MODELS FOR OMNI-CHANNEL RETAILING

LINA ZHANG

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

FEBRUARY 2020

ABSTRACT

Though many retailers have embarked on the omni-channel journey, most of them are grappling with the difficulties of order fulfilment and delivery in an omni-channel environment. As customers have become more demanding regarding delivery flexibility and more intolerant of long delivery waiting times, it is imperative for retailers to revise their supply chains and operational systems to embrace omni-channel and guarantee high customer service levels. The main objective in this thesis is to explore key supply chain management (SCM) decisions faced by retailers in transition to omni-channel and to use quantitative methods to examine some of these decisions. Based on extensive review from both academic studies and the business literature, we present a conceptual map pinpointing critical supply chain and operations decisions. The conceptual map encapsulates the main contextual factors including retailer characteristics, demand characteristics, order characteristics and supplier characteristics.

In light of the conceptual map, we examine three research questions. The first research question investigates the operational decisions for fulfilling same day 'Buy-Online-Pickup-in-Store' (BOPS) orders from retail stores. Given the order cut-off time and the deadline (i.e., the earliest collection time promised to customers), analytical models are developed to find the Best Performance Frontiers (BPFs), along which the best combination of the time to commence picking and the minimum picking rate needed can be found for each stipulated service level. Two scenarios have been examined, i.e., when the retailer picks a single wave per cycle and when the retailer organises multiple picking waves per cycle. We also conduct sensitivity analyses to explore the impact of the timing and the magnitudes of demand rate flunctuations on the picking rate required. Our model results and analyses can assist retailers with the decisions on scheduling order picking and setting the picking rate in stores for BOPS fulfilment to achieve a specific service level.

The second research question studies the decisions on the ordering window for BOPS services and the corresponding fulfilment operations. Given a deadline, a longer ordering window leads to a shorter window for order processing and transportation, which requires higher processing rate and/or more late deliveries. However, a longer ordering window can be convenient to customers and retailers can capture last minute shoppers. The strategic decision on the optimal ordering window for BOPS services is clearly a problem at the marketing-operations interface. We develop cost minimisation models to jointly optimise BOPS service offerings (i.e., the order receipt window length) and BOPS fulfilment operations (i.e., the time to start picking and the picking rate needed). We consider different retail scenarios when BOPS orders are

fulfilled from local stores with zero transportation window and when BOPS orders are fulfilled from more central facilities (e.g., warehouses) before being shipped to the collection stores. We apply our model results to analyse and explain the current service offerings of five major UK-based retailers. Accordingly, we provide insights into the key trade-offs for the centralisation level of BOPS fulfilment operations and present two frameworks to delineate the impact of product characteristics and retailer characteristics on BOPS collection speed to be offered, i.e., same- or next-day or over two days.

The third research question examines the strategic decision on from where to fulfil online orders. We first provide a typology of commonly used fulfilment network configurations, as retailers from different product sectors use different combinations of online fulfilment options. We then focus our study on the online groceries fulfilment network and present analytical models to determine how to best fulfil online orders. In our research setting, online demand is further divided into BOPS demand and home delivery (HD) demand. The total cost in the proposed analytical model is expressed as a function of the centralisation level of online picking, which allows the derivation of clear insights into the number of online fulfilment facilities needed. Numerical experiments have been conducted to investigate the impact of different problem parameters, such as the demand density and the maximum travel distance allowable for BOPS collection, on the configuration of the optimal online fulfilment network. The analytical expression of the total cost provide further insights. For example, the size of the retailer' service region plays no role in determining the centralisation level of online fulfilment, as it only affects the magnitude of cost savings from such decisions.

Keywords: Omni-channel retailing; Buy-Online-Pickup-in-Store (BOPS); Home delivery (HD); ordering receipt window; the picking rate; online order fulfilment

ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my supervisors, Professor Bart MacCarthy and Dr. Luc Muyldermans, whose knowledge, experience, vision, intuition, attention to details, patience and encouragement have all been indispensable to the completion of this thesis. I am greatly indebted to them for their support and advice both professionally and personally. It was so inspiring to work with them and there is still much more to learn from them.

I am also grateful to my committee members: Dr Anshuman Chutani, Professor Sanja Petrovic and Professor Kim Hua Tan. They have given many constructive comments on my work at my annual review presentations and their feedback has been helpful in shaping this thesis. I would like to thank Professor Neil Ashworth, the former chief executive officer of CollectPlus and former supply chain director of Tesco.com, for sharing the current practices of the UK retailers with me.

I am deeply appreciative of the unconditional love and support from my parents and my little brother. They are the rock upon which I stand. I would also like to thank my dearest friends including Dr. Lin Wu, Yumeng Zhang, Sera Goktas, Howard Lee and Daphne Lee. Their unwavering love and faith never failed to keep me going.

Contents

ABSTRACT i
ACKNOWLEDGEMENTS iii
LIST OF TABLES vii
LIST OF FIGURES viii
MASTER LIST OF NOTATIONS x
LIST OF ABBREVIATIONS xii
PUBLICATIONS/CONFERENCE PAPERS xiii
Chapter 1 1
Introduction 1
1.1 Omni-channel retailing 1
1.2 Omni-channel supply chain and operational challenges 3
1.3 Research objectives 4
1.4 Problems and solution methodologies 4
1.5 Organisation of the thesis 8
Chapter 2 9
Literature review 9
2.1 Scope and structure 10
2.1.1 Scope 10
2.1.2 Structure 10
2.2 Categorisation of multiple channel strategies 11
2.3 Planning areas for omni-channel SCM 15
2.4 Inventory management 16
2.4.1 Inventory ownership 16
2.4.2 Inventory integration level 19
2.4.3 Inventory control policies 22
2.5 Warehousing and fulfilment 26
2.5.1 Warehousing design and processes 26
2.5.2 Order fulfilment 28
2.6 Last mile delivery 34
2.7 Returns management 38
2.8 A conceptual map of omni-channel SCM 47
2.9 Justification of research questions in the thesis 52
Chapter 3 54
Best Performance Frontier for Buy-Online-Pickup-In-Store order fulfilment 54
3.1 Introduction 54
3.2 Best Performance Frontiers (BPFs) for BOPS order fulfilment 57
3.2.1 Problem description 57
3.2.2 Single wave picking 59
3.2.3 Best Performance Frontier (BPF) for BOPS with single wave picking 62

3.2.4 Key relationships for single wave in-store picking 65
3.3 Single wave picking with two different arrival rates 66
3.3.1 The timing of demand surge 67
3.3.2 The difference between the two demand rates 68
3.4 A generalised store wave picking model 69
3.5 Managerial insights and implications 74
3.6 Concluding remarks 77
Chapter 4 78
Deadline-driven fulfilment: the optimal online ordering window for retail store collection 78
4.1 Introduction 78
4.2 Model formulation and analysis 82
4.2.1. Problem description and assumptions 82
4.2.2 The order fulfilment model when $L = 0$ 86
4.2.3 The order fulfilment model when $L > 0$ 89
4.3 Retail operating strategies 94
4.3.1 Estimation of model parameters for major UK retailers 94
4.3.2 Theoretical optimality for retailers 97
4.3.3 Guidance frameworks for BOPS speed offerings 100
4.4 Concluding remarks 104
Chapter 5 107
Configuring the online order fulfilment networks for 107
omni-channel retailers 107
5.1 Introduction 107
5.2 Relevant literature 109
5.2.1 Physical Network Design (PND) 109
5.2.2 Continuous approximation (CA) technique 111
5.3 A typology of online fulfilment networks 113
5.3.1 Grocery network 113
5.3.2 Non-food fulfilment network 1 115
5.3.3 Non-food fulfilment network 2 116
5.3.3 Non-food fulfilment network 3 118
5.4 Models for the online groceries fulfilment network 119
5.4.1 Problem description 119
5.4.2 The formulation of the total cost function 125
5.4.3 Numerical experiments and results 129
5.4.4 Online fulfilment from DCs or EFCs 138
5.5 Concluding remarks 142
Chapter 6 145
Conclusions and future research 145

6.1 Summary of contributions 145
6.2 Future research 150
A Appendix for Chapter 3 153
A.1 The derivation of NSD and the critical point when the retailer encounters two arrival rates per cycle. 153
A.2 Proof of proposition 4 155
B Appendix for chapter 4 157
B.1 The retailer starts picking before the order cut-off time, i.e., $tw \leq tc$. 157
B.2 KKT necessary conditions for problem (P1) 158
B.3 KKT necessary conditions for problem (P2) 163
B.4 KKT necessary conditions for problem (P3) 167
B.5 Parameter estimation for other retailers 170
Bibliography 173

LIST OF TABLES

Table 1 Differentiation between multi-, cross- and omni-channel retailing. 12
Table 2 Research agendas for key SCM decisions in transition to omni-channel retail. 49
Table 3 Notation for the single and multi-wave store picking problem. 59
Table 4 Workload for each picking wave in the generalised store picking model. 69
Table 5 Generalised store wave picking model - a numerical illustration. 73
Table 6 Examples of BOPS service offerings from British retailers. 79
Table 7 Notation for the cost minimisation models 85
Table 8 Optimal solution when $tw \geq tc$ and $L = 0$. 88
Table 9 Optimal solution for $L > 0$ - case 1 ($L \leq Te \leq \lambda T\mu$). 91
Table 10 Optimal solution for $L > 0$ - case 2 ($\lambda T\mu < Te \leq \lambda T\mu + L$) 93
Table 11 Model parameters 95
Table 12 Model results for retailer service offerings (NSD = 0.98) 98
Table 13 Key features of each picking facility under different centralisation level k. 123
Table 14 Notation for the online order fulfilment problem 123
Table 15 Parameter values in the base case scenario. 130
Table 16 Parameter values for the illustrative example 141
Table 17 Results of the numerical example. 141

