-\theta c_{nj}/\phi_n}}{\sum\limits_{k\in\mathcal{O}_n} e^{-\theta c_{nk}/\phi_n}}.$$

The parameter ϕ_n, for each service $n \in \mathcal{N}$, is the nesting or correlation degree between options in the nest n. It is ranged between 0 and 1. The MNL model corresponds to the case where $\phi_n = 1$ for all $n \in \mathcal{N}$, while options are more and more correlated as $\phi_n \to 0$.

Considering two options i and j in two nests respectively n and m, the ratio of probabilities is :

$$\frac{p_{ni}}{p_{mj}} = \frac{e^{-\theta c_{ni}/\phi_n}\big(\sum\limits_{k\in\mathcal{O}_n} e^{-\theta c_{nk}/\phi_n}\big)^{\phi_n-1}}{e^{-\theta c_{mj}/\phi_m}\big(\sum\limits_{k\in\mathcal{O}_m} e^{-\theta c_{mk}/\phi_m}\big)^{\phi_m-1}}.$$

When the two options are in the same nest ($n = m$) the two sum terms are removed and the ratio is independent of all others options. So if one option's utility

varies, other option in the same nest varies according to the proportional substitution. This behaviour is realistic because individuals who would change their choice but remain in the same nest, will only consider the attractiveness of the nest's options.

The behaviour is different if we analyse the relation between two options of different nests ($n \neq m$). The ratio depends on the options present in the two nests. We see that the probabilities ratio depends on all the options of the two considered nests. When options i varies, say it decreases, the attractiveness of nest n decreases, while the attractiveness of other nests, including m, increases. This makes indirectly the probability of option j increase.

4.5 Stochastic user equilibrium

In order to determine the SUE with the nested logit model presented in section 4.4, we have to solve a system of non-linear equations given by $p_{nj} = Pr(n)Pr(j|n)$ for all $n \in \mathcal{N}, j \in \mathcal{O}_n$. Such system is difficult to solve in closed-form, mainly because of the complex congestion functions based on queueing systems metrics. We give in the following a formulation of the SUE as a convex optimization problem. Then we describe an efficient method to compute the SUE.

4.5.1 Equivalent optimization problem

We give in this section an optimization formulation for the nested SUE. The utility functions c_{nj}, $\forall n \in \mathcal{N}, \forall j \in \mathcal{O}_n$ are increasing functions and separable, meaning that c_{nj} depends only on option's probability p_{nj}. Our formulation is an adaptation of the general formulation of Sheffi [Sheffi 1985]. The mathematical formulation enables the computation of the nested SUE. Furthermore it is used to develop a sensitivity analysis for the SUE in section 4.6.

Proposition 4 *Let $\boldsymbol{p}^*$ a solution of the following minimization problem:*

$$
\begin{aligned}
\textit{[N-SUE]} \quad \min_{\boldsymbol{p}} \quad & Z(\boldsymbol{p}) = Z_1(\boldsymbol{p}) + Z_2(\boldsymbol{p}) + Z_3(\boldsymbol{p}) && \text{(4.5a)}\\
s.\ t. \quad & Z_1(\boldsymbol{p}) = \sum_{n\in\mathcal{N}} \sum_{j\in\mathcal{O}_n} \int_0^{p_{nj}} c_{nj}(s)ds && \text{(4.5b)}\\
& Z_2(\boldsymbol{p}) = \sum_{n\in\mathcal{N}} \frac{\phi_n}{\theta} \sum_{j\in\mathcal{O}_n} p_{nj} \ln(p_{nj}) && \text{(4.5c)}\\
& Z_3(\boldsymbol{p}) = \sum_{n\in\mathcal{N}} \frac{1-\phi_n}{\theta} ((\sum_{j\in\mathcal{O}_n} p_{nj}) \ln(\sum_{j\in\mathcal{O}_n} p_{nj})) && \text{(4.5d)}\\
& \sum_{n\in\mathcal{N}} \sum_{j\in\mathcal{O}_n} p_{nj} = 1 && \text{(4.5e)}\\
& p_{nj} \geq 0 \quad \forall n \in \mathcal{N}, \forall j \in \mathcal{O}_n && \text{(4.5f)}
\end{aligned}
$$

Then $\boldsymbol{p}^*$ is a solution of the non-linear system (4.4) and there exists a SUE considering a Nested Logit DCM with congestion cost functions.

Proof The term Z_1 comes from the formulation of the Deterministic equilibrium [Sheffi 1985]. Z_2 is an entropy term that is related to the Logit model. It is modified here compared to Fisk's formulation [Fisk 1980] to include the correlation coefficients. Finally we introduce a second entropy term Z_3 that corresponds to the nested choice structure.

We demonstrate that the optimality conditions of the proposed mathematical formulation correspond to the nested SUE (4.4).

Let us consider the following Lagrangian function :

$$L(\mathbf{p},\nu) = Z_1(\mathbf{p}) + Z_2(\mathbf{p}) + Z_3(\mathbf{p}) + \nu(1 - \sum_{n\in\mathcal{N}}\sum_{j\in\mathcal{O}_n} p_{nj}),$$

where ν is the Lagrange coefficient associated to the constraint (4.5e). The first-order conditions of the Lagrangian function are obtained by looking at the partial derivatives of L with respect to decision variable p_{nj}:

$$\forall n \in \mathcal{N}, \forall j \in \mathcal{O}_n, \quad \frac{\partial L(\mathbf{p},\nu)}{\partial p_{nj}} = c_{nj}(p_{nj}) + \frac{\phi_n}{\theta}(\ln(p_{nj})+1) + \frac{1-\phi_n}{\theta}(\ln(\sum_{k\in\mathcal{O}_n} p_{nk})+1) - \nu = 0,$$

and also

$$\frac{\partial L(\mathbf{p},\nu)}{\partial \nu} = 0.$$

To simplify mathematical notations, $c_{nj}(p_{nj})$ is written as c_{nj} for the rest of the proof. After some manipulations of the equations we get:

$$\forall n \in \mathcal{N}, \forall j \in \mathcal{O}_n, \quad p_{nj}(\sum_{k\in\mathcal{O}_n} p_{nk})^{\frac{1-\phi_n}{\phi_n}} = e^{(\nu\theta-1)/\phi_n} e^{-\theta c_{nj}/\phi_n}. \tag{4.6}$$

Summing equation (4.6) over all options $j \in \mathcal{O}_n$ for the nest $n \in \mathcal{N}$, we obtain:

$$\forall n \in \mathcal{N}, \quad (\sum_{k\in\mathcal{O}_n} p_{nk})^{\frac{1}{\phi_n}} = e^{(\nu\theta-1)/\phi_n} \sum_{j\in\mathcal{O}_n} e^{-\theta c_{nj}/\phi_n}. \tag{4.7}$$

Elevating both sides to ϕ_n, we get:

$$\forall n \in \mathcal{N}, \quad \sum_{k\in\mathcal{O}_n} p_{nk} = e^{\nu\theta-1}(\sum_{j\in\mathcal{O}_n} e^{-\theta c_{nj}/\phi_n})^{\phi_n}. \tag{4.8}$$

Then we can derive the probability $Pr(n)$ for a nest n to be chosen by:

$$\forall n \in \mathcal{N}, \quad Pr(n) = \frac{\sum_{k\in\mathcal{O}_n} p_{nk}}{\sum_{m\in\mathcal{N}}\sum_{k\in\mathcal{O}_m} p_{mk}} = \frac{(\sum_{k\in\mathcal{O}_n} e^{-\theta c_{nk}/\phi_n})^{\phi_n}}{\sum_{m\in\mathcal{N}}(\sum_{k\in\mathcal{O}_m} e^{-\theta c_{mk}/\phi_m})^{\phi_m}}. \tag{4.9}$$

The probability for an option $j \in \mathcal{O}_n$ to be chosen, given that nest n was chosen, is obtained by dividing equation (4.6) by equation (4.7).

$$\forall n \in \mathcal{N}, \forall j \in \mathcal{O}_n, \quad Pr(j|n) \;=\; \frac{p_{nj}}{\sum\limits_{k\in\mathcal{O}_n} p_{nk}} = \frac{e^{-\theta c_{nj}/\phi_n}}{\sum\limits_{k\in\mathcal{O}_n} e^{-\theta c_{nk}/\phi_n}}. \tag{4.10}$$

Equations (4.9) and (4.10) correspond to the nested Logit equilibrium given by the system 4.4. In this sense, the mathematical formulation (4.5) corresponds to a nested Logit SUE formulation. ■

Proposition 5 *The strategic decision problem considering the Nested Logit DCM admits a unique SUE solution.*

Proof Since for all $n \in \mathcal{N}, j \in \mathcal{O}_n$, the cost functions $c_{nj}(p_{nj})$ are positive and increasing in p_{nj}, the function $Z_1(\mathbf{p})$ is convex. Note also that the feasible region is convex. Moreover, twice differentiating the function $Z_2(\mathbf{p})$ and $Z_3(\mathbf{p})$ gives the following results:

$$\frac{\partial^2 Z_2(\mathbf{p})}{\partial p_{nj}\partial p_{mk}} = \begin{cases} \frac{\phi_n}{\theta p_{nj}}\ , & \text{if n=m and j=k,} \\ 0\ , & \text{otherwise,} \end{cases} \tag{4.11}$$

$$\frac{\partial^2 Z_3(\mathbf{p})}{\partial p_{nj}\partial p_{mk}} = \begin{cases} \frac{1-\phi_n}{\theta}\frac{1}{\sum\limits_{l\in\mathcal{O}_n} p_{nl}}\ , & \text{if n=m,} \\ 0\ , & \text{otherwise.} \end{cases} \tag{4.12}$$

First, note that in DCM models, each alternative has a strictly positive probability, i.e. for all $n \in \mathcal{N}, \forall j \in \mathcal{O}_n\ p_{nj} > 0$. Second, as θ is finite, the Hessian matrix of $Z_2(\mathbf{p})$ is diagonal with positive elements. Thus the function $Z_2(\mathbf{p})$ is strictly convex. The Hessian matrix of function $Z_3(\mathbf{p})$ is bloc diagonal. It is positive semidefinite (all elements in the matrix are equal to one another, and the determinant is equal to zero). All arguments together ensure strict convexity of the whole objective function $Z(\mathbf{p})$, and hence the solution is unique in terms of choice probabilities p_{nj}. ■

4.5.2 SUE computation

A classical method for equilibrium solution is the method of successive averages (MSA). This method has been proved to be efficient for solving the Logit-based SUE [Sheffi 1985, Damberg 1995, Maher 1998], and for the nested SUE [Bekhor 2003]. It is an iterative descent method that involves at each iteration i descent direction $\mathbf{d}^i$ with respect to the objective function and a step size τ^i. A sequence of points $\mathbf{p}^i = (p^i_{nj}), \forall n \in \mathcal{N}, \forall j \in \mathcal{O}_n$ with $i = 1, 2, \ldots$, that converges to the solution of (4.5) is constructed following the following recursive pattern:

$$\forall i = 1, 2, \ldots, \quad \mathbf{p}^{i+1} = \mathbf{p}^i + \tau^i \mathbf{d}^i.$$

This process converges to a local minimizer which is global in the case of convex problem. At a given iteration, given a current choice probabilities $\mathbf{p}^i = (p^i_{nj})$, utility functions are fixed for all $n \in \mathcal{N}, \forall j \in \mathcal{O}_n$ at the value $C^i_{nj} := c_{nj}(p^i_{nj})$. An auxiliary choice probabilities vector $\mathbf{y}^i = (y^i_{nj})$ is then computed by applying the formulae of the nested SUE(4.4). A descent direction $\mathbf{d}^i$ is then given by $\mathbf{d}^i := \mathbf{y}^i - \mathbf{p}^i$.