LIST OF FIGURES

Figure 1. Omni-channel development progress by country (scale: 1-100). Source: PwC (2017) 2
Figure 2.Categorisation tree in multi-, cross- and omni-channel retailing for retailers and retailing, from Beck and Rygl (2015) 11
Figure 3. Trajectory of retailers moving from multi-channel to omni-channel, adapted from DHL (2016). 15
Figure 4. Omni-channel retailing online order fulfilment processes 16
Figure 5. Inventory ownership decisions for omni-channel retailers.....17
Figure 6. Inventory integration strategies for retailers with multiple channels. 19
Figure 7. An online fulfilment network with four fulfilment options, adapted from Ishfaq and Raja (2018). 29
Figure 8. Classification of delivery services in omni-channel retailing. 34
Figure 9. Product returns options in omni-channel retailing. 39
Figure 10. The conceptual map of key supply chain and operations management decisions for omni-channel retailing. 48
Figure 11. Illustration of the same day BOPS service promise by Sainsbury's. 57
Figure 12. Different response policies to a constant order arrival process. 58
Figure 13. Single wave picking: only the picking wave in the current cycle contributes to NSD in (a) and (b), and both the picking waves in the current and previous cycle contribute to NSD in (c) and (d). 60
Figure 14. BPF for single wave picking with pre-specified order cut-off and deadline times – a same day BOPS example. 62
Figure 15. Single wave picking model with a late demand surge.....66
Figure 16. The minimum required picking rate in an early and a late surge scenario for single wave picking.....67
Figure 17. The minimum picking rate when the arrival rate difference is 300, 500 and 900. 68
Figure 18. A generalised store wave picking model - illustration of 3 waves per cycle. 69
Figure 19. The shift in BPF as the number of picking waves per cycle increases.....74
Figure 20. Trade-offs in BOPS order fulfilment (adapted from Supply Chain 24/7 (2015)). 80
Figure 21. The evolution of order accumulation, WIP inventories, and late orders over time when $t_w > t_c$ and L = 0. 84
Figure 22. The evolution of order accumulation, WIP inventories and late orders over time when L = 2h. 90
Figure 23. The impact of STS transportation window L on the number of late orders. 91
Figure 24. The evolution of order accumulation, WIP inventories and late orders over time when L = 2days. 93
Figure 25. Total fulfilment cost change (%) for Tesco when the total BOPS demand is picked in fewer stores. 100
Figure 26. Optimal BOPS service speed for different retailers 101
Figure 27. Optimal retail operating strategies based on product profile.....103

Figure 28. Optimal retail operating strategies based on retailer profile.....103
Figure 29. The online order fulfilment network for groceries. 115
Figure 30. The online fulfilment network configuration 1 for non-food. 116
Figure 31. The online fulfilment network configuration 2 for non-food. 117
Figure 32. The online fulfilment network configuration 3 for non-food. 118
Figure 33. The regular hexagon-shaped primary influence area of a collection store. 121
Figure 34. Regular hexagonal tessellation of collection stores in the service region **R**. 121
Figure 35. Each picking store handles more online orders as k increases 122
Figure 36. The cost improvement when varying the maximum travel distance d_{max}. 131
Figure 37. The total cost when varying the maximum travel distance d_{max}. 132
Figure 38. The cost improvement when varying the demand density ρ. 132
Figure 39. The cost improvement when varying the BOPS demand ratio θ. 134
Figure 40. The total cost when varying the BOPS demand ratio θ. 134
Figure 41. The cost improvement when varying the delivery vehicle capacity Q. 135
Figure 42. The cost improvement when varying the unit variable transportation cost c_v. 136
Figure 43. The cost improvement when varying the congestion rate α. 137
Figure 44. The cost improvement when varying the fixed facility cost F. 137
Figure 45. Illustration of the numerical example results. 142
Figure 46. Picking wave starts before the demand shift point and (a) finishes before the deadline, (b) finishes after the deadline. 153
Figure 47. Picking wave starts after the demand shift point and (a) finishes before the deadline, (b) finishes after the deadline. 154
Figure 48. The evolution of order accumulation, WIP inventories and late orders when $t_w \leq t_c$. 157

MASTER LIST OF NOTATIONS

Thesis chapter	Notation	Explanation
Chapter 3	N	The number of picking waves for a cycle of orders
	T	The cycle length (one day, i.e., $T = 1$ for BOPS)
	t_c	The order cut-off time
	t_d	The deadline when BOPS orders should be ready for customers to collect
	t_f	The finishing time of the (last) picking wave for a cycle of orders
	t_w	The starting time (i.e., release time) for picking using a single wave
	t_w^*	The starting time for picking using a single wave on the BPF
	$t_{w,j}$	The starting time for picking the *j*-th wave (j= 2, … N)
	$t_{w,j}^*$	The starting time for picking the *j*-th wave (j= 2, … N) on the BPF
	$t_{w,1}'$	The starting time for picking the first wave with *N-1* picking waves
	t^*	Demand rate switch point in the case with two different arrival rates
	λ	BOPS demand arrival rate
	μ	The picking rate in stores using a single wave
	μ^*	The minimum picking rate in stores using a single wave
	μ_N	The picking rate in stores with *N* picking waves for a cycle of orders
	μ_N^*	The minimum picking rate in stores with *N* picking waves
Chapter 4	λ	Order arrival rate, i.e., the number of BOPS orders arrived per time unit
	μ	Order picking rate, i.e., the number of BOPS orders picked per time unit
	c_h	Inventory holding cost, £ per order per time unit
	c_p	Penalty cost, £ per late order per lateness time
	$c_\mu(\mu)$	Unit variable order picking cost, £ per order picked
	$c_\mu'(\mu)$	Marginal variable order picking cost; the first derivative of $c_\mu(\mu)$
	$c_\mu''(\mu)$	The second derivative of $c_\mu(\mu)$
	K	Fixed cost for setting up order picking activities
	t_c	The cut-off time for customers to place orders online
	t_d	The deadline when orders placed by the cut-off time will be ready for collection
	t_w	The time to start processing BOPS orders in a cycle
	t_f	The finish time of overall picking activities in a cycle
	t_p	The time to finish late orders from the previous cycle
	L	The STS transportation window
	T	The order receipt window, i.e., the length of a cycle
	T_e	The fulfilment service window, i.e., $T_e = t_d - t_w$
	f	The fulfilment speed as a fraction of a cycle, i.e., $f = (t_d - t_c)/T$
	k	The centralisation level of online picking operations
	N_k	The number of picking facilities given the centralisation level *k*
	$\boldsymbol{A_k}$	The influence area of each picking facility given the centralisation level *k*
	A_k	The size of $\boldsymbol{A_k}$
	r_k	The radius of $\boldsymbol{A_k}$
	$\boldsymbol{R}$	The retailer's entire service region
	R	The size of $\boldsymbol{R}$
	$\boldsymbol{S}$	The primary influence area of each collection store
	S	The size of $\boldsymbol{S}$
	d_{max}	The maximum travel distance for customers to collect BOPS orders

Chapter 5	ρ	Online customer density per square mile across $\boldsymbol{R}$
	θ	The ratio of BOPS demand in total online demand
	$\overline{d_k}$	The average STS travel distance from a picking facility to a collection store
	α	The picking congestion rate
	λ_0	The baseline order volume that can be handled without being affected by congestion in a picking facility
	Q	The delivery vehicle capacity, i.e., the maximum number of orders/vehicle
	M_k	The maximum picking capacity for each picking facility
	L_{HD}	The total expected all-vehicle travel distance for each picking facility to deliver HD orders
	L_{BOPS}	The total expected all-vehicle travel distance for each picking facility to deliver BOPS orders via STS
	$\bar{l}$	The average distance from a picking facility to a customer point in its influence area
	n	The number of customer points for a picking facility in its influence area
	F	The fixed facility cost
	a	The reorder cost
	h	The unit inventory holding cost
	c_p	The unit order picking cost
	c_0	The unit order picking cost when the congestion rate is zero
	c_f	The fixed cost to conduct a vehicle route
	c_v	The unit variable transportation cost for each mile travelled to deliver HD orders
	c_l	The unit variable transportation cost for each mile travelled to ship each BOPS order
	TF	The total fixed facility cost
	TI	The total inventory holding cost
	TP	The total order picking cost
	$TD1$	The total delivery cost for HD orders
	$TD2$	The total delivery cost for BOPS orders via STS
	TC	The total cost for fulfilling online demand

LIST OF ABBREVIATIONS

ADI: Advance Demand Information

AOV: Average Order Value

BC: Bricks-and-Clicks

BM: Bricks-and-Mortar

BOPS: Buy-Online-Pickup-in-Store

BORS: Buy-Online-Return-in-Store

BPF: Best Performance Frontier

CEP: Courier, Express and Parcel

CM: Clicks-and-Mortar

DC: Distribution Centre

EFC: E-commerce Fulfilment Centre

FCFS: First-Come-First-Serve

HD: Home Delivery

KKT: Karush-Kuhn-Tucker

LICQ: Linear Independence Constraint Qualification

MDP: Markov Decision Process

NSD: Next Schedule Deadline

O2O: Online-to-Offline

PND: Physical Network Design

RC: Returns Centre

ROPS: Reserve-Online-Pick-up-and-Pay-in-Store

SCM: Supply Chain Management

SFS: Ship-From-Store

STS: Ship-To-Store

VRP: Vehicle Routing Problem

WIP: Work-in-progress

PUBLICATIONS/CONFERENCE PAPERS

This thesis includes three original papers that have been previously published/submitted for publication in peer-reviewed journals and several conference papers, as follows:

Peer-reviewed journal articles:

- MacCarthy, B.L., **Zhang, L.,** Muyldermans, L., 2019[1]. Best Performance Frontiers for Buy-Online-Pickup-in-Store order fulfilment. *International Journal of Production Economics*, 211, 251-264.
- Muyldermans, L., **Zhang, L.,** MacCarthy, B.L., 2020[1]. Deadline-driven fulfilment: the optimal online ordering window for retail store collection. Submitted to European Journal of Operational Research.
- **Zhang, L.,** MacCarthy, B.L., Muyldermans, L., 2020. Is There a Sweet Spot in Omni-channel Retailing? Convergence in the UK Omni-channel Retailing Market. Submitted to Journal of Retailing and Consumer Services.

Conference papers:

- **Zhang, L**., Muyldermans, L., MacCarthy, B.L., 2020. Configuring the online order fulfilment networks for omni-channel retailers. *European Operations Management Association (EurOMA) 2020 Managing operations for impact*, 28-30 June, University of Warwick, UK
- **Zhang, L**., MacCarthy, B.L., Muyldermans, L., 2018. Extending online order cut-off time – how late is too late? *20th International Working Seminar on Production Economics(IWSPE)*, 19-23 February, Innsbruck, Austria

[1] Journal articles emanating from the contents of this thesis.

Chapter 1
Introduction

1.1 Omni-channel retailing

The retail sector has always been dynamic with constant reinvention. In today's retail world, the pace of evolution has never been faster due to the advancement of technologies and the proliferation of social media such as Facebook. From the hostile battle between pure online players and traditional bricks-and-mortar (BM) retailers to the expansion of traditional retailers into online channels and vice versa, the line between physical and online retailing has become blurred. The move towards omni-channel retailing with channel-agnostic customers and seamless integration across retail channels is transforming the retail landscape (Brynjolfsson et al., 2013; Forbes, 2019a).

Verhoef et al. (2015) defines omni-channel retailing as the synergetic management of numerous available channels and customer touchpoints in such a way that the customer experience across channels and the performance over channels are optimised. In the business literature, Deloitte defines omni-channel as a direct to consumer business model, where all sales channels ranging from online, mobile, telephonic, mail order, self-service and physical retail stores are aligned and fulfilment processes are integrated to provide consumers with a seamless shopping experience aligned with the retailer's brand proposition (Deloitte, 2015). Both definitions convey the essence of omni-channel retailing, which strives for ultimate flexibility for front-end customer interactions and back-end retail operations. Two common examples of omni-channel initiatives are Buy-Online-Pickup-in-Store (BOPS) and Ship-From-Store (SFS) services (Bayram and Cesaret, 2017). The former allows customers to browse and purchase products online and collect online orders from retail stores, while the latter uses store inventories to deliver online orders directly to customers (Gao and Su, 2017).

Most retailers in many countries have embarked on the journey to be omni-channel to some extent. On one hand, traditional BM retailers such as Walmart in the US and Tesco in the UK have expanded their online businesses and sought to integrate online and offline channels in order to provide omni-channel services such as BOPS and SFS. On the other hand, online retailers such as Amazon and Alibaba have invested in physical stores for BM retailing as well (Bloomberg, 2018). Strategy&, a subsidiary of PricewaterhouseCooper (PwC), has analysed the status of omni-channel execution worldwide. The company has examined retailers across 9 categories from 28 countries in 2017 and ranked them on a scale of 1 to 100 based on four dimensions: the degree of digitalisation for retail sales in the country or category, omni-channel potential, consumer behaviour, and technology infrastructure (PwC,

2017) (see Figure 1). The message from their analysis is clear: omni-channel retail is fast developing across the world.

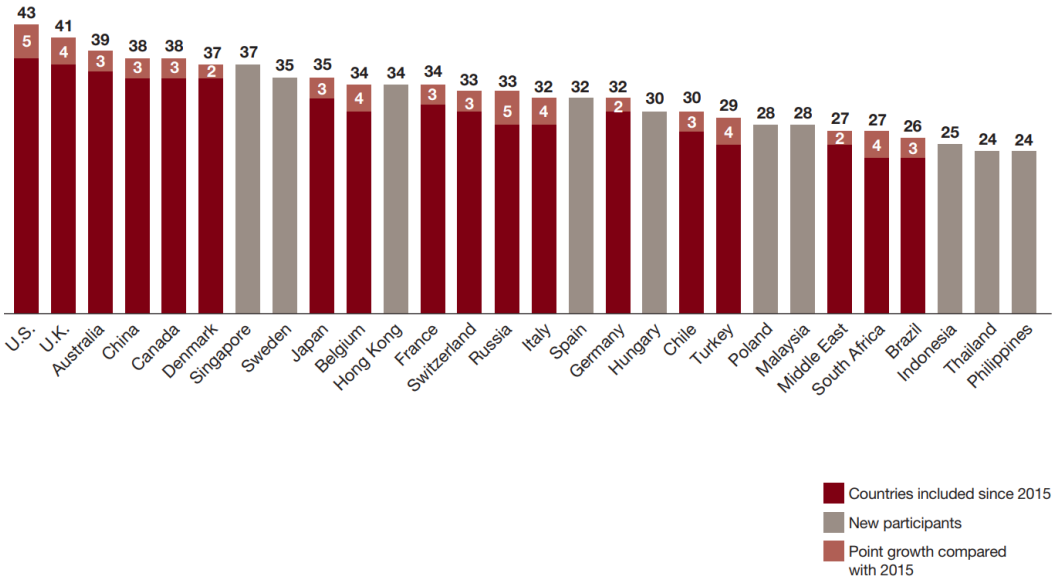


Figure 1. Omni-channel development progress by country (scale: 1-100). Source: PwC (2017).

Omni-channel retailing seeks to make it possible for customers to shop whenever, wherever, and however they wish. It also offers great opportunities for retailers to reach more customers and tap into new markets, since some barriers such as geography and distance are less dominant. In addition, studies have shown that omni-channel customers using multiple retail touchpoints spend more than customers who shop through a single channel (Sopadjieva et al., 2017). Retailers may also benefit from channel synergies when integrating operations and resources across different channels (Avery et al., 2012; Verhoef et al, 2015). For example, retailers may avoid product markdowns in physical stores by using excess store inventories for online fulfilment.

The evolving retail landscape and customer expectations have prompted retailers to adapt to the omni-channel era. The pain in this retail transformation has been felt widely. Well-known names such as Toys R Us and Maplin collapsed into administration in the UK (The Guardian, 2018a); the prestigious department store, House of Fraser, went into administration before being rescued by Sports Direct (BBC News, 2018); the well-known British multinational retailer, Marks & Spencer (M&S), has been closing a large number of its high-street stores (BBC News, 2019a). The casualties from the disruption of digitalisation and new retail models continue to grow. Additionally, last few years have witnessed various changes attempted by retailers to reposition themselves in the transformation to omni-channel. For example, the major British supermarket, Tesco, has shut its loss-making non-food website

Tesco Direct that sold general merchandise such as clothing and homeware. This move led to the closure of its online fulfilment centre in Fenny Lock for Tesco Direct orders (The Guardian, 2018b). The low cost German discounter, Lidl, is set to launch an online delivery service in the UK to offer multiple channel retailing (The Sun, 2019). This dynamic omni-channel retail landscape is exciting and calls for more attention from the research community.

1.2 Omni-channel supply chain and operational challenges

A core concept in omni-channel retailing is to offer a high degree of flexibility to customers including various options for purchase, order collection and returns. To achieve such flexibility, retailers need to integrate and coordinate supply chain activities for a multitude of sales channels, which poses a significant challenge. Ernest & Young (EY) surveyed 42 senior supply chain executives from the world's largest consumer goods and retail companies in 2015 and reported that 81% of the respondents believed that the supply chain operations in their organisations were not fit for omni-channel purpose (EY, 2015). Fulfilling and delivering the increasing volume of online orders has put high financial burdens on retailers. The shipping cost and fulfilment cost in Amazon in 2017 were $21.7 billion and $25.2 billion, respectively. The sum of these two costs amounted to 26.4% of Amazon's net sales, compared to 16.6% in 2007 (Supplychain 24/7, 2018). With razor-thin margins and expensive fulfilment and delivery costs for online demand, many retailers are struggling to provide omni-channel services profitably.

Demand forecasting and planning is also more complex in an omni-channel environment with different demand signals across all sales channels (PwC, 2016). Demand planning decisions have repercussions on inventory management, which in turn affects the possible locations available to fulfil online orders. To offer cross-channel functions such as BOPS and SFS, inventories in one stocking location are often pooled to fulfil demand from both online and offline channels (Ishfaq and Raja, 2018). This requires the sharing of inventory data across various channels with high accuracy and decisions on from where to fulfil each online order. Inventory control policies such as ordering decisions and inventory rationing decisions in a stocking location to satisfy online and offline demands are more complicated than in traditional BM retailing or in pure e-commerce retailing. Compared to the more predictable high volume offline demand, online demand is generally more sporadic with lower volumes, higher variety and higher requirements on demand fulfilment speed (e.g., same day delivery/collection services) (Metters and Walton, 2007). Therefore, fulfilment and delivery for both offline and online demands need to be considered and reconciled. For example, a retail distribution centre (DC) handling both traditional store replenishment and

online order fulfilment must consider the implications of fulfilling both streams of demand when making routing and scheduling decisions for delivery vehicles.

A further issue relates to returns handling in the omni-channel retail. Product returns have been rising across retail sectors as online customer spending continues to grow. It is estimated that returns cost UK retailers £60 billion a year, £20 billion of which is generated by unwanted online purchases (Financial Times, 2016a). By allowing online customers to return items to retail stores, it is more onerous for retailers to efficiently handle traditional store returns, online returns to stores, and online returns via Courier, Express and Parcel (CEP) services. Other supply chain challenges include capacity management in warehousing facilities, the routing of delivery vehicles to ship both online and offline demand, and supplier management when retailers outsource certain activities to 3rd parties (Agatz et al., 2008; Hübner et al., 2015).

Given the increasing popularity of the omni-channel retail strategy and the associated opportunities and challenges, it is important and timely to investigate how retail supply chains, operations and logistics management should be reshaped to embrace an omni-channel strategy, striking the right balance between cost and service.

1.3 Research objectives

Retailing is a highly competitive landscape and retailers must avoid being outpaced by more adaptable competitors. The emerging omni-channel retailing landscape has motivated our research and defined our research objectives. The main objective of this research is:

- Provide guidelines for retailers to make supply chain, operations and logistics decisions when transitioning to omni-channel through the development of decision models.

The focus in this work is on the development of quantitative decision models that can contribute to informing an omni-channel retailing strategy. First, we seek to develop and present a comprehensive conceptual map to identify all of the key supply chain and operations decisions that omni-channel retail fulfilment raises. We do this by critically reviewing a wide range of literature from both the academic and practitioner communities. Second, we develop new decision models for three critical problems highlighted in the conceptual map. We introduce each of these in the following section.