For the Nested Logit DCM, given current costs C^i_{nj}, the auxiliary point $\mathbf{y}^i = (y^i_{nj})$ is computed as follows:

$$\forall n \in \mathcal{N}, \forall j \in \mathcal{O}_n, \forall i = 1, 2, \dots \quad y^i_{nj} = \frac{(\sum\limits_{k \in \mathcal{O}_n} e^{-\theta C^i_{nk}/\phi_n})^{\phi_n}}{\sum\limits_{m \in \mathcal{N}} (\sum\limits_{k \in \mathcal{O}_n} e^{-\theta C^i_{nk}/\phi_m})^{\phi_m}} \frac{e^{-\theta C^i_{nj}/\phi_n}}{\sum\limits_{k \in \mathcal{O}_n} e^{-\theta C^i_{nk}/\phi_n}} \quad (4.13)$$

The standard MSA suggests a step size $\tau_i = 1/i$ that is shown to be efficient [Sheffi 1985]. The convergence of MSA is reached when new probabilities vector is ε closed to the previous one, where $\varepsilon > 0$ is a given threshold. The detailed steps of (MSA) are as follows.

- **Step 0:** *Initialization,* $i = 0$. Find initial choice probabilities $\mathbf{p}^0$ and compute initial costs $C^0_{nj} = c_{nj}(p^0_{nj})$, $\forall n \in \mathcal{N}, j \in \mathcal{S}_n$.
- **Step 1:** *Direction finding.* Apply equation (4.13) with fixed costs C^i_{nj} in order to get the auxiliary points $\mathbf{y}^i$.
- **Step 2:** *Move.* Find the new solution $\mathbf{p}^{i+1} = (p^{i+1}_{nj})$ by:

$$\forall n \in \mathcal{N}, \forall j \in \mathcal{O}_n, \quad p^{i+1}_{nj} = p^i_{nj} + \frac{y^i_{nj} - p^i_{nj}}{i+1}.$$

- **Step 3:** *Convergence criterion.* Compute the infinite norm difference as $||\mathbf{p}^{i+1} - \mathbf{p}^i||_\infty := \max\limits_{n \in \mathcal{N}, j \in \mathcal{O}_n} |p^{i+1}_{nj} - p^i_{nj}|$. If $||\mathbf{p}^{i+1} - \mathbf{p}^i||_\infty < \varepsilon$ then stop, else set $i = i + 1$ and go to step 1.

The costs at the initialization step are obtained considering the free-flow costs, i.e. for all $n \in \mathcal{N}, \forall j \in \mathcal{O}_n$, $C^0_{nj} = c_{nj}(0)$. Note that MSA can be improved by performing a line-search to have optimal step size at step 2 [Chen 1991]. As our problem is similar to a traffic assignment problem with a particular parallel links topology, we use the standard MSA that is shown to be efficient for the examples we consider in section 4.8.

4.6 Sensitivity analysis

The SUE that rules customer behaviour depends on the system parameters and characteristics such as options tariff, pick-up relay capacities, delivery capacity, etc... It is then interesting to understand how the resulting SUE evolves when one or more of those parameters change.

The absence of closed expression of the dependence of SUE on the control parameter motivates the use of sensitivity analysis like in [Yang 1994a]. "Sensitivity analysis identifies how the variability in an output quantity of interest is connected to an input in the model"[2].

It is generally based on sensitivity derivatives i.e., the gradient of the output of interest with respect to input variables. We develop in this subsection a sensitivity analysis of the nested SUE with respect to the provider's control parameters. Let I by the number of set of provider's control parameters indexed in the set $\mathcal{I} = 1, .., I$, $\mathbf{u} = (u_i), \forall i \in \mathcal{I}$.

4.6.1 General result

We first remind basic results of sensitivity analysis for nonlinear optimization problem [Fiacco 1983]. Consider a general parametric nonlinear programming problem defined for a given $\varepsilon > 0$ as follows:

$$Q(\varepsilon) = \min_x f(x, \varepsilon),$$

subject to

$$g_i(x, \varepsilon) \leq 0, \quad i = 1, ..m,$$

$$h_i(x, \varepsilon) = 0, \quad i = 1, ..n.$$

The Lagrangian function associated with $P(\varepsilon)$ is defined by:

$$L(x, \nu, \mu, \varepsilon) = f(x, \varepsilon) + \sum_{i=1}^{m} \nu_i g_i(x, \varepsilon) + \sum_{j=1}^{n} \mu_j h_j(x, \varepsilon)$$

where $\nu \in \mathcal{R}^m$, $\mu \in \mathcal{R}^n$. Suppose that the second order sufficient conditions for a strictly local minimum of $P(\varepsilon)$ hold at x with the associated Lagrange multiplier vectors ν and μ. Then we have the following theorem.

Theorem 1 *[Fiacco 1983] If the following conditions are verified:*

- *the functions defining problem $Q(\varepsilon)$ are twice continuously differentiable neighborhood of a couple $(x^\star, \varepsilon^\star)$,*
- *the second order sufficient conditions for a local minimum of $Q(\varepsilon^\star)$ holds at $x^\star$, with associated Lagrange multipliers $\nu^\star$ and $\mu^\star$,*
- *the gradients, $\bigtriangledown_x g_i(x^\star, \varepsilon^\star)$ for i such that $g_i(x^\star, \varepsilon^\star) = 0$, and $\bigtriangledown_x h_j(x^\star, \varepsilon^\star)$ for $j = 1, ..., n$ are linearly independent,*
- *the strict complementary slackness condition: $\nu_i > 0$ when $g_i(x^\star, \varepsilon^\star) = 0$, is satisfied,*

[2] Dimitri Papadimitriou, Plenary talk "Network modeling and design for unpredictability and uncertainty" in the 7th international network optimization conference, 20/05/2015.

then we have:

1. *$x^\star$ is a strict local minimum of $Q(\varepsilon^\star)$ and the associated Lagrange multiplier vectors $\nu^\star$ and $\mu^\star$ are unique,*

2. *For ε in a neighborhood of $\varepsilon^\star$, there exists a unique once continuously differentiable function $[x(\varepsilon), \nu(\varepsilon), \mu(\varepsilon)]$ satisfying the second-order sufficient conditions for a local minimum of $P(\varepsilon)$ such that*

$$[x(\varepsilon^\star), \nu(\varepsilon^\star), \mu(\varepsilon^\star)] = [x^\star, \nu^\star, \mu^\star]$$

and hence, $x(\varepsilon)$ is a locally unique solution to $Q(\varepsilon)$ and $\nu(\varepsilon)$ and $\mu(\varepsilon)$ are the unique associated multipliers.

3. *For $\varepsilon = \varepsilon^\star$ and $(x, \nu, \mu) = (x^\star, \nu^\star, \mu^\star)$, the first order Kuhn-Tucker conditions are:*

$$\begin{aligned} \nabla_x L(x, \nu, \mu, \varepsilon) &= 0 \\ \nu_i g_i(x, \varepsilon) &= 0 \quad i = 1, ..., m \\ h_i(x, \varepsilon) &= 0 \quad j = 1, ..., n \end{aligned} \tag{4.14}$$

Let the Jacobian matrix of the system of equations (4.14) with respect to $y = (x, \nu, \mu)$ be denoted as $M(\varepsilon)$ and with respect to ε as $N(\varepsilon)$. Then the matrix $M(\varepsilon)$ is non-singular and the partial derivatives of $[x, \nu, \mu]$ with respect to ε are given by:

$$\nabla_\varepsilon y(\varepsilon) = -M^{-1}(\varepsilon) N(\varepsilon) \tag{4.15}$$

This theorem is at the core of the sensitivity analysis in nonlinear optimization problems. We apply this theorem in our setting in order to make a sensitivity analysis of the nested Logit SUE. The perturbation occurs in provider's tariffs $\mathbf{t} = (t_{nj})$. Thus, option utility functions are of the form $c_{nj}(t_{nj}, p_{nj})$, for all $n \in \mathcal{N}, \forall j \in \mathcal{O}_n$. As required for the theorem application, the utilities functions are once continuously differentiable in $\mathbf{t}$ and $\mathbf{p}$.