1.4 Problems and solution methodologies

Problem 1 – How and when should a retailer pick online orders in stores?

The use of BOPS services has proliferated worldwide with a continuously growing trend (Barilliance, 2019). The increasing popularity and prevalence of BOPS services is mainly attributed to their benefits for both customers and retailers, which we elaborate in chapter 3. In practice, retailers typically advertise an earliest time for customers to collect their online orders in a chosen store, if customers place their orders by the cut-off time specified. The earliest collection time promised by retailers is therefore a deadline for their order fulfilment operations, by which orders should be ready for collection to avoid disappointing customers. Store fulfilment operations for online orders are generally manual and it is well-acknowledged that they are inefficient in comparison to warehouse picking operations, which translates into high labour costs (Hübner et al., 2016a). Therefore, it is in every retailer's interest to maintain a picking rate needed only to guarantee a high service level. We develop a model to support store picking operations for fulfilling same day BOPS orders in our first research problem.

Given the advertised order cut-off time and the corresponding deadline (i.e., earliest store collection time), retailers need to decide when to start picking orders in stores and the picking rate necessary to achieve a target service level. In this problem, we consider the common practice that retailers offer a daily order cut-off time. If retailers commence picking activities earlier before the cut-off time, the processing window will be longer and a lower picking rate is required to handle a certain number of orders. However, orders arriving after the cut-off time will not be processed by the deadline in a single-wave picking system. In contrast, if retailers start processing orders later, fewer orders arriving late will be missed. However, the later that picking commences, the shorter the processing window until the designated deadline and therefore the higher the picking rate that would be needed. Retailers need guidance to make such decisions to unlock the potential of BOPS services, but this problem has not been considered in the research literature. We aim at helping retailers to decide the timing to commence picking, the minimum picking rate needed, and the number of picking waves per day to achieve a pre-defined service level, given the order cut-off time, the deadline and the demand rate. We develop analytical models to solve variants of this problem and derive Best Performance Frontiers (BPFs) to show the best combinations of the picking rate needed and the time to start picking in order to achieve a specific service level.

Problem 2 – What is the optimal online ordering window a retailer should offer for in-store collection?

Offering a one-day ordering window for BOPS services is a common practice for retailers, since it is straightforward to implement a daily order cut-off time. However, many variations exist among retailers regarding the BOPS order receipt window, which is the order

acceptance period for a promised earliest collection time. For example, the order receipt window of Tesco's same day BOPS service is shorter than one day, whilst Superdrug sets the order receipt window as one week for its free BOPS services (www.tesco.com; www.superdrug.com). At holiday times with given delivery/collection deadlines (e.g., Christmas Eve), retailers also compete on providing late cut-off time (i.e., longer order receipt window) to capture last-minute shoppers. Asos, a British online fashion and cosmetic retailer, allowed customers to order until midnight December 19 for standard click-and-collect and December 21 for next day click-and-collect in 2018. However, the company faced an onslaught of bad reviews after failing to deliver orders on time (The independent, 2018; Metro, 2018). The BOPS order receipt window affects customers' perceptions on service but also impacts back-end operations to process orders accumulated during the receipt window in a cost and time efficient manner.

The strategic decision on the order receipt window for BOPS services is the second problem we address. Retailers can fulfil BOPS orders from local stores without incurring any transportation. Alternatively, they can fulfil BOPS orders from other distribution facilities such as central warehouses before shipping these orders to the chosen stores for customer collection. Such a fulfilment strategy is known as 'Ship-To-Store' (STS) and requires a transportation window from the picking facility to the collection store (Akturk et al., 2018). The speed for order fulfilment hinges on the interaction of three time windows: the order receipt window, the order process window and the STS transportation window. Given the delivery deadline, a longer order receipt window may capture last-minute shoppers and reap more revenue. However, it results in shortened order processing window and/or STS transportation window, which requires higher processing rate and/or generates more late deliveries.

The decision on the BOPS order receipt window is a challenge at the marketing-operations interface and has not been examined previously in the literature. We present cost minimisation models to determine the optimal online order receipt window for BOPS services, the time to commence order processing and the minimum required processing rate. Our analyses include the decentralised picking scenario with zero STS transportation window (i.e., picking BOPS orders from local stores) and a more centralised scenario with a given STS transportation window. We apply our model results to explain differences observed in contemporary retail operating strategies and present frameworks to show the impact of product and retailer characteristics on service offerings for store collection.

Problem 3 – Where should online orders be fulfilled from in a retailer's store and distribution network?

In the first two problems, the facilities used for online order fulfilment are assumed as a priori knowledge. In practice, retailers in different sectors are using different fulfilment strategies to serve online customers. The strategic decision on the fulfilment location to satisfy online orders could make or break a business. The comparison between the now-defunct US online grocer Webvan and the long established grocery retailer Tesco provides powerful evidence of the online fulfilment strategy's effect on business performance. Webvan promised groceries delivery in a specific 30-minute window and invested heavily in large and highly automated fulfilment centres to ensure timely delivery (Delaney-Klinger et al., 2003). In contrast, Tesco adopted a store-based strategy and picked online orders from physical stores, which allowed Tesco to roll out its delivery services quickly without hefty upfront investment (Delaney-Klinger et al., 2003). Fulfilling and shipping groceries are challenging operationally due to their typical features, including large order size (approximately 40 items per order), very slim profit margin (typically 1% - 3%), high perishability and various temperature control requirements for storage and transportation (e.g., ambient, chilled and frozen) (Punakivi et al., 2001; Boyer et al., 2003; Hübner et al., 2016b). The high transaction cost and the insufficient order volume made it impossible for Webvan to survive and the company filed for bankruptcy in 2001 after 3 years of operation (Lunce, et al., 2006). Tesco, in contrast, now has the largest share of the online grocery market in the UK having adopted a more flexible strategy and adapted it over time (Mintel, 2019).

Given its strategic importance, we examine the decision on the fulfilment location for online orders for retailers with a store network. We consider both BOPS orders and direct home delivery (HD) orders for online demand. Typically, there are four online fulfilment options for retailers with multiple channels - fulfil from stores, from DCs, from dedicated e-commerce fulfilment centres (EFCs) and from suppliers' warehouses, a practice known as drop-shipping. We observe that retailers in different product sectors are using different combinations of these online fulfilment options. Therefore, we first develop a typology of online fulfilment network configurations for both groceries and non-food products. We then focus on the optimal configuration of the online fulfilment network for the important category of groceries, taking account of the characteristics and features of this product category.

The commonly used online fulfilment network for leading grocers comprises a series of retail stores and potentially dedicated EFC(s). With more stores used to fulfil online orders, travel distance for online shipments can be reduced and the excessive delivery costs can be curbed.

However, if fewer stores are deployed as online fulfilment facilities, inventories for online demand are stocked at a smaller number of locations and the inventory holding cost can be lowered. The fixed cost associated with upgrading and operating each store as an online fulfilment facility also decreases. The decision on the number of stores used for online fulfilment facilities is associated with complex trade-offs between various types of costs. We derive the total cost of the online groceries fulfilment problem analytically when a retailer fulfils online orders from retail stores. The decision variable in the analytical model is the centralisation level of online order picking, which defines the number of retail stores to be converted as picking facilities in order to satisfy online demand from the retailer's entire service region. Using computational experiments, we provide managerial insights on the number of picking stores needed under different business scenarios. We also demonstrate how to adapt the model to the case in which online orders are fulfilled from DCs and/or EFCs.

1.5 Organisation of the thesis

The thesis is organised as follows: Chapter 2 discusses existing literature on supply chain and operations related decision-makings for retailers transforming to omni-channel. A conceptual map encapsulating critical supply chain and operations decisions for omni-channel fulfilment is presented by analysing and synthesizing relevant literature. Chapter 3 studies the first problem of fulfilling BOPS orders in stores, given an advertised cut-off time and the corresponding deadline. Relationships between the time to start picking, the picking rate needed, the number of picking waves per cycle, and the achievable service level are derived to aid decision-making. Chapter 4 examines the second problem and develops cost minimisation models to optimise jointly the order receipt window for BOPS service offerings, the time to commence order processing and the processing rate needed. Chapter 5 investigates the third problem and derives analytical expression for the total cost to determine the number of retail facilities (e.g., stores and DCs) to be converted as online fulfilment facilities. For each model developed, we discuss in Chapters 3, 4 and 5 the relevance and managerial implications derived from the model results and the analyses respectively, thereby having no specific discussion chapter. Finally, Chapter 6 presents conclusions summarising the major contributions in the omni-channel research of the conceptual map and the models for the three problems. Limitations of the thesis are also discussed with suggestions for future work.

Chapter 2

Literature review

Though supply chain and operations management is critical for retailers in transition to omni-channel, the literature addressing the related issues is limited and reviews specifically focussing on operational issues and decision making in this field are scarce. To the best of our knowledge, there are only four review papers in the omni-channel domain, out of which only two specifically touch on supply chain management (SCM) issues. Mirsch et al. (2016) present a review of literature associated with the definition, strategy conceptualisation and requirements, benefits and obstacles of omni-channel management and the transformation trajectory towards it. Cummins et al. (2016) provide a review on omni-channel in the context of sales and sales management. Galipoglu et al. (2018) discuss existing studies focusing on omni-channel retailing and structure them based on dimensions such as research methodology, research themes, and the geographical and sectoral focus. Their review also debates the intellectual foundation of omni-channel retailing research. As valuable as these reviews are, they cast only a limited light on the key supply chain decisions to embrace omni-channel and the necessary guidance to make these decisions optimally. The most relevant review paper is over 12 years old. Agatz et al. (2008) discusses key supply chain issues and related quantitative models particularly for online demand fulfilment in the emerging multi-channel retail landscape of that time. Our contribution is distinctive from Agatz et al. (2008) in several ways. First, we consider the contemporary retailing landscape that has emerged in the last decade and include the latest literature after 2008 that focuses particularly on omni-channel retailing. Second, we consider the impact of critical contextual factors including the characteristics of retail, market, product and supplier that impact on omni-channel decisions. Third, we develop a conceptual framework to map out critical decisions for retailers to execute the omni-channel strategy from both front-end customer experiences and back-end operational and SCM perspectives.

Considering the increasing prevalence of omni-channel and the importance of SCM, this chapter aims at providing an overview of operational and SCM related issues arising in the transformation to omni-channel retailing. Specifically, we answer the following research questions:

1. What are the key decisions in transition to omni-channel retailing from both customer service and SCM perspectives?
2. How will contextual factors (product, market, retail and supplier characteristics) affect these decisions?

3. What are the contributions of quantitative research models relevant to omni-channel decision-making?
4. What are the research agendas for the future?

The investigation of the questions above will also be used to justify the choice of the three main research problems addressed in the thesis.

2.1 Scope and structure

In this section we clarify the scope of our review and describe the structure of our discussion throughout this chapter.

2.1.1 Scope

Our prime objective is to outline the current state of research that uses quantitative methods to tackcle challenges in supply chain and operations management to embrace omni-channel retailing. The scope of our study has two boundaries. First, we focus only on store-based retailers who integrate both online and offline businesses, while pure online retailers or manufacturers who sell directly to customers via online channels are beyond the scope of this study as they do not adopt omni-channel strategies. Second, we are interested in presenting an overview of relevant quantitative decision models for retailers transitioning to omni-channel. Hence, we mainly discuss the contribution of studies using quantitative methods for operational and supply chain decision-making.

2.1.2 Structure

After comparing and categorising three different channel strategies, namely multi-, cross- and omni-channel retailing, we identify four planning areas for omni-channel supply chain and operations management. The four planning areas are inventory management, fulfilment and warehousing, last mile delivery and returns management. For each planning area, we first highlight the emerging issues and underline the critical decisions in the context of omni-channel retailing. This is followed by a discussion on the contributions made by extant studies, which helps identify research gaps that deserve further investigation. Note that in separating the discussion on each planning area we do not imply that decisions in each planning area are, or should be, isolated from each other. On the contrary, we emphasise the importance of managing the back-end supply chain for effective omni-channel services as an integrated system.

Based on the discussion in each planning area, we present the conceptual map of critical supply chain and operations decisions for retailers in transition to omni-channel. The conceptual map depicts the interface between front-end customer services offered by retailers

and back-end operations to deliver such services. The interface is moderated by contextual factors such as retailer characteristics (e.g., the number and location of retail facilities) and market characteristics (e.g., demand distribution).

2.2 Categorisation of multiple channel strategies

Multiple channels include both traditional store-based channel and online channels such as online website, social media and mobile apps (Salmani et al., 2018). To proceed with the literature review, it is necessary to clarify key terminologies regarding retail channel strategies, as the three terminologies (multi-channel, cross-channel and omni-channel) are often used interchangeably. Beck and Rygl (2015) present a highly cited taxonomy of categorisation for multiple channel retailing based on two dimensions, namely the party (retailer or customer) who controls or triggers channel interaction and the scope of channels considered (partial or full range of existing sales channels). Their taxonomy describes eight different channel strategies: two types of multi-channel, four types of cross-channel, and two types of omni-channel retailing (see Figure 2). For multi-channel retailing, though the retailer operates its business in more than one channel, the customer cannot trigger interaction (multi-channel I) nor the retailer control integration between channels (multi-channel II). Cross-channel III and IV enable the customer to trigger partial integration and the retailer to control partial integration of at least two channels, respectively. Cross-channel V and VI offer the possibility for the customer to trigger full integration and the retailer to control full integration of at least two but not all channels, respectively. Omni-channel VII (VIII) allows the customer (retailer) to trigger (control) full integration of all the widespread retailing channels, such as physical stores, mobile shop and online shop.

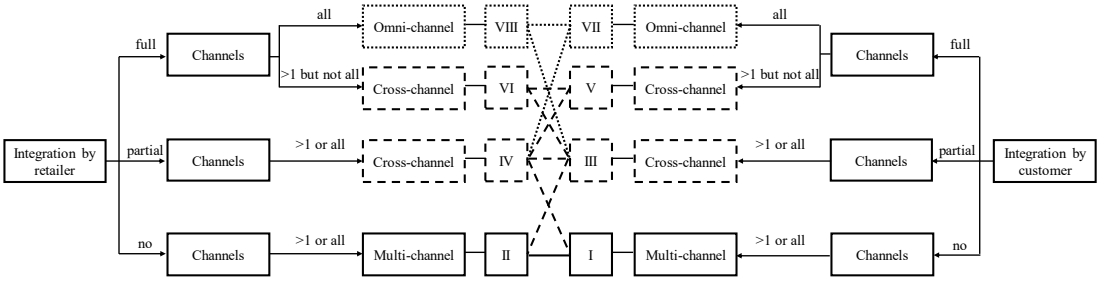


Figure 2.Categorisation tree in multi-, cross- and omni-channel retailing for retailers and retailing, from Beck and Rygl (2015)

With a 'customer-centric' mentality, today's retailers only qualify as cross- or omni-channel when the corresponding channel integration synergy effects are reflected in the improvement of customer services. The synergetic management of mutliple channels focuses on delivering a superior customer experience across channels (Verhoef et al., 2015). Therefore, full integration controlled by retailers without account of the customer experience may not be

fully justified to be omni-channel, as in Beck and Rygl (2015). For example, retailers may achieve full functional integration in the back-end operations to fulfil online orders flexibly from any inventory stocking location (i.e., full integration controlled by the retailer). However, the front-end service offerings may not mirror such synergy effects. Short delivery times may not be offered in front-end customer services even though it may be viable with the SFS capabilities, or wider assortments may not be provided despite the fact that online orders could be routed and fulfilled by suppliers who have a wider product range. In this case, the multiple channel strategy cannot be deemed as omni-channel from a customer perspective, though full integration can be controlled by the retailer. As suggested in Beck and Rygl (2015), the key difference between multi-, cross- and omni-channel retailing lies in the channel integration level. However, the classification in Beck and Rygl (2015) does not encapsulate and manifest the channel integration from both front-end customer experience and back-end SCM perspective. To this end, we explain the channel integration level and the difference between multi-, cross- and omni-channel considering both downstream customer service offerings and upstream supply chain and operations management.

A specialist report provided by DHL (2016) presents the different maturity levels of omni-channel capabilities using two dimensions: front-end customer promise and back-end supply chain capabilities. The omni-channel capability maturity levels for retailers range from level 1 (e-commerce beginner) to level 4 (omni-channel champion), based on the front-end customer promise (assortment, delivery options, returns options) and back-end supply chain capabilities (inventory management, fulfilment & warehousing, final mile delivery and return handling). However, the purpose of the report is not to precisely define and differentiate the three multiple channel strategies and the focus of the two dimensions used are diferent from what we propose in this study. Built on Beck and Rygl (2015) and DHL (2016), we classify and delineate the three commonly used multiple channel strategies in Table 1.

Table 1 Differentiation between multi-, cross- and omni-channel retailing.

		Multi-channel	**Cross-channel**	**Omni-channel**
Front-end customer experience	**Integrated transaction**	Customers cannot switch from one channel to another during transaction	Customers can switch between certain but not all channels	Customers can switch between all sales channels
	Integrated product information	No access to product pricing and availability across channels	Access to product pricing and avaibility between some but not all channels	Full access to product pricing and availability across all channels

	Integrated promotion	Separately promotion management for each channel and coupons cannot be redeemed across channels	Partially integrated promotion management for channels and coupons can be redeemed between certain channels	Fully integrated promotion management across channels and coupons can be redeemed across all channels
	Integrated delivery	Only HD option available for online orders	Several delivery options including click & collect for online orders	Flexible delivery/collection options for online orders
	Integrated reverse logistics	Online purchases can only be returned via carrier, express and parcel provider (CEP)	In-store returns are allowed for online orders, together with CEP option	Returned products can be accepted flexibly across channels, regardless of the original purchasing channel
	Overall shopping experience	Distinctive customer experience in different channels and no channel switching allowed	Some identical experience in different channels and switching between certain channels are possible, but still not seamless	Channel-agnostic and seamless experience as customers can shop anywhere, anyhow and anytime
Back-end SCM	**Data management**	No customer, pricing and inventory data is shared across channels	Customer, pricing and inventory data are shared between some but not all channels	Customer, pricing and inventory data are shared across all channels
	Assortment planning	Independent assortment planning across channels	Some collaboration of assortment planning across channels	Coordinated assortment optimisation across all channels
	Pricing	Independent pricing decisions across channels	Certain coordination across channels	Synergistically optimise pricing across all channels
	Inventory management	Separate inventories for demand from different channels	Partially integrated inventories across channels	Fully integrated inventories across channels, i.e, single stock for demand from all channels
	Order Fulfilment	Fulfil from dedicated depot for online orders	Fulfil online orders from some retail stores and/or DCs	Flexible fulfilment within the retail network, including suppliers, DCs, stores and online depots
	Last mile delivery	Separate online delivery and store replenishment	Certain collaboration between online delivery and store replenishment	Coordinated online delivery and store replenishment, e.g., shared delivery trucks
	Returns management	Returns from different channels are received and handled separately	Certain collaboration in receiving and handling returns across channels	Optimising returns handling across all channels

As seen from Table 1, multi-channel and cross-channel retailing are essentially different stages for retailers with multiple channels in their evolution to omni-channel (Hübner et al., 2015). At the beginning of expanding to online businesses, traditional BM retailers often keep online and offline channels separate without any interaction, which signifies the stage of multi-channel. Considering the requirements of customers on more flexible service offerings (e.g., delivery options and synced data across channels) and the promise of higher back-end operational efficiency (e.g., single inventory stocks for demand across different channels), retailers may be motivated to have some integration and coordination between online and offline channels. This is the intermediate stage of cross-channel retailing. Ultimately, retailers can achieve full integration across all channels for back-end operational

efficiency and evolve to be omni-channel so that customers can research and purchase anywhere, anytime and anyhow.