4.6.2 The case of nested Logit SUE

In this section, based in the theorem 1 we develop a sensitivity analysis of the nested SUE, in order to compute the derivatives of choice probabilities at equilibrium $\bar{\mathbf{p}}$ with respect to the provider's tariffs $\mathbf{t}$. The parameterized [N-SUE] problem is

$$\begin{aligned} [\text{N-SUE}(\mathbf{t})] \quad \min_{\mathbf{p}} \quad & Z(\mathbf{t}, \mathbf{p}) \\ s.t \quad & \sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{O}_n} p_{nj} = 1 \\ & p_{nj} \geq 0 \qquad \forall j \in \mathcal{O}_n, n \in \mathcal{N} \end{aligned}$$

We assume that for a given tariffs vector $\mathbf{t}^0 = (t^0_{nj})$ we have an equilibrium solution $\bar{\mathbf{p}}^0 = (\bar{p}^0_{nj})$ solution of $[N - SUE(\mathbf{t})]$, and it is unique. Let L be the Lagrangian of problem [N-SUE($\mathbf{t}^0$)]:

$$L(\mathbf{t}^0, \mathbf{p}, \nu, \mu) = Z(\mathbf{t}^0) - \sum_{n\in\mathcal{N}}\sum_{j\in\mathcal{O}_n} \nu_{nj}p_{nj} - \mu(\sum_{n\in\mathbb{N}}\sum_{j\in\mathcal{O}_n} p_{nj} - 1)$$

where μ and ν are the Lagrange multipliers associated with constraints of [N-SUE]. Let y be the vector ($\mathbf{p}$, μ, ν). An equilibrium solution ($\bar{\mathbf{p}}^0$, $\mu^\star$, $\nu^\star$) must satisfy the first order necessary conditions:

$$\begin{array}{lll}
\frac{\partial Z(\mathbf{t}^0,\mathbf{p})}{\partial p_{nj}} - \mu^\star - \nu^\star_{nj} & = 0 & \forall j \in \mathcal{O}_n, \forall n \in \mathcal{N} \\
\sum_{n\in\mathcal{N}}\sum_{j\in\mathcal{O}_n} \bar{p}^0_{nj} - 1 & = 0 & \\
\nu^\star_{nj}\bar{p}^0_{nj} & = 0 & \forall j \in \mathcal{O}_n, \forall n \in \mathcal{N} \\
\bar{p}^0_{nj} \geq 0, \quad \nu^\star_{nj} \geq 0 & & \forall j \in \mathcal{O}_n, \forall n \in \mathcal{N}
\end{array}$$

We can simplify the previous system. We know that in a solution $\bar{\mathbf{p}}$ of [N-SUE($\mathbf{t}$)] the probability of each service is strictly positive. We thus have $\forall j \in \mathcal{O}_n, \forall n \in \mathcal{N}$ $\bar{p}_{nj} > 0$ and $\nu^\star_{nj} = 0$. The same situation is of course obtained after any perturbation in $\mathbf{t}$. Hence the derivatives of $\nu^\star_{nj}$ with respect to $\mathbf{t}$ are null. And the system becomes

$$\begin{array}{lll}
\frac{\partial Z(\mathbf{t}^0,\mathbf{p})}{\partial p_{nj}} - \mu & = 0 & \forall j \in \mathcal{O}_n, \forall n \in \mathcal{N} \\
\sum_{n\in\mathcal{N}}\sum_{j\in\mathcal{O}_n} \bar{p}_{nj} - 1 & = 0 &
\end{array}$$

we write the first equation evaluated for a given $\bar{p}_{nj}$

$$\frac{\partial Z(\mathbf{t}^0, \mathbf{p})}{\partial p_{nj}} = c_{nj}(\bar{p}_{nj}) + \frac{\phi_n}{\theta}(\ln(\bar{p}_{nj}) + 1) + \frac{1-\phi_n}{\theta}(\ln(\sum_{k\in\mathcal{O}_n} \bar{p}_{nk}) + 1) \ .$$

We note that the first and the second terms are functions of $\bar{p}_{nj}$ while the third depends on the probabilities of all options in nest n. Let $\nabla_p Z(\mathbf{t}^0, \bar{\mathbf{p}}^0) = [\frac{\partial Z(t_0,p)}{\partial p_{nj}}]$ be the Jacobian matrix of $Z(\mathbf{t}^0, \mathbf{p})$ with respect to $\mathbf{p}$ and evaluated at $\bar{\mathbf{p}}^0$. and M is a vector of size J with all elements equal to one. M^T denotes the matrix transpose of M. The system can be written in the matrix form:

$$\nabla_p Z(\mathbf{t}^0, \bar{\mathbf{p}}^0) - \mu M = 0 \tag{4.16a}$$

$$M^T\bar{\mathbf{p}}^0 - 1 = 0 \tag{4.16b}$$

Now we derive the system (4.16) with respect to variable $y = (\mathbf{p}, \mu)$. Each equation is derived with respect to each variable, we obtain the following Jacobian matrix

$$J_y = \begin{bmatrix} \nabla^2_p Z(\mathbf{t}^0, \bar{\mathbf{p}}^0) & -M \\ M^T & 0 \end{bmatrix}$$

where $\nabla^2_p Z(\mathbf{t}^0, \bar{\mathbf{p}}^0)$ is the Jacobian matrix of $\nabla_p Z(\mathbf{t}^0, \mathbf{p}^0)$ with respect to $\mathbf{p}$ and evaluated at $\bar{\mathbf{p}}^0$. The matrix $\nabla^2_p Z(\mathbf{t}^0, \bar{\mathbf{p}}^0)$ is bloc diagonal and it is given by :

$$\nabla^2_p Z(\mathbf{t}^0, \bar{\mathbf{p}}^0) = D + B$$

where D is diagonal matrix and B is bloc diagonal matrix. Each bloc of matrix B corresponds to a service $n \in \mathcal{N}$, it is a square matrix of size $|\mathcal{O}_n|$ and all its elements are equal.

$$D = diag([..., d_{nj}, ...]) \quad B = \begin{bmatrix} b_1 & \dots & b_1 & & & & \\ \vdots & \vdots & \vdots & & 0 & & 0 \\ b_1 & \dots & b_1 & & & & \\ & & & b_2 & \dots & b_2 & \\ & 0 & & \vdots & \vdots & \vdots & 0 \\ & & & b_2 & \dots & b_2 & \\ & & & & & & \\ & 0 & & & 0 & & \ddots \\ & & & & & & \end{bmatrix}$$

where $d_{nj} = c'_{nj}(p_{nj}) + \frac{\phi_n}{\theta p_{nj}}$, $\forall j \in \mathcal{O}_n, n \in \mathcal{N}$ and $b_n = \frac{1-\phi_n}{\theta \sum\limits_{k \in \mathcal{O}_n} p_{nk}}$, $\forall n \in \mathcal{N}$. Here $c'_{nj}(p_{nj})$ is the derivative of $c_{nj}(p_{nj})$ with respect to p_{nj}.

Assume that

$$J_y^{-1} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

To simplify notation, $\nabla_p^2 Z(\mathbf{t}^0, \bar{\mathbf{p}}^0)$ is just written $\nabla_p^2 Z$ for the rest of the proof. Denoting by $\nabla_p^2 Z^{-1}$ the inverse matrix of $\nabla_p^2 Z$, it can be obtained easily that

$$\begin{aligned} B_{11} &= \nabla_p^2 Z^{-1}[I - M^T[M\nabla_P^2 Z^{-1} M^T]^{-1} M\nabla_p^2 Z^{-1}] \\ B_{12} &= \nabla_p^2 Z^{-1} M^T[M\nabla_p^2 Z^{-1} M^T]^{-1} \\ B_{21} &= -[M\nabla_p^2 Z^{-1} M^T]^{-1} M\nabla_p^2 Z^{-1} \\ B_{22} &= [M\nabla_p^2 Z^{-1} M^T]^{-1} \end{aligned}$$

where I denotes an identity matrix of appropriate dimensions. The Jacobian matrix of the system (4.16) of equations with respect to $\mathbf{t}$ is

$$J_t = \begin{bmatrix} \nabla_t \nabla_p Z(\mathbf{t}, \mathbf{p}) \\ 0 \end{bmatrix}$$

From the expression of $\frac{\partial Z(\mathbf{t},\mathbf{p})}{\partial p_{nj}}$ and $c_{nj}(p_{nj})$ we obtain that $\nabla_t \nabla_p Z(\mathbf{t}, \mathbf{p})$ is equal to the identity matrix. From (4.15) we have

$$\begin{bmatrix} \nabla_t P \\ \nabla_t \mu \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} -I \\ 0 \end{bmatrix}$$

Therefore, the derivatives of choice probabilities with respect to $\mathbf{t}$ and evaluated at $(\mathbf{t}^0, \bar{\mathbf{p}}^0)$ are given by

$$\nabla_{\mathbf{t}} \bar{\mathbf{p}} = -B_{11}. \tag{4.17}$$

The derivatives informations $\frac{\partial \bar{p}_{nj}}{\partial t_{mk}}$, $\forall n, m \in \mathcal{N}, \forall j \in \mathcal{O}_n, \forall k \in \mathcal{O}_m$ give informations on how the SUE evolves when one or more tariffs vary. They are also used to build approximation of the SUE local to a tariffs set, the quality of this approximation is evaluated in section 4.8.

4.7 The delivery services pricing problem

We studied in the previous section the concept of nest SUE used to model the behaviour of customers in a system where they interact through congestion effects. The service provider is in general looking for achieving a particular goal or objective function. The provider is then interested in finding optimal set of the system parameters, in our context, the design of delivery services. We denote by SDP the bi-level services design problem addressed by the provider.

4.7.1 Problem formulation

Without loss of generality and for simplicity of analysis, we limit in this first study the provider control to service tariffs t_{nj} for all $n \in \mathcal{N}, j \in \mathcal{O}_n$. Then, for a particular set of tariffs, customers decide their option. This hierarchical process leads to a SUE that depends on the tariffs set by the provider. Denote by $\mathbf{t}$ the tariffs vector. The bi-level problem with SUE constraint can be formulated as follows:

$$\begin{aligned}
\text{SDP} \quad (\text{U}) \qquad & \max_{\mathbf{t},\mathbf{p}} \quad F(\mathbf{t},\mathbf{p}) && && (4.18a)\\
& \text{s. t.} \quad L_{nj} \leq t_{nj} \leq U_{nj} && \forall n \in \mathcal{N}, j \in \mathcal{O}_n && (4.18b)\\
(\text{L}) \qquad & \min_{\mathbf{p}} \quad Z(\mathbf{t},\mathbf{p}) && && (4.18c)\\
& \text{s. t.} \quad \sum_{n,j} p_{nj} = 1 && && (4.18d)\\
& \qquad\quad p_{nj} \geq 0 && \forall n \in \mathcal{N}, j \in \mathcal{O}_n && (4.18e)
\end{aligned}$$

The bi-level problem consists in two sub-problems: (U) which is defined as the upper level problem, and (L) the lower level problem. The upper level decision maker or system manager is called the leader in bi-level terminology, while the lower-level decision maker(s) is(are) called the follower(s). The function $F(\mathbf{t},\mathbf{p})$ is the objective function of the upper-level problem and $\mathbf{t}$ is the decisions or variables vector of the leader. The leader's objective function F is generally a global metric which depends on the entire system as a whole. We consider in this paper that this objective function is the revenue and it is defined by:

$$F(\mathbf{t},\mathbf{p}) = \sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{O}_n} \lambda t_{nj} p_{nj}.$$

Note that our model and results are easily applied to other example of leader objective function like the social welfare for example. Constraints (4.18b) are called the

upper level constraints and set bounds on the leader variables. Constraints (4.18d) and (4.18e) are the lower level constraints related to the lower-level problem. For any control $\mathbf{t}$ of the leader, lower-level problem (L) determines a response based on the minimization in $\mathbf{p}$ of the function $Z(\mathbf{t}, \mathbf{p})$ defined in section 4.5. The leader then influences followers decision by setting tariffs t, thus changing the objective function of the lower level problem, or (in different problems than ours) restricting the feasible set for the followers decisions. The second interaction is that the objective function of the upper-level problem depends on the followers decisions.