Figure 3 illustrates the trajectory of retailers from multi-channel through cross-channel to ultimately omni-channel offerings using the dimensions of front-end offerings and back-end operations, based on the work of DHL (2016). The size of the circle represents the number of retailers currently at that stage. Most retailers have achieved certain integration across channels and provided services such as BOPS and in-store returns for online purchases (i.e., cross-channel stage). The complete separation (i.e., multi-channel stage) or full integration (omni-channel stage) of multiple channels are comparatively less common in the retail industry. Although we do not investigate the issue of the best channel strategy, it is worth noting that there has been considerable debate on channel cannibalisation and synergy effects as each channel has potential advantages and disadvantages from both market and operations perspectives (Kollmann et al., 2012). The marketing literature has many studies on the nature of channel effects. Some study empirically the cannibalisation effect on sales between the traditional, online and catalogue channels (Deleersnyder et al., 2002; Weltevreden, 2007; Avery et al., 2012; Pauwels and Neslin, 2015). However, many studies have demonstrated that channel conflicts can be mitigated and channel synergy effects can be exploited with appropriate channel coordination and integration (Gulati and Garino, 2000; Lee and Whang, 2001; Avery et al., 2012; Herhausen et al., 2015; Saghiri et al., 2017; Verhoef et al., 2015). Various strategies and mechanisms are proposed to achieve cross-channel synergies through customer management, integrated marketing, pricing, assortment planning, investment allocation, profit sharing contract, and the choice of the ordering time (Avery et al., 2012; Berry et al., 2010; Choi et al., 2019; Li et al., 2015b; Neslin et al., 2006; Nyquist, 2016; Payne et al., 2017; Verhoef et al., 2007; Van Baal, 2014; Zhang et al., 2010). The benefits of adding another channel and offering cross-channel functionalities such as BOPS to firm performance (e.g., sales growth and customer base) have been explored and validated extensively (Xia and Zhang, 2010; Cao, 2014; Gallino and Moreno, 2014; Fornari et al., 2016; Chen and Chen, 2017; Gao and Su, 2017; Akturk et al., 2018).

Both Table 1 and Figure 3 illustrate a progression from multi-channel to cross-channel and omni-channel with respect to the degree of flexibility for downstream customer service and upstream operational efficiencies.

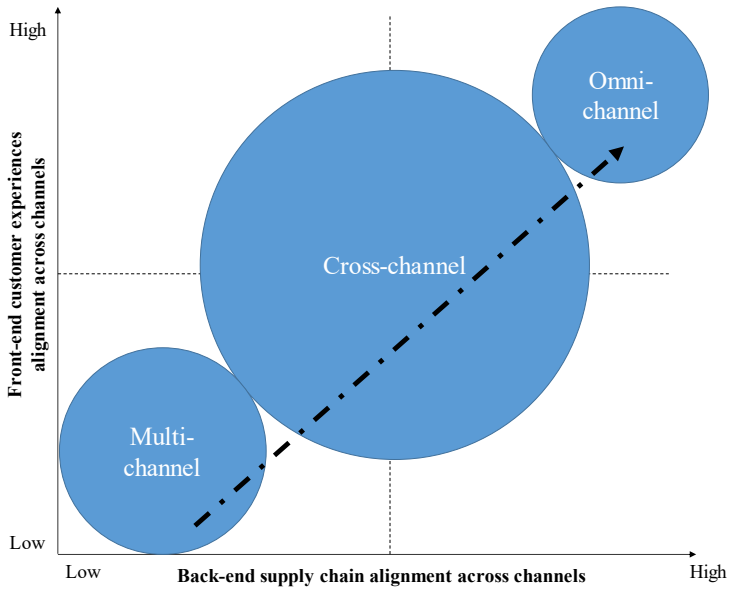


Figure 3. Trajectory of retailers moving from multi-channel to omni-channel, adapted from DHL (2016).

2.3 Planning areas for omni-channel SCM

Based on both academic studies and business literature, we depict the typical flowchart of online fulfilment process in Figure 4. After capturing orders from various channels such as physical stores and mobile apps, retailers need to decide how to allocate orders across different stocking locations in their systems (e.g., physical stores and DCs) and how to fulfil orders based on the inventory locations and the delivery/collection services chosen by customers. Last mile delivery is entailed from a fulfilment location to the customer's designated destination (e.g., the customer's home or a collection store) and unwanted parcels could be returned by customers back to the retailer system. Figure 4 highlights four specific planning areas for omnichannel retail operations, namely, (1) inventory management, (2) warehousing and fulfilment, (3) last mile delivery, and (4) returns management. These four key areas are also discussed in Saghiri et al. (2017) and DHL (2016) to exemplify omni-channel integration. The following sections overview the decisions that have been investigated as well as decisions that need to be examined in each planning area. We then summarise our results and present the conceptual map highlighting critical SCM decisions for retailers in transit to omni-channel.

Our focus in this study is on operations and supply chain management. Product pricing is not considered explicitly in this thesis. Harsha et al. (2019) propose price optimisation models for different sales channels (i.e., online or retail stores) and for stores in different locations, taking account of demand sensitivity to price and the competitive effect. Except for Harsha et al. (2019), pricing decisions are often jointly optimised with other decisions such as

inventory control and order fulfilment to balance demand and supply. Therefore, we discuss relevant models for pricing decisions whenever they are simultaneously considered with other decisions in the four planning areas.

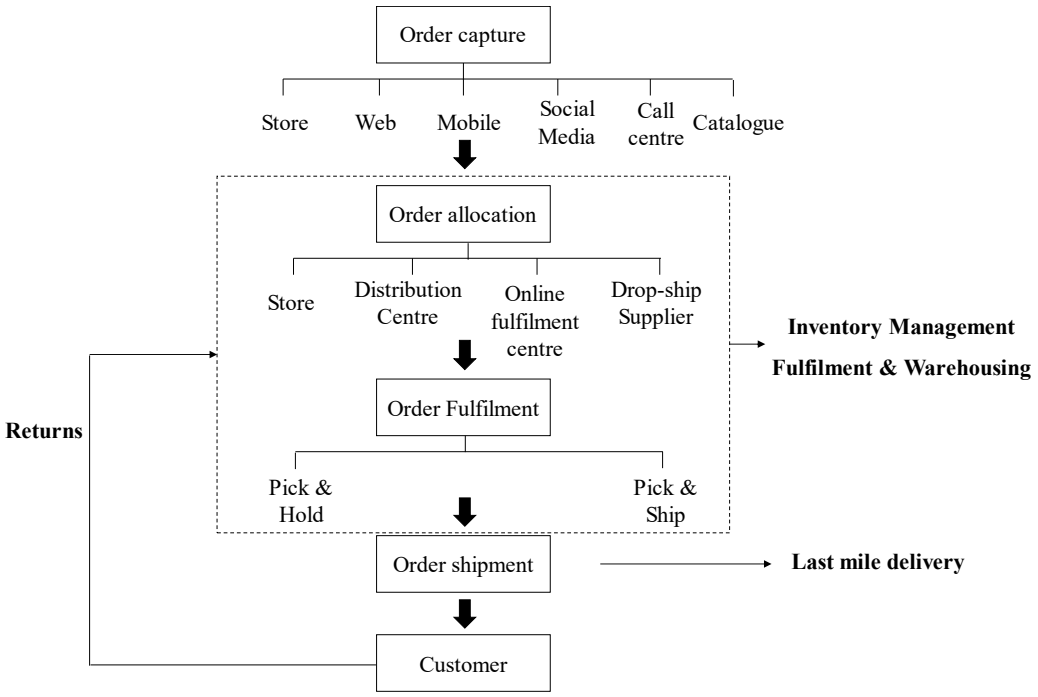


Figure 4. Omni-channel retailing online order fulfilment processes

2.4 Inventory management

Omni-channel retailing is distinguished from single channel retailing by the potential to aggregate online orders and fulfil them flexibly from many possible locations (Chopra and Meindl, 2016), including self-owned inventory facilities such as retail stores, fulfilment centres, DCs, and also suppliers' warehouses. The latter practice is often known as 'drop-shipping' when retailers simply forward online orders to their upper-echelon suppliers who then fulfil directly from their own inventories (Zhao et al., 2016). As a result, the interdependence of different fulfilment locations (within the network of a retailer and potentially its suppliers) adds further complexity to inventory management. The location and amount of inventories within the supply chain network directly affect the inventory and warehouse investment, delivery planning efficiency, and customer service level with respect to product availability. Therefore, it is vital for retailers to consider and decide where to put inventories and how much inventory each stocking location should carry to fulfil demand from across all their sales channels.

2.4.1 Inventory ownership

The time delay from order placement to order receipt by customers makes it viable for retailers to adopt the drop-shipping option and therefore maintain low or zero inventories for

online demand. In this way, retailers avoid investment in e-fulfilment capabilities and eliminate inventory related risks (Khouja and Stylianou, 2009; Randall et al., 2006). Therefore, retailers can potentially offer a wider range of assortments from the manufacturer's product portfolio without physically bearing the stocking costs (Brynjolfsson et al., 2011). Additionally, drop-shipping involves fewer intermediate operations through the retailer's systems, which may lower the end-to-end cost and alleviate the product deterioration problem (DHL, 2016). However, drop-shipping may result in fragmented deliveries and longer delivery time, and retailers have less control over the customer experience (Ayanso et al., 2006; Chopra and Meindl, 2016).

Given the upsides and downsides of drop-shipping, an evident question is: whether and how should retailers adopt the drop-shipping option? If the drop-shipping option is not considered, retailers rely completely on their in-house inventories for online fulfilment (see Figure 5(a)). The other extreme is to route all online orders for drop-shipping (see Figure 5(b)). Alternatively, retailers could adopt a mix of in-house inventory and drop-shipping for online fulfilment (see Figure 5(c)). Several studies have empirically investigated the contextual operating conditions under which the drop-shipping option is more preferable for retailers. For example, Randall et al. (2006) investigate empirically the linkage of the decision on inventory ownership to nine factors including product characteristics such as product size/weight, retailer characteristics such as firm age and mean revenues, and supplier characteristics such as the ratio of retailers to the supplier. Their results show that drop-shipping is more beneficial for retailers when product variety is high, when demand uncertainty is high, when the ratio of retailers to the supplier is high, when the relative gross margin is low, when the age of the retail firm is young and when the product size/weight is high.

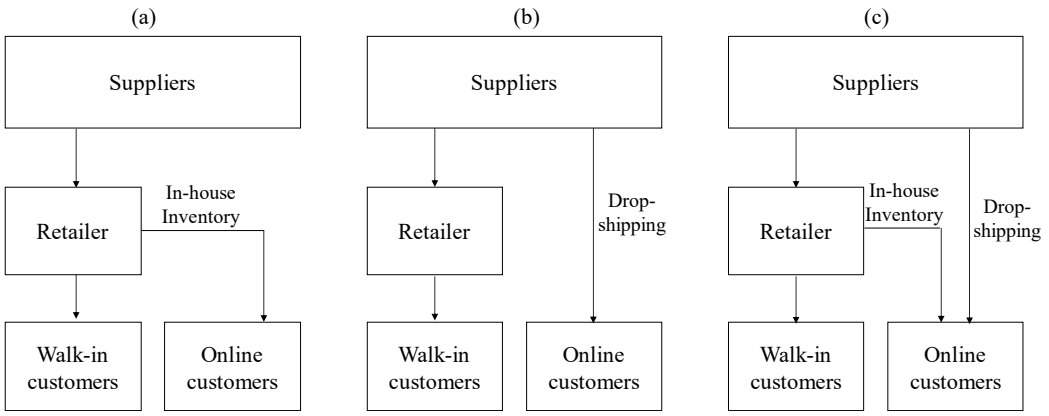


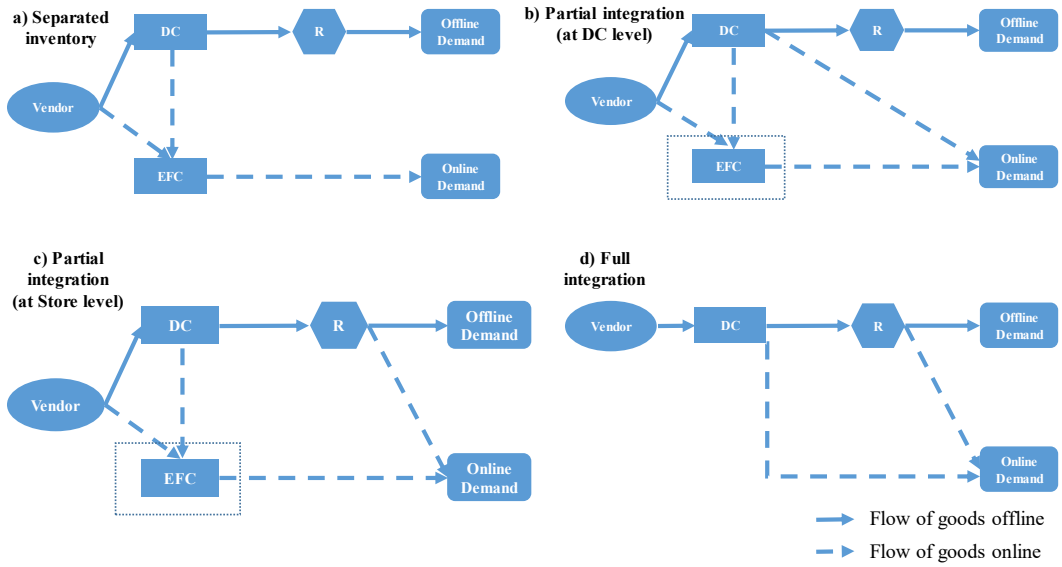
Figure 5. Inventory ownership decisions for omni-channel retailers.

To reap the benefits and avoid the drawbacks of drop-shipping, researchers have studied the optimal mix of in-house inventory and drop-shipping for retailers with multiple sales channels. Khouja (2001) was one of the first to evaluate the mix of in-house inventory and drop-shipping for online fulfilment. The author presents a single period newsvendor model to identify the optimal order quantity for in-house inventory for stochasic online demand, with the assumption of unlimited fulfilment capacity from drop-shipping. In his research setting, online fulfilment from in-house inventories incurs a shorter lead time and therefore a higher price for the service, compared to the drop-shipping option. The results demonstrate that if all online customers are willing to switch from in-house fulfilment to drop-shipping, it is optimal for retailers to carry zero inventory and rely on drop-shipping completely. For products with high value and therefore high unit purchasing cost, drop-shipping is also more preferable. Focusing on internet book retailing, Bailey and Rabinovich (2005) compared two inventory location strategies: speculation (fulfil from in-stock inventory) and postponement (drop-shipping). They develop analytical models that yield closed-form expressions for the amount of online demand fulfilled from in-store inventories and drop-shipped. Their results show that retailers will rely more heavily on both in-stock inventories and drop-shipping (i.e., more reliance on both forms of inventory management strategies) when product popularity and retailers' market share increase.

Netessine and Rudi (2006) propose noncooperative game-theoretic models to compare three types of inventory ownership for online fulfilment: complete reliance on in-house inventories (Model T in their paper), complete reliance on drop-shipping (Model D), in-house inventories as primary source and drop-shipping as a back-up option (Model C). They highlight that the stocking decisions for retailers' in-house inventories and for the wholesaler who handles drop-shipping hinge on the trade-off between profit margin and inventory risk, which depends mainly on the drop-shipping markup and the transportation cost differential between in-house fulfilment and drop-shipping. Their results suggest that both retailers and the wholesaler prefer the utilisation of drop-shipping over complete reliance on in-house inventory fulfilment for a wide range of problem parameters. Chiang and Feng (2010) develop a game-theoretic economic order quantity (EOQ) model to compare the profitability and efficiency of two channel structures for online fulfilment: a traditional channel using in-house inventories and a drop-shipping channel where the retailer maintains no inventory. Assuming that demand rate is price dependent, they jointly optimise the pricing and lot-sizing decisions for these two channel structures. They identify scenarios under which the retailer is better off to rely on drop-shipping for online fulfilment without investing in any inventory holdings for online demand.

2.4.2 Inventory integration level

If some or all online orders are fulfilled from retailers' in-house inventories, the next issue is to decide whether to separate or integrate inventories for demand from different channels. In addition to separated inventory, three channel integration levels can be identified for inventory management in practice (see Figure 6).



Note: Distribution Centre (DC), E-commerce Fulfilment Centre (EFC), Retail Stores (R)

Figure 6. Inventory integration strategies for retailers with multiple channels.

Separated inventory. As shown in Figure 6(a), retailers keep the inventories for demand from the traditional BM channel and online channels separated. This approach has been used by retailers in the initial stage of transition to omni-channel, since they wish not to be distracted from their long-standing and fine-tuned traditional store replenishment inventory systems. Additionally, if retailers use the online channels to target different customer groups with differentiated product offerings, it may be rational to keep separate inventory locations for online demand.

Note that the EFC in Figure 6(a) receives products from suppliers and dedicates its stocks to online demand only. A further variant of the EFC is the so-called dark store that receives replenishments from DCs and only serves online customers, which explains the product flow from DC to EFC (Grant et al., 2014). However, separated inventories could adversely impact both the operating cost and the service level for retailers. For the back-end operations, no integration means duplicated inventories and therefore higher costs (DHL, 2016). For the front-end customer interface, a separated inventory system may prevent retailers from providing a seamless and channel–agnostic shopping experience. For example, customers

cannot choose a SFS option to get their orders quickly since store inventories cannot be used to fulfil online orders.

Partially integrated inventory. In this case, stocks are partially integrated at either the DC or store level. Depending on the capacity of those integrated facilities and the online demand volume, retailers might still need dedicated EFC(s) (the dashed line box may or may not exist in Figures 6(b) and 6(c)). For example, Tesco used a store-based fulfilment strategy at the beginning of its expansion to online businesses (Fernie and Sparks, 2014). Inventories are pooled at the store level so that retail stores handle both online and offline demand without deploying any dedicated facility for online channels. As online demand grows, the existing store capacities tend to be inadequate and therefore dedicated EFCs are invested and put in use (Wulfraat, 2014). Figure 6(c) represents the inventory integration strategy adopted by Tesco. In contrast, retailers such as British department store John Lewis have integrated inventories at the DC level (see Figure 6(b)), so that their DCs fulfil online demand while also handling traditional store replenishment (eDelivery, 2015).

Fully integrated inventory. As illustrated in Figure 6(d), inventories are completely integrated at any location at any level within the retailer' distribution network. Therefore, stocks in both stores and DCs are pooled for demand from online and offline channels. One example is the British electrical and white goods retailer Currys PC World who deploy store inventories for SFS and BOPS services (Dixons Carphone, 2017). Also, DCs can be used to fulfil online orders for STS and/or direct HD services.

Related literature

Separated VS integrated inventory. A number of earlier studies have investigated and demonstrated the cost benefits of achieving a certain level of inventory integration over separated inventories across channels. Seifert et al. (2006) compare the ordering quantities for two supply chain strategies. One is a dedicated supply chain where online orders are fulfilled from a specific online depot. The other one is an integrated supply chain where retail stores continue to serve all walk-in customers but use excess inventories for online orders. They develop mathematical models to compare the ordering quantities for these two supply chains and quantify the cost reduction attainable from the integrated supply chain. Their numerical results show that the cost savings from an integrated supply chain increase when the number of retail stores is high, when the transshipment-penalty cost is low, when demand variability is high, and when the replenishment lead time is high.