In some applications, it is possible to obtain analytically the followers response function with respect to leader decision. In our problem, the followers interact between each others, moreover their response is not entirely rational. Thus, the leader cannot predict exactly followers response set for a given control. We present in the following, three heuristics to solve the bi-level services pricing problem.

4.7.2 Gradient descent algorithm (GDA)

This first algorithm for optimizing the provider's objective function, taking into account the underlined SUE, computes a descent direction along which the upper-level objective function is gradually improved. Intuitively the descent direction is the gradient of the upper-level objective function. For example, considering the revenue maximization problem, this gradient direction is given by:

$$\forall n \in \mathcal{N}, j \in \mathcal{O}_n, \quad \frac{\partial F(\mathbf{t}, \bar{\mathbf{p}})}{\partial t_{nj}} = \lambda(\bar{p}_{nj} + \sum_{m \in \mathcal{N}} \sum_{k \in \mathcal{O}_m} t_{mk} \frac{\partial \bar{p}_{mk}}{\partial t_{nj}}).$$

We observe the use of the sensitivity analysis performed previously, in order to determine the right-hand side term. We then next elaborate the GDA method as follows for the revenue optimization purpose. Figure 4.2 illustrates the general schema of GDA. The heuristic sequentially handles the upper-level problem and the lower-level problem:

- Upper level phase: given a follower equilibrium probabilities, a descent direction for upper-level is computed and tariffs are updated.

- Lower level phase: given the tariffs, the SUE and the sensitivity derivatives are computed.

The detailed steps of GDA are as follows:

- **Step 0:** *Initialization.* Determine an initial set of control variables $\mathbf{t}^0 = (t^0_{nj})_{n \in \mathcal{N}, j \in \mathcal{O}_n}$. Set counter $i = 0$.

- **Step 1:** *Lower-level.* Solve the lower-level problem for t^i using MSA and obtain corresponding SUE $\bar{\mathbf{p}}^{i+1}(\mathbf{t}^i)$.

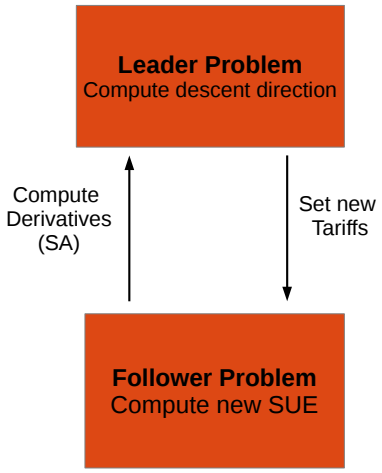

Figure 4.2: GDA main steps

- **Step 2:** *Sensitivity analysis.* Compute the derivatives $\frac{\partial \bar{p}_{nj}^{i+1}}{\partial t_{mk}^{i}}$, $\forall n \in \mathcal{N}, m \in \mathcal{N}, j \in \mathcal{O}_n, k \in \mathcal{O}_m$ at t^i using the sensitivity analysis method.
- **Step 3:** *Upper-level.* Compute descent direction $\forall n \in \mathcal{N}, j \in \mathcal{O}_n$ $d_{nj}^{i+1} := \bar{p}_{nj}^{i+1} + \sum_{m \in \mathcal{N}} \sum_{k \in \mathcal{O}_m} t_{mk} \frac{\partial \bar{p}_{mk}^{i+1}}{\partial t_{nj}^{i}}$.
- **Step 4:** *Move.* Compute $\forall n \in \mathcal{N}, j \in \mathcal{O}_n, t_{nj}^{i+1} = t_{nj}^{i} + d_{nj}^{i+1}$.
- **Step 5:** *Convergence.* If $|F(\mathbf{t}^{i+1}, \bar{\mathbf{p}}^{i+1}) - F(\mathbf{t}^{i}, \bar{\mathbf{p}}^{i})| < \varepsilon$ then stop, else go to step 1 and $i = i + 1$.

The GDA is intuitive and does not need a particular parametrization. To evaluate the quality of GDA, and particularly the efficiency of gradient based descent direction, both in terms of solution and computation cost, we implemented a second heuristic. The latter performs local search and we call it bi-level local search (BLS).

4.7.3 Bi-level local search (BLS)

Local search is a classical heuristic method in combinatorial optimization problems [Talbi 2009]. The idea of a local search is to move from solution to solution in the space of candidate solutions (the search space) by applying minor local changes. This process runs until a solution deemed optimal is found or a time bound is elapsed. Concretely speaking, such algorithm starts from an initial set of control variables $\mathbf{t}_0$. Then, at each iteration i, a neighborhood set V^i of candidate solutions is built. Then the algorithm evaluates the leader's function $F(\mathbf{v}, \bar{\mathbf{p}}(v))$, for all $\mathbf{v} \in V^i$ and pick

the neighbor with the best evaluation. This step is a direction finding step similar to step 3 in GDA, but without computing the gradient of the Leader's objective function. If the best neighbor candidate improves the objective function then it becomes the new current solution. These steps are repeated until no improvement can be obtained.

Figure 4.3 illustrates the progress of a local search for an example with two leader's variables. The numbered points designate the current solution at a given iteration, the grey points surrounding a numbered point are the corresponding neighbors.

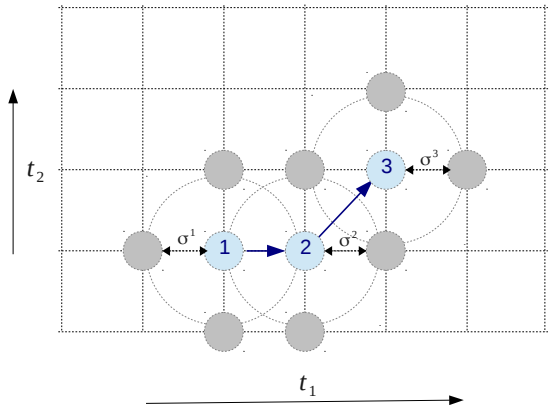

Figure 4.3: Local search progress.

Different neighborhood structures are possible for local search, the size and the construction method of the neighborhood need to be well defined in a way to have the appropriate trade-off between exploration and computation cost. For our services design problem, the neighborhood V^i of a current solution $\mathbf{t}^i$ at iteration i is a set of tariff vectors, having the same size then $\mathbf{t}^i$. Each neighborhood is constructed by locally modifying $\mathbf{t}^i$, by adding or subtracting a radius σ^i to some of the tariffs of $\mathbf{t}^i$. Thus the neighborhood V^i contains all possible tariffs vector $\mathbf{v} = (v_{nj})_{n\in\mathcal{N},j\in\mathcal{O}_n}$ of the form $v_{nj} = t^i_{nj} + \Delta, \forall n \in \mathcal{N}, j \in \mathcal{O}_n$ where $\Delta \in \{-\sigma^i, 0, \sigma^i\}$. The radius σ^i is a non-negative decreasing sequence. Thus if the total number of tariffs is I, the size of the neighborhood is I^3. This neighborhood structure is slightly bigger than the one depicted in figure 4.3 where some neighbors are not shown (corner points) for more clarity. The steps of the BLS algorithm can be described as follows:

- **Step 0:** *Initialization.* Determine an initial set of control variables $\mathbf{t}^0 = (t^0_{nj})_{n\in\mathcal{N},j\in\mathcal{O}_n}$. Set counter $i = 0$.
- **Step 1:** *Neighborhood.* Build the neighborhood set V^i of the solution $\mathbf{t}^i$.
- **Step 2:** *Evaluation.* At each neighbor $\mathbf{v}$, perform MSA to get the corresponding SUE $\bar{\mathbf{p}}(\mathbf{v})$ and evaluate leader objective function $F(\mathbf{v}, \bar{\mathbf{p}}(\mathbf{v}))$.

- **Step 3:** *Selection.* Select the best neighbor $\mathbf{t}^{i+1} = \text{argmax}_{\mathbf{v}\in V^i} F(\mathbf{v}, \bar{\mathbf{p}}(\mathbf{v}))$.
- **Step 4:** *Convergence.* A stopping criterion can be when the leader objective can no more be improved. If $|F(\mathbf{t}^{i+1}, \bar{\mathbf{p}}^{i+1}) - F(\mathbf{t}^i, \bar{\mathbf{p}}^i)| < \varepsilon$ then stop, else go to step 1 and $i = i + 1$.

The BLS algorithm, like the GDA, iteratively explores the solution space from a point to another in order to find the best one. BLS works with a randomly generated neighborhood of candidates. While one may expect that evaluating different candidates around the current solution helps finding a better descent direction, the major inconvenient of this heuristic is that at each neighbor of a current solution, a SUE is computed using the MSA. As the size of problem instance increases, at each iteration of BLS the neighborhood is larger, increasing the number of calls to MSA, that moreover, requires more computation time. We are thus faced to classical issues of local search related to how to find a good compromise between the exploration strength (size of the neighborhood) and the computation expense.

Contrary to GDA, the BLS algorithm needs parameter configuration which is the radius (σ^i) for the construction of the neighborhood V^i. At the first iterations, the radius should lead to rapidly get close to the optimum. As the algorithm approaches the optimum, the radius has to decrease. We use a predetermined decreasing sequence of decreasing radius $\sigma^i = \frac{1}{i}$. Experimentally, as we will see in section 4.8, more efficient radius sequence can be used for improving both the quality of the solution and the computation time of the algorithm.

In order to reduce computation time, we introduce in the next section, a third heuristic that combine the local search technique and useful informations obtained by the sensitivity analysis.