Hovelaque et al. (2007) compare three different supply chain models for bricks-and-clicks (BC) retailers, namely store-picking, dedicated warehouse-picking and drop-shipping. The first model integrates online and offline inventories at the store level, whilst the latter two separate online and offline inventories. The authors develop a single period newsvendor model in a supply chain with only a single product, single supplier and single retailer. With the assumption of normally distributed demand for both online and offline channels, they compute the optimal stocking quantities for each model and obtain the corresponding profit using Monte-Carlo simulation. Their numerical results suggest that the store-picking model outperforms the dedicated warehouse-picking model regardless of the online market size, which illustrates the inventory pooling benefits. They also found a threshold with respect to the demand uncertainty for the choice between the store-picking model (optimal below the threshold) and the drop-shipping option (optimal above the threshold), in order to maximise the profit of the entire supply chain. Bendoly et al. (2007) develop a mathematical model to determine the optimal position in the spectrum of separate inventory and integrated inventory (i.e., retail stores fulfil all online orders) in a two-echelon inventory system. The level of inventory integration is captured by the fraction of online demand fulfilled via the dedicated EFC. Their results show that there exists a threshold for the ratio of online demand to total demand, below which separated inventory is preferred, and above which integrated inventory is more desirable.

Alishah et al. (2015) investigate inventory management for an omni-channel retailer selling a single product from one online store and one offline store. The retailer could adopt a NP (No Pooling) fulfilment structure where online demand is fulfilled from dedicated inventories. Alternatively, the retailer could pool inventories at the offline store for both online and offline demand, which is called the P (pool) structure. They develop a newsvendor model to decide the inventory stocking level and profit for both NP and P structures. Their analyses show that the optimal structure for online fulfilment depends on the cross-channel fulfilment cost and the mix of demand between online and offline channels. In the same spirit, Govindarajan et al. (2017) compare three fulfilment structures, namely no integration (separate inventories for online and offline channels), partial integration where online customers can only be served by the nearest stores, and full integration where any stores can be used as a backup source for online demand if needed. They develop analytical models to obtain the optimal ordering quantities for the three fulfilment structures in a two-store, single-period setting first, before extending it to a generalised multi-store setting. The authors also investigate the effect of channel correlation and location correlation, apart from other factors such as online demand volume. Channel correlation reflects customers' channel switching behaviour between online and offline and therefore captures the channel

cannibalisation effect. Their results show that the inventory pooling benefits decrease when the channel correlation increases.

The abovementioned papers have contributed to the understanding of decision-making for the right inventory integration level for retailers with both online and offline channels. However, it is worth pointing out that they focus only on the partial integration at the store level and transshipment between stores for online demand fulfilment is often not considered. From omni-channel retailing, it is important to consider the full integration case when any location within retailers' network, including stores and DCs, can handle demand from different channels. Further, in the case of inventory integration, the extant research considers using the stocks previously reserved for offline demand to fulfil online demand, which is known as 'Online-to-Offline (O2O). However, for high maturity omni-channel retailers, stocks should be able to be used more flexibly. Unsatisfied walk-in customers can be encouraged to switch to online channels and their demand satisfied from another facility with stocks, be it at the store level or at the DC level, which is fast becoming common practice in many traditional large scale retailers.

2.4.3 Inventory control policies

Inventory control problems for traditional retailing has inspired an extraordinarily rich and extensive body of literature. Related inventory management mathematical models provide formulations that vary in complexity from a single item, single echelon system and deterministic models to multiple items, multi-echelon system and stochastic models. Readers are referred to review papers such as Bakker et al. (2012), Clark (1972) and Minner (2003) for inventory models in the traditional retail settings. The implementation of drop-shipping poses the necessity for retailers to revise their inventory control policies, as it is viable for them to use their suppliers' inventories. Inventory integration for demand across multiple channels also poses new challenges for stocking decisions. The matter becomes more complicated when the returned products flow into the system and become resalable inventories, a trend that cannot be neglected due to the high return rates from online channels. Therefore, retailers moving towards omni-channel face several problems. What is the optimal order quantities for the in-house inventories taking account of the drop-shipping option? How does the incorporation of returned products affect the ordering decisions? What is the inventory rationing policy for demand from different channels?

Khouja and Stylianou (2009) present two continuous-review order quantity reorder-point (Q, R) inventory models, in which retailers can utilise the drop-shipping option for online fulfilment if a shortage occurs during the lead time. The shortage can be backordered in the

first model whilst sales are lost in the second model. They provide closed-form expressions for the optimal order quantity Q and reorder point R. Their results indicate that the use of drop-shipping will decrease both optimal Q and R, especially when the ratio of unit drop-shipping penalty cost to the backorder cost is small, when the unit holding cost is high and when the lead time is long. In such cases, it could even be optimal to have a zero reorder point. Ma et al. (2017) incorporate the option of drop-shipping and reselling returned products, with the objective to maximise the expected profit. They consider a newsvendor who uses store inventories for walk-in customers, while online demand can be met by store inventories or by suppliers via drop-shipping. Their results demonstrate that the expected profit can be 14.4% higher by considering resalable returns and the optimal store inventory level can be 31.2% lower by allowing the drop-shipping option for online demand.

He et al. (2020) examine the pricing and stocking decisions before and after the introduction of the 'Buy-Online-and-Return-in-Store' (BORS) option for omni-channel retailers. The consideration of BORS captures the impact of resalable returns on inventory control and pricing policies. They divide customers into two types: customers with high value (accept the product) and customers with low value (return the product). They also compute the utility of consumer purchase from new product via online, new product via stores, returned product via online and returned product via stores respectively. Game theoretic models are developed to jointly optimise the product price and stocking quantities for online and store channels with and without the BORS option. Their results show that the introduction of BORS increases the optimal order quantity for online channels and decreases the stocking level in stores, whilst the price for new products in both channels remain unchanged. Naturally, products returned to stores can be viewed as replenishment for store inventories and then store order quantities can be reduced. Therefore, when the store demand is sufficiently low, the store order quantity could be zero. Their work also highlights the impact of the price of returned products (i.e., salvage value) on the ordering quantity for online and store channels. Specifically, they show that when returned products are sold at a higher price, the order quantity increases for online channel and decreases for stores. Clearly, they consider separate inventory control for different channels as stocks in one channel cannot be used for another.

Zhang et al. (2018) compare the optimal pricing, the inventory stocking quantity and the expected profit for the online-only strategy and the for omni-channel strategy, with account for possible consumer returns and order cancellation. Compared to the online-only strategy, the omni-channel retail strategy increases the total demand, enables the offering of both BOPS and HD services, and yields cross-selling benefits. For both channel strategies, the authors develop newsvendor models to jointly optimise the retail price and stocking quantity

decisions in order to maximise the total profit. Their sensitivity analyses describe the impact of problem parameters, such as the cross-selling benefit, BOPS demand fraction and return shipping fee, on the pricing and stocking decisions. Such analyses provide insights on the conditions under which the value of moving from the online-only to the omnichannel strategy is significant. The results also suggest that the omni-channel strategy decreases the consumer return rate and increases the cancellation rate compared to the online-only strategy.

Considering inventory-level-dependent demand, Li et al. (2015a) examine the inventory replenishment decisions for both the physical store and the online store operated by one vendor. Dual channels in their model compete through inventory level and unmet demand is lost for the physical store and backlogged for the online store. Neither demand nor inventories can be transferred between channels (i.e., no cross-channel fulfilment or inventory transshipment). Demand in one channel increases linearly with its inventory level and decreases linearly with the inventory level of the other channel. The authors propose a multi-period stochastic dynamic programming model and show that under mild conditions, the myopic inventory policy is optimal for the infinite horizon problem. Their model results and analysis reinforce their suggestion that demand dependency on inventory levels cannot be neglected.

Compared to the extensive research on the single channel retailing (traditional or pure online), research on the inventory replenishment planning for retailers with multiple channels is relatively limited. One reason could be that the well-known inventory control policies for a single channel can be adjusted to the context of multiple channels. Schneider and Klabjan (2013) provide conditions for known inventory control policies (e.g., the base stock policy and (s, S) policy) to be optimal when there are two sales channels. Specifically, they examine the optimal ordering policies for a typical multi-channel retail scenario: periodic review, single-item lost sales model with zero lead time, fixed set up costs and two demand classes from two sales channels without inventory rationing. They focus on the scenario when product prices in two channels are different and analyse the ordering policies for three different demand arrival processes (e.g., Poisson distribution or no particular distribution). Xu and Cao (2019) develop a finite horizon, periodic review inventory model to dynamically optimise the inventory ordering and allocation policy for a franchisee retail store handling its own in-store demand and online demand from its franchisor. Assuming zero replenishment lead time and zero fixed ordering cost, the authors formulate the inventory control and rationing problem as a Markov decision process (MDP) with the objective to maximise the expected total discounted profit during the time horizon. Once the retail store decides the percentage of inventories reserved for online orders, inventories allocated to one

channel cannot be utilised to fulfil demand from the other channel. Their analyses prove a base-stock inventory policy to be optimal. They also show that the optimal base-stock level is independent of the initial inventory at hand and the optimal stocking quantity reserved for offline demand is non-decreasing on the base-stock level.

Xu et al. (2017) present dynamic lot-sizing models with incorporation of advance demand information (ADI) for retailers with online channels (i.e., applies to both pure online retailers and BC retailers). Since there is a natural gap between the time of order placement and the time for deliveries, retailers obtain ADI for the time-varying demand. Note that they classify online demand into BOPS demand and HD demand. Stores only handle BOPS demand while HD demand is processed in dedicated online depots. They formulate the inventory replenishment problem as a mixed-integer linear programming (MILP) model and design a dynamic programming algorithm with polynomial running time to solve the model. Numerical experiments are carried out with real data from several online retailers. Their results report that the proposed model could reduce the total inventory cost by 2.3% on average. In the same vein, Alawneh and Zhang (2018) propose a multi-item inventory model for an integrated warehouse with space constraint, demand and lead time uncertainty. The integrated warehouse handles both online order fulfilment and traditional store replenishment. They provide a solution algorithm when a warehouse space constraint exists and derive closed-form solutions for the optimal reorder point and the order quantity when there is no warehouse space constraint.

Gallino et al. (2017) study the impact of introducing an STS service on sales dispersion and the subsequent effect on the cycle and safety inventories needed by the retailer. The contribution of the 90% lowest-selling products to total sales for the retailer increased by 0.75% points (i.e., increased sales dispersion) due to the channel integration strategy via STS. This increased sales dispersion in turn requires the retailer to increase