4.7.4 Sensitivity analysis based local search (SLS)

The SLS is proposed in order to overcome the highlighted drawback of the BLS which is the computation of a the SUE for each of the candidate solution in the neighborhood of the current solution (step 2 of BLS algorithm). We suggest instead to use the sensitivity analysis result and to compute an approximation of the SUE. Specifically, at iteration i, given the current tariffs set $\bar{t}^i = (t^i_{nj})$, MSA is performed once in order to compute the SUE $\bar{p}^i = (\bar{p}^i_{nj})$. Then, for a given neighbor $\mathbf{v} = (v_{nj})$, the corresponding SUE is approximated by the following formulae:

$$\forall n \in \mathcal{N}, j \in \mathcal{O}_n, \quad \bar{p}_{nj} = \bar{p}^i_{nj} + \sum_{m\in\mathcal{N}} \sum_{k\in\mathcal{O}_m} \frac{\partial \bar{p}_{nj}}{\partial t_{mk}} (v_{mk} - t^i_{mk}). \tag{4.19}$$

The approximated choice probabilities are then injected in the upper level objective function to evaluate each neighbor. Thus, at each iteration MSA is called one time for (SLS) while it is called for each neighbor for BLS. The SLS algorithm is composed of the same steps as the BLS algorithm but step 2, which is replaced by the following :

- **Step 2':** *Evaluation.* Run MSA and compute the corresponding SUE for the current solution. For each neighbor, compute an approximation of the SUE using (4.19) and evaluate the leader objective function.

These three methods for solving SDP are evaluated and compared in different numerical scenarios within the next section.

4.8 Numerical results

We conduct in this section several experimentations that focus on two main features of the delivery services system. First the correlation between options has impacts on customer choice process. This will be shown by the effect of the nested coefficient on options probabilities at equilibrium in subsection 4.8.1. Consequently the correlation between options impacts also the leader objective function as illustrated in subsection 4.8.2. The second feature is technical and is related to the benefit of combining local search technique with sensitivity analysis in the solution of the services design problem. The comparison of the three heuristics is conducted in subsection 4.8.2 for two delivery system configurations.

All scripts and codes were written in Scilab 5.3.3, experimentations were made on a laptop running with processor Intel Core 2 DUO CPU T7100 1.80 GHZx2 and 2.9 Go RAM.

We study first the example 1 described by figure 4.4. We consider two nests , i.e. $\mathcal{N} = 2$, that are Home delivery (D), indexed by 1, and Pick up (P), and the second by 2. Each nest contains two options. The conversion coefficients of congestion effect that depends on the service type, is $\beta_1 = 100$ for (D) service, and $\beta_2 = 10$ for (P) service. The arrival rate of demand is $\lambda = 40$ and the rate at which a parcel stays in a pick up location is $\mu = 1$. The storage capacities pick up locations are respectively $K_{21} = 15$ and $K_{22} = 10$ parcels. For (D) services, delivery capacities are respectively $K_{11} = 40$ and $K_{12} = 30$ parcels per unit of time.

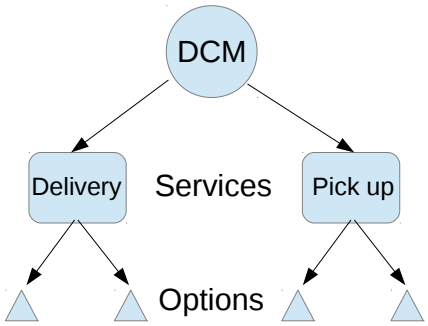

Figure 4.4: Delivery system with two services

4.8.1 Stochastic user equilibrium

The SUE is computed by an implementation of MSA. The threshold for stopping criteria is set to $\varepsilon = 10^{-3}$. In the example of Figure 4.4 , we observed that the MSA converges quickly to the SUE (in less than 10 iterations).

First we analyse the effect of the nesting coefficients. Figure 4.5 depicts the variations of the SUE when one tariff, here t_{21}, increases from 0 to 20. Two situations are considered: $\phi_2 = 1$ (no correlation and the choice model corresponds to the multinomial Logit), $\phi_1 = \phi_2 = 0.5$ (with correlation). Every option is associated to two curves, each curve corresponds to a nesting coefficient.

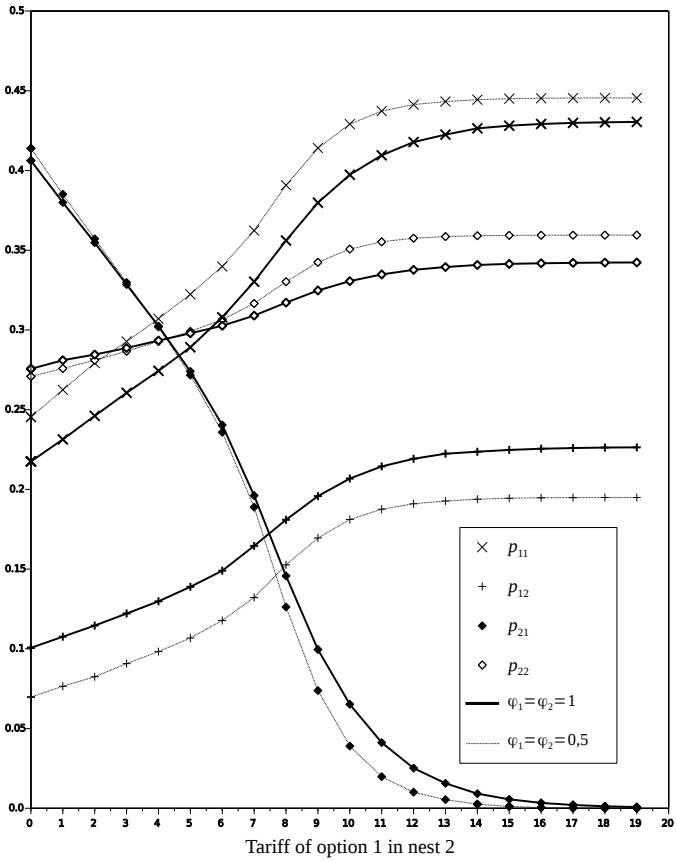

Figure 4.5: Options comparaison in Logit and nested Logit models

Naturally as tariff t_{21} increases, option 1 in nest 2 is less and less attractive compared to others options and the probability p_{21} decreases. The decrease is greater in the nested situation than in the logit situation. Indeed, this decrease is related to two facts. First, the nest 2 has less weight compared to other nests. Second, option 1 in nest 2 has less weight compared to other options in the same nest.

Considering the other options, we see that all probabilities increase. Again, this

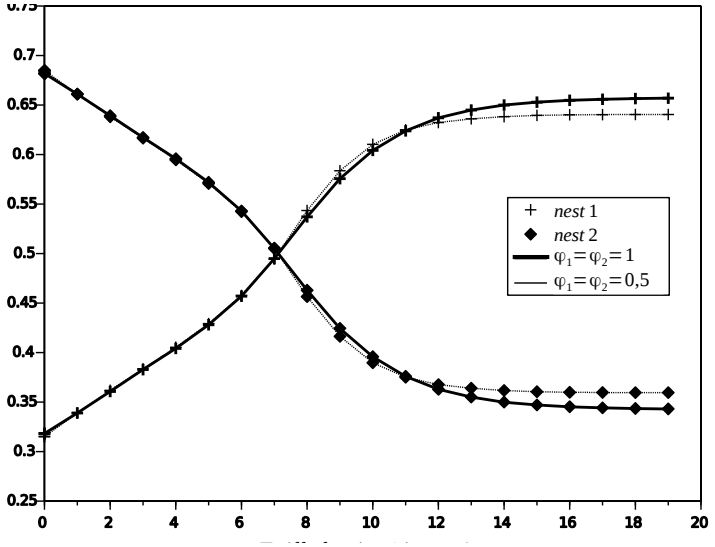

Figure 4.6: services comparaison in Logit and nested Logit models

increase however depends on the nesting coefficient. We see that p_{22} increases more in the nested situation than in the logit situation. That means that option 2 in nest 1, captures more customers who leave option 1 in nest 2 in the nested situation than in the logit situation. In the same time the increase of option 2 is far from counterbalancing the decrease of the option 1, it is then clear that globally customers are moving from nest 2 to nest 1. Figure 4.6 depicts the probabilities of each service, computed as the sum of the probabilities of its options, in the two nesting situations.

In nest 1, we observe clearly that option 1 captures more customers in the nested situation than in the logit one. At the opposite option 1 is depreciated in the nested situation. As mentioned before, in the logit situation all options compete at the same level, and thus an option attracts customers as it outperforms any other option. In the nested situation, options in different nests compete through the weights of their corresponding nests. Options in the same nest compete directly for the distribution of customers attracted by that nest.

4.8.1.1 Sensitivity analysis

We evaluate now the quality of the derivatives of the equilibrium probabilities $\bar{p}$ with respect to tariff, obtained by the sensitivity analysis described in section 4.5. Table (4.1) shows the computed derivatives of choice probabilities with respect to tariffs, and evaluated at a given tariffs set (here $t_{11} = t_{12} = t_{21} = t_{22} = 3$).

The derivatives computed using the sensitivity analysis can be used to estimate the choice probabilities given the tariff of each option. First we compute using MSA the SUE corresponding to the tariffs set $t_{11} = t_{12} = t_{21} = t_{22} = 3$. The values of resulting options probabilities are given by column "Initial" in table (4.2). Then for a perturbation of one tariff the new exact SUE computed using the MSA is given by column "Exact" and the estimated SUE (first order approximation using derivatives

Options Probabilities	Derivatives of choice probabilities			
	t_{11}	t_{21}	t_{12}	t_{22}
p_{11}	- 0.1174628	0.0509520	0.0486570	0.0178537
p_{12}	0.0509520	- 0.0794786	0.0208691	0.0076575
p_{21}	0.0486570	0.0208691	- 0.0883142	0.0187881
p_{22}	0.0178537	0.0076575	0.0187881	- 0.0442993

Table 4.1: Sensitivity analysis at $t_{11} = t_{12} = t_{21} = t_{22} = 3$.

by applying (4.19)) is given by column "Estimated".

Choice Probabilities	Initial	Variation with $\delta t_{11} = +1.00$		Variation with $\delta t_{21} = +1.00$	
		Exact	Estimated	Exact	Estimated
p_{11}	0.2454	0.1315	0.1279	0.2892	0.2940
p_{12}	0.0697	0.1257	0.1207	0.0885	0.0906
p_{21}	0.4139	0.4620	0.4625	0.3271	0.3256
p_{22}	0.2708	0.2906	0.2887	0.2810	0.2896

Table 4.2: Comparison of estimated SUE and exact SUE.

We can observe that the estimated SUE are very close to the exact SUE. This means that the sensitivity analysis gives good result to approximate the impact of control parameters like tariffs on the equilibrium of the customer choices.

4.8.2 Services design problem

We now illustrate the algorithms proposed in section 4.7 in order to optimize the revenue of the delivery provider by controlling the tariffs of the different options in each service. For more clarity, we consider a second example made by removing an option from nest 1 which now contains only one option, while the nest 2 contains 2 options. We also assume that one option is not controllable in the sense that the provider cannot modify the tariff (otherwise the solution is trivial). This assumption holds in a competitive market between several providers that induces some tariff constraint.

Figure 4.7 illustrates the leader revenue as a function of tariffs t_{11} and t_{21}. We observe that this function has good properties of concavity and a single maximum.

The table 4.3 illustrates the comparison of the three heuristics GDA, BLS and SLS. The sequence of radius for BLS and SLS is $\sigma^i = \frac{1}{i}$. The algorithm BLS gives the best solution in terms of revenue. But this algorithm requires more computation time than the others. We expect that we can reduce the computational time of BLS and SLS algorithms by choosing appropriately the radius. We then consider different radius for BLS and SLS of the form $\sigma^i = \frac{a}{b+i}$. For each we found a particular radius that improves the corresponding heuristic. The improved radius $\sigma_b^i = \frac{2}{1.5+i}$ gives the

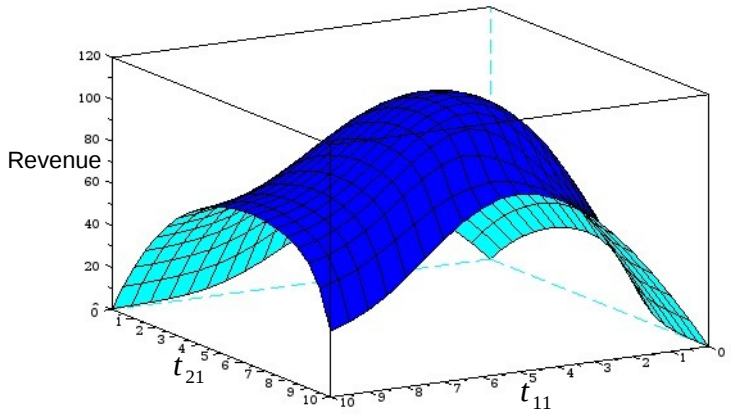

Figure 4.7: Leader revenue depending on tariffs t_{11} and t_{21}.

	Ex 1: 3 services		
	GDA	SLS	BLS
Revenue	89.534	89.536	89.553
Tariffs	3.928 5.575	3.926 5.577	3.948 5.456
Nb iter	69	149	132
Time (s)	0.636	1.492	6.97

Table 4.3: Heuristics comparison with default radius.

best results for BLS while the improved radius $\sigma_s^i = \frac{8}{6+i}$ gives the best result for SLS. Those results are summarized in table 4.4.

	Ex 1: 3 services		
	GDA	SLS	BLS
Revenue	89.534	89.571	89.571
Tariffs	3.928 5.575	3.862 5.646	3.872 5.661
Nb iter	69	34	30
Time (s)	0.636	0.376	1.656

Table 4.4: Heuristics comparison with improved radius.

We observe that these improved radius help both BLS and SLS algorithms reach a better solution in a reduced time. The sequence of improved radius are obtained empirically and the optimization of these parameters is an important improvement of BLS and SLS which is out of the scope of this work.

An important feature of our bi-level services design problem is the nested model parameters (the correlation and the dispersion factors) that describes the DCM customer choice at the lower level. Figure 4.8 illustrates the impact of the nested correlation factor ϕ_2 on the provider's revenue. We compare the revenue obtained with the Nested model with the multinomial Logit (i.e. $\phi_2 = 1$). For each value the optimal tariffs are computed using (GDA) for example. The solid curve represents the revenue obtained if tariffs are fixed to what is obtained assuming (MNL) model ($\phi_2 = 1$). In others words, the nested structure is ignored. While the dotted curves represent the revenue computed with the nested model for each value of ϕ_2. We can notice that the more correlation (ϕ_2 decrease) the bigger the gap between the two curves. That means that the computed tariffs based on (MNL) are more and more sub-optimal when ignoring the nested choice structure.

We examine in a second analysis, the effect of the dispersion parameter θ. This parameter characterizes the degree of rationality of customers in their decision. We see in figure 4.9 that for low value of θ (weak rationality) the revenue is high, and as θ increases, the revenue decreases. This is intuitive since the more customers are rational the more they are influenced by perceived costs and then by the provider's tariffs. Moreover the impact of θ is slightly the same for different levels of nesting coefficient.

The last numerical test studies the scalability issue of our algorithms. We consider the same 2 services but with more options (3 options for each service). For BLS and SLS, we found empirically that the best values for radius are respectively $\sigma_b^i = \frac{6}{7+i}$, $\sigma_s^i = \frac{5}{8+i}$. For those settings, we report the results in table 4.5 along with the results of GDA. We observe particularly that the BLS algorithm is very time consuming and does not scale well. Therefore, for large instances, the SLS algorithm seems to give the best trade-off between quality of the solution and scalability.

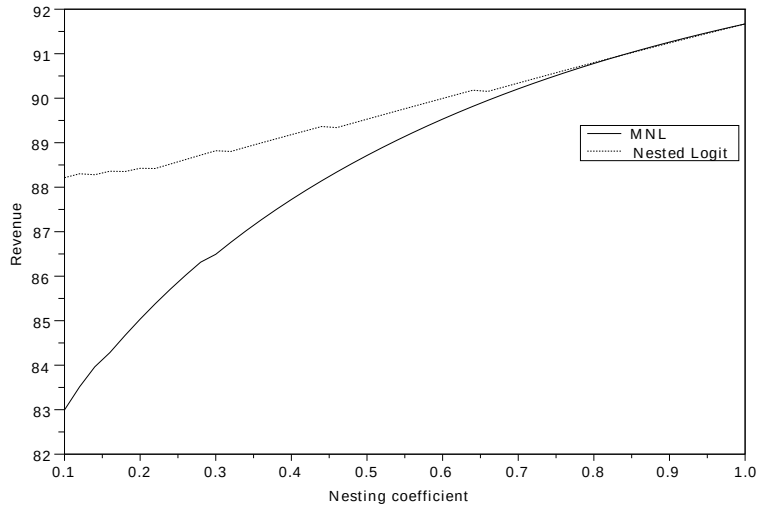

Figure 4.8: Nesting coefficient ϕ_2 on revenue

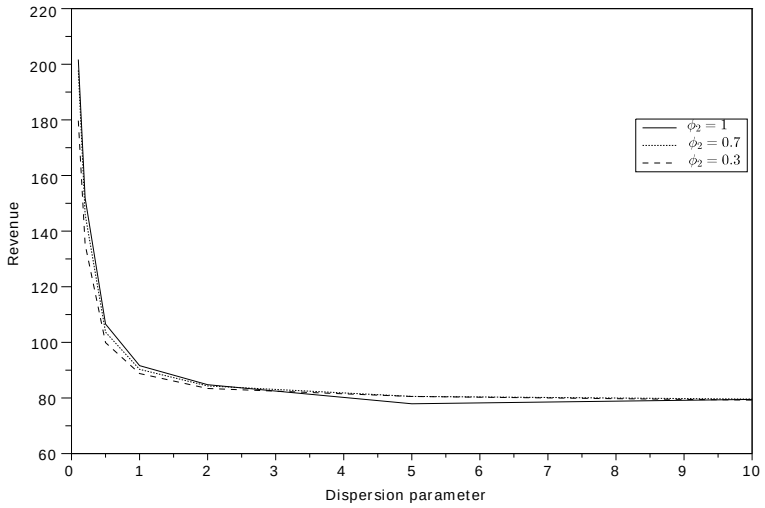

Figure 4.9: Dispersion parameter θ on revenue

	Ex 1: 6 services		
	GDA	SLS	BLS
Revenue	78.964	79.311	83.142
Nb iter	527	32	24
Time (s)	14	1.3	41

Table 4.5: Heuristics comparison for a large instance with 6 options.

4.9 Conclusion

In this chapter, we study a pricing problem for e-commerce last mile delivery system. The problem includes several decision makers having complex interactions and different, even conflicting, goals. These interactions are divided in two level. The first interaction is between the services provider and all the customers. The second interaction is between each customer and all the others.

We propose a bi-level model that reflects the first hierarchical interaction. At the upper level, the provider controls and optimizes service tariffs in order to achieve a certain goal like the overall revenue. At the lower level, users react by choosing selfishly their delivery option according to their utility function. A service utility function incorporates the provider's tariff and a congestion measure.

The second interaction, between customers, is modelled using a particular discrete choice model, the nested logit model. The nested model captures the hierarchical choices structure since services are grouped by type. The steady state customers choices are obtained by computing the stochastic user equilibrium associated to the nested Logit model. We use for this purpose the method of successive averages. We also present a sensitivity analysis for the SUE that enables to compute the derivatives of equilibrium probabilities with respect to services tariffs.

A gradient descent algorithm and a local search heuristic are proposed for the bi-level pricing problem with stochastic equilibrium constraint. Moreover, based on the sensitivity analysis of the equilibrium, we provide an improved local search method that gives encouraging results.

In future works we suggest to investigate the possible extension of the model to additional parameters of the delivery system that can be controlled and optimized by the provider. Possible parameters are the delivery capacities and the relay station capacities. This extension is challenging since it involves discrete leader's variables. Thus the resulting model would be a mixed integer bi-level problem with equilibrium constraint, and new solution methods need to be investigated.

Conclusions and perspectives

Conclusion

All over the world, with a continuous growth, the e-commerce has imposed itself as a major retail market. The supply chain in e-commerce is a main factor of its development and success. It has known consecutive evolutions impacting the involved actors, interactions and process. The maturity of e-commerce is today shown by the emergence of large e-fulfillment networks and the development of a global supply chain. The wide success of e-commerce has been achieved by overcoming plenty of crucial issues arising in the optimization of e-fulfillment process and in the management of the logistic platforms.

This thesis aims to contribute in the understanding of the logistic challenges in e-fulfilment and to propose relevant models and methods to help decision making. In a competitive environment, the supply chain is a complex system that has to deal with fluctuating situations under high requirement of reliability and profitability. Our contributions are related to two distinct phases: Orders picking and shipping, and last-mile delivery. We precisely propose:

- An integrated approach for resources design and operations planning in picking and shipping phases.

- A bi-level approach for last mile delivery services pricing with quality-sensitive customers.

In the first part, we study the process of order picking and shipping in order to investigate the potential benefit of two novel features: the coordination between order picking and shipments schedules, and the integration of resources design and picking and shipping planning. This integrated approach aims to enhance the flexibility of the supply chain and to reduce global cost while fitting the reliability requirement.

We first define a multi period model for the integrated picking and shipping problem. The problem looks for an operational plan that minimizes over the planning horizon a global cost function including labour cost, trucks cost, penalty and docks occupation. The optimal solution corresponds to a trade-off between several costs with different structure under complex operational constraints.

To solve the proposed model, we have designed a matheuristic that determines gradually decisions in three phases involving the solution of a MILP. The method is enhanced by speed-up techniques that exploit lower bounds. It has been tested on real-size instances based on data provided by a major logistic company. The three-phase matheuristic provides encouraging results compared to commercial solver in finding good solutions in a reasonable amount of time.

Our second contribution presented in chapter 3 is an advanced model that aims to capture the information acquisition and the decision making dynamics of e-fulfillment. It is based on a rolling horizon procedure where decisions are determined period by period based on a sequence of solutions of the integrated picking and shipping problem over a reduced horizon.

Moreover, we proposed several approaches that deal with demand uncertainty in order to determine efficiently the quantities of postponed orders. One first approach is based on the estimation of the impact of order postponements using a linearised resources cost. We also propose different mono-scenario techniques that construct a solution based on a representative scenario of the uncertain future demand. The proposed approaches are compared and evaluated based on a-priori and a-posteriori bounds.

The two models proposed for the integrated picking and shipping problem can be coupled to form a global tool. While the second model focuses on tactical decisions (number of workers and trucks) under uncertainty, the first model, based on these decisions and on deterministic information of the current day, determines the operations planning for each order.

The second part of the thesis is related to the last mile delivery phase. In e-commerce, this phase is crucial for the success of the delivery and for the e-retailer image. An efficient delivery system should offer various services and predict customers behaviour. In a highly competitive environment, customers are sensitive to the tariff of a delivery service, but also to its quality. The latter can be damaged by congestion effect induced by customer decisions. We introduce a bi-level model for last-mile delivery services pricing with two family of services. At the upper level the service provider determines services tariffs in order to maximize its revenue. At the lower level, customers choose selfishly their service according to the tariff and to the congestion. The reaction of customers is computed as the equilibrium state based on the nested logit model.

Our study includes computation and sensitivity analysis of the equilibrium state corresponding to a given set of tariffs. We illustrate the impact of the nested structure of services on the customer choices. For the bi-level services pricing problem, we provide two classical heuristics: a gradient descent algorithm and a local search heuristic. We introduce a third heuristic that combines the local search technique with the sensitivity analysis of the lower level equilibrium. This heuristic shows encouraging results in term of solution quality and computation time.

Perspectives

Perspectives for the presented works can be drawn on different levels. Short term perspectives related to specific extensions of each work are presented in the corresponding chapter.

On a long term, concerning the integrated picking and shipping problem, a first perspective is to investigate alternative solution methods based on orders aggregation/disaggregation techniques. The idea is to define an appropriate method to cluster orders. Then IPSP is solved with the grouped orders as input. Finally, a disaggregation step uses the solution and determines decisions related to each order. The process of aggregated process needs to be defined including lot-sizing, postponement and mode change. Of course, the optimization phase will be facilitated and solutions will be obtained in shorter time. The challenges here are twofold. First, the aggregation/disaggregation method should not prevent information tracking on orders process. Second, it will be interesting to investigate if the aggregation leads to a deterioration of the solution value. In fact, the process of aggregated orders has reduced number of possibilities compared to the process of orders individually. It would be then crucial to look for an efficient aggregation method that realizes the best trade-off between optimality and computation cost.

The multi-period procedure with rolling horizon appears to be a good framework to deal with data uncertainty. In this thesis we propose approaches to treat orders arrival uncertainty. However, other uncertainty are inherent to e-fulfilment process like order process time or the workers availability. We then suggest to include these uncertainties to the framework and to adapt accordingly the proposed solution methods.

Regarding the second part of the thesis dedicated to the last-mile delivery services, the discrete choice model can be refined by considering a multi-class representation. Indeed in real life, customers perceive in different way the same service utility according to their revenue or habits for example. The resulting equilibrium where customers perceive the congestion differently, is challenging and it will be interesting to investigate if the *method of successive averages* remains efficient or there will be a need for an advanced method.

A second perspective related to the services pricing problem is to consider a dynamic version where the provider optimises tariffs frequently, day by day, regarding the occurrence of influencing events like big fluctuations in demand or services capacities change. Moreover, the services pricing problem should also take into account the influence of the competition. Indeed, the competition imposes to the delivery company to offer attractive tariffs. In the same time, the company designs services capacities in order to manage the congestion and remain attractive. Thus, an approach for services design and pricing with competition will help the company to find the most profitable strategy for capacities investments and market share.

Bibliography

[Agatz 2008] Niels Agatz, Moritz Fleischmann and Jo Van Nunen. *E-fulfillment and Multi-Channel Distribution*. European Journal of Operational Research, vol. 187, no. 2, pages 339–356, 2008. (Cited on pages 3 and 10.)

[Albareda-Sambola 2014] Maria Albareda-Sambola, Elena Fernandez and Gilbert Laporte. *The dynamic multiperiod vehicle problem with probabilistic information*. Computers and Operations Research, vol. 48, no. 1, pages 31–39, 2014. (Cited on page 51.)

[Alp 2006] O. Alp and T. Tan. *Tactical capacity management under capacity flexibility in make to stock systems*. IEE Transactions, vol. 40, no. 3, pages 221–237, 2006. (Cited on page 18.)

[Alptekinoglu 2005] Aydin Alptekinoglu and Christopher S. Tang. *A model for analysing multi-channel distribution systems*. European Journal of Operational Research, vol. 163, no. 3, pages 802–824, 2005. (Cited on page 11.)

[Anthony 1965] R. Anthony. Planning and control systems : A framework for analysis. studies in management control. Division of Research, Graduate School of Business Administration, Harvard University,, 1965. (Cited on page 2.)

[Asdemir 2008] Kursad Asdemir, Varghese S. Jacob and Ramayya Krishnan. *Dynamic pricing of multiple home delivery options*. European Journal of Operational Research, vol. 196, no. 2009, pages 246–257, 2008. (Cited on page 74.)

[Atamturk 2001] A. Atamturk and D.S. Hochbaum. *Capacity acquisition, subcontracting and lot sizing*. Management Science, vol. 47, no. 8, pages 1081–1100, 2001. (Cited on page 17.)

[Ball 2011] M. O. Ball. *Heuristics based on mathematical programming*. Surveys in Operations Research and Management Science, vol. 16, no. 1, pages 21–38, 2011. (Cited on page 26.)

[Baptiste 2008] P. Baptiste, E. Gaudimier and E Alsene. *Integration of production and shipping planning : A cooperative approach*. Journal of Production, Planning and Control, vol. 19, no. 7, pages 645–654, 2008. (Cited on page 16.)

[Bekhor 2003] Shlomo Bekhor, Lena Reznikova and Tomer Toledo. *Application of cross-nested logit route choice model in stochastic user equilibrium traffic assignment*. Transportation Research Record, vol. 2003, no. 1, pages 41–49, 2003. (Cited on pages 81 and 84.)

[Birge 2011] John R. Birge and Fraonçois Louveaux. Introduction to stochastic programming. Springer, 2011. (Cited on page 68.)

[Bouhtou 2007] M. Bouhtou, G. Erbs and M. Minoux. *Joint optimization of pricing and resource allocation in competitive telecommunications networks.* Networks, vol. 50, no. 1, pages 37–49, 2007. (Cited on page 76.)

[Boulanouar 2014] Mohamed Boulanouar. *Game theoretical models and algorithms applied to transportation networks. Master thesis.* Avignon university, 2014. (Cited on page 63.)

[Braekers 2016] K. Braekers, K. Ramaekers and I. Van Nieuwenhuyse. *The vehicle routing problem: State of the art classification and review.* Computers and Industrial Engineering, 2016. (Cited on page 74.)

[Brotcorne 2000] L. Brotcorne, M. Labbe, P. Marcotte and G. Savard. *A bilevel model and solution algorithm for a freight tariff-setting problem.* Transportation science, vol. 34, no. 3, pages 289–302, 2000. (Cited on page 76.)

[Brotcorne 2008] L. Brotcorne, P. Marcotte and G. Savard. *Bilevel Programming: The Montreal School.* Information systems and operational research, 2008. (Cited on page 76.)

[Cantarella 2005] G.E Cantarella, G. Pavone and A. Vitetta. *Heuristics for urban road network design: Lane layout and signal setting.* European journal of operational research, vol. 175, pages 1682–1695, 2005. (Cited on page 77.)

[Cattaruzza 2015a] D. Cattaruzza, N. Absi, D. Feillet and J. Gonzalez-Feliu. *Vehicle routing problems for city logistics.* Euro Journal on Transportation and Logistics, pages 1–29, 2015. (Cited on page 5.)

[Cattaruzza 2015b] D. Cattaruzza, B. Tounsi, L. Brotcorne and F. Semet. *A Matheuristic for the Packaging and Shipping Problem.* In Odysseus, 2015. (Cited on page 9.)

[Cetinkaya 2005] S. Cetinkaya, F. Mutlu and C. Y. Lee. *A comparison of outbound dispatch policies for integrated inventory and transportation decisions.* European Journal of Operational Research, vol. 171, no. 2006, pages 1094–1112, 2005. (Cited on page 17.)

[Ceylan 2003] H. Ceylan and M.G.H. Bell. *Traffic signal timing optimisation based on genetic algorithm approach, including drivers' routing.* Transportation research part B, vol. 38, pages 329–342, 2003. (Cited on page 77.)

[Chen 1991] M. Chen and A.S. Alfa. *Algorithms for solving fisk's stochastic traffic assignment model.* Transportation Research, vol. 25, no. 6, pages 405–412, 1991. (Cited on page 85.)

[Chen 2010] Z. L. Chen. *Integrated production and outbound distribution scheduling: Review and extensions.* Operations Research, vol. 58, no. 1, pages 130–148, 2010. (Cited on page 16.)

[Colson 2005] B. Colson, P. Marcotte and G. Savard. *Bilevel programming : A survey.* 4OR, vol. 3, pages 87–107, 2005. (Cited on page 75.)

[Cordeau 2015] J. F. Cordeau, M. Dell'Amico, S. Falavigna and M. Lori. *A rolling horizon algorithm for auto-carrier transportation.* Transportation Research Part B, vol. 76, no. 1, pages 68–80, 2015. (Cited on page 51.)

[Crainic 2003] T.G. Crainic. Long haul freight transportation, in handbook of transportation science. Randolph W. Hall, 2003. (Cited on page 4.)

[Crainic 2009] T.G. Crainic, N. Ricciardi and G. Storchi. *Models for evaluating and planning city logistics systems.* Transportation Science, vol. 43, pages 432–454, 2009. (Cited on page 5.)

[Damberg 1995] O. Damberg, J.T. Lundgren and M. Patriksson. *An algorithm for the stochastic user equilibrium problem.* Transportation Research, vol. 30, no. 2, pages 115–131, 1995. (Cited on page 84.)

[Erlander 1975] S. Erlander and N.F. Stewart. *The gravity model in transportation analysis: theory and extensions.* VSP Utrecht, 1975. (Cited on page 75.)

[Facchinei 1999] F. Facchinei, H. Jiang and L. Qi. *A smoothing method for mathematical programs with equilibrium constraints.* Mathematical programming, vol. 85, pages 107–134, 1999. (Cited on page 76.)

[Fiacco 1983] A.V. Fiacco. *Introduction to sensitivity and stability analysis on nonlinear programming.* Programming Academic Press New york, vol. 29, pages 125–139, 1983. (Cited on page 86.)

[Fisk 1980] C. Fisk. *Some developments in equilibrium traffic assignment.* Transportation Research Part B, vol. 14, no. 3, pages 243–255, 1980. (Cited on page 83.)

[Friesz 1990] TL Friesz, RL Tobin, HJ Cho and NJ Mehta. *Sensitivity analysis based heuristic algorithms for mathematical programs with variational inequality constraints.* Mathematical Programming, vol. 48, no. 1, pages 265–284, 1990. (Cited on page 76.)

[Gallo 2010] M. Gallo, L. D'Acierno and B. Montella. *A meta-heuristic approach for solving the urban network design problem.* European journal of operational research, vol. 201, pages 144–157, 2010. (Cited on page 77.)

[Gendron 1999] B. Gendron, T. G. Crainic and A. Frangioni. *Multicommodity capacitated network design*. Telecommunications Network Planning, pages 1–19, 1999. (Cited on page 33.)

[Gicquel 2008] Celine Gicquel, Michel Minoux and Yves Dallery. *Capacitated Lot Sizing models: a literature review*. Open Access Article hal-00255830, Hyper Articles en Ligne, 2008. (Cited on page 15.)

[Gu 2006] Jinxiang Gu, Marc Goetschalckx and Leon F. McGinnis. *Research on warehouse operation: A comprehensive review*. European Journal of Operational Research, vol. 177, no. 2007, pages 1–21, 2006. (Cited on pages 10 and 11.)

[Hayel 2016] Y. Hayel, D. Quadri, T. Jimenez and L. Brotcorne. *Decentralized optimization of last-mile delivery services with non-cooperative bounded rational customers*. Annals of Operations Research, vol. 239, no. 2, pages 451–469, 2016. (Cited on page 77.)

[Helber 2013] S. Helber, F. Sahling and K. Schimmelpfeng. *Dynamic capacitated lot sizing with random demand and dynamic safety stocks*. OR Spectrum, vol. 35, no. 1, pages 75–105, 2013. (Cited on page 15.)

[Judice 2004] J. Judice, P. Martins and J. Nunes. *Workforce planning in a lotsizing mail processing problem*. Computer and operations research, vol. 32, no. 1, pages 3031–3058, 2004. (Cited on page 18.)

[Karimi 2003] B. Karimi, S.M.T. Fatemi Ghomia and J.M. Wilson. *The capacitated lot sizing problem: a review of models and algorithms*. The International Journal of Management Science, vol. 31, no. 2003, pages 365–378, 2003. (Cited on page 15.)

[Lee 2003] C. Y. Lee, S. Cetinkaya and W. Jaruphongsaa. *A dynamic model for inventory lot sizing and outbound shipment scheduling at a third-party warehouse*. Operations Research, vol. 51, no. 5, pages 735–747, 2003. (Cited on page 17.)

[Lee 2013] Amy H.I. Lee, He-Yau Kang, Chun-Mei Lai and Wan-Yu Hong. *An integrated model for lot sizing with supplier selection and quantity discounts*. Applied Mathematical Modelling, vol. 37, no. 1, pages 4733–4746, 2013. (Cited on pages 15 and 16.)

[Liu 2015] H. Liu and D.Z.W. Wang. *Global optimization method for network design problem with stochastic user equilibrium*. Transportation Research Part B, vol. 35, no. 7, pages 20–39, 2015. (Cited on page 76.)

[Luo 1996] Zhi-Quan Luo, Jong-Shi Pang and Daniel Ralph. Mathematical programs with equilibrium constraints. Cambridge university press, 1996. (Cited on page 76.)

[Maher 1998] M. Maher. *Algorithms for logit-based stochastic user equilibrium assignment*. Transportation research part B, vol. 32, no. 8, pages 539–549, 1998. (Cited on page 84.)

[Meng 2001] Q. Meng, H. Yang and M.G.H. Bell. *An equivalent continuously differentiable model and a locally convergent algorithm for the continuous network design problem*. Transportation research part B, vol. 35, pages 83–105, 2001. (Cited on page 77.)

[Meng 2011] Lingyun Meng and Xuesong Zhou. *Robust single-track train dispatching model under a dynamic and stochastic environment: A scenario-based rolling horizon solution approach*. Transportation Research Part B, vol. 45, no. 7, pages 1080–1102, 2011. (Cited on page 51.)

[Mincsovics 2009] G. Mincsovics, O. Alp and T. Tan. *Integrated capacity and inventory management with capacity acquisition lead times*. European Journal of Operational Research, vol. 196, no. 3, pages 949–958, 2009. (Cited on page 18.)

[Pac 2009] M.F Pac, O. Alp and T. Tan. *Integrated workforce capacity and inventory management under labor supply uncertainty*. International Journal of Production Research, vol. 47, no. 15, pages 4281–4304, 2009. (Cited on page 18.)

[Pieper 2001] H. Pieper. Algorithms for mathematical programs with equilibrium constraints with applications to deregulated electricity markets. ph.d. thesis, department of management science and engineering,. Stanford University, Stanford, CA., 2001. (Cited on page 76.)

[Pillac 2013] V. Pillac, M. Gendreau, C. Gueret and A.L. Medaglia. *A review of dynamic vehicle routing problems*. Computers and Industrial Engineering, vol. 225, no. 1, pages 1–11, 2013. (Cited on page 74.)

[Pinker 2001] E.J. Pinker and R.C. Larson. *Optimizing the use of contingent labor when demand is uncertain*. European Journal of Operational Research, vol. 1, no. 1, pages 39–55, 2001. (Cited on page 18.)

[Pironet 2014] Thierry Pironet. *Multi-period Stochastic Optimization Problems in Transportation Management*. PhD thesis, 2014. (Cited on pages 51, 54 and 63.)

[Rakke 2011] Jorgen Glomvik Rakke, Magnus Stalhane, Christian Rorholt Moe, Marielle Christiansen, Henrik Andersson, Kjetil Fagerholt and Inge Norstad. *A rolling horizon heuristic for creating a liquefied natural gas annual delivery program*. Transportation Research Part C, vol. 19, no. 5, page 896–911, 2011. (Cited on page 51.)

[Sheffi 1985] Y. Sheffi. Urban transportation networks: Equilibrium analysis with mathematical programming methods. Prentice-Hall, Inc., 1985. (Cited on pages 75, 82, 83, 84 and 85.)

[Stevanovic 2006] Dalibor Stevanovic. *Application de modéle logit mixte emboîté dans le cadre de l'estimation de la demande de transport.* PhD thesis, Laval university, 2006. (Cited on pages 80 and 81.)

[Takács 1962] Lajos Takács. Introduction to the theory of queues, volume 584. Oxford University Press New York, 1962. (Cited on page 78.)

[Talbi 2009] E. G. Talbi. Metaheuristics: from design to implementation. John Wiley and Sons, 2009. (Cited on page 92.)

[Tounsi 2015] B. Tounsi, Y. Hayel, D. Quadri and L. Brotcorne. *Mathematical programming with stochastic equilibrium constraints applied to optimal last-mile delivery services*. In International network optimization conference, 2015. (Cited on page 71.)

[Tounsi 2016a] B. Tounsi, Y. Hayel, D. Quadri and L. Brotcorne. *Mathematical programming with stochastic equilibrium constraints applied to optimal last-mile delivery services*. Electronic Notes in Discrete Mathematics, pages 5–12, 2016. (Cited on page 71.)

[Tounsi 2016b] B. Tounsi, F. Semet, L. Brotcorne and D. Cattaruzza. *An integrated multi-period stochastic problem in e-fulfillment optimization*. In Tristan, 2016. (Cited on page 49.)

[Wang 2005] H. Wang and C.Y. Lee. *Production and transport logistics scheduling with two transport mode choices*. Naval research logistics, vol. 52, no. 8, pages 796–809, 2005. (Cited on page 16.)

[Wang 2010] David Z.W Wang and Hong K. Lo. *Global optimum of the linearized network design problem with equilibrium flows*. Transportation research part B, vol. 11, pages 482–492, 2010. (Cited on page 76.)

[Wen 2001] C.H. Wen and F.S. Koppelman. *The generalized nested logit model*. Transportation Research Part B: Methodological, vol. 35, no. 7, 2001. (Cited on page 75.)

[Yang 1994a] H. Yang and S. Yagar. *Traffic assignment and traffic control in general freeway arterial corridor systems.* Transportation Research B, vol. 28, no. 6, pages 463–486, 1994. (Cited on pages 76 and 86.)

[Yang 1994b] H. Yang, S. Yagar, Y. Iida and Y. Asakura. *An algorithm for the inflow control problem on urban freeway networks with user-optimal flows.* Transportation Research, vol. 28, no. B, pages 123–139, 1994. (Cited on page 76.)

[Ying 2003] J. Q. Ying and H. Yang. *Sensitivity analysis of stochastic user equilibrium flows in a bi-modal network with application to optimal pricing.* Transportation research Part B, vol. 39, pages 769–795, 2003. (Cited on page 77.)

[yu Kevin 2005] Chiang Wei yu Kevin and George E. Monahan. *Managing inventories in a two-echelon dual-channel supply chain.* European Journal of Operational Research, vol. 162, no. 2, pages 325–341, 2005. (Cited on page 11.)

[Zhou 2010] Z. Zhou, A. Chen and S. Bekhor. *C-logit stochastic user equilibrium model: formulations and solution algorithm.* Transportmetrica, vol. 8, no. 1, pages 17–41, 2010. (Cited on page 75.)

[Z.L. Chen 2005] G. L. Vairaktarakis Z.L. Chen. *Integrated scheduling of production and distribution operations.* Management science, vol. 51, no. 4, pages 614–628, 2005. (Cited on page 16.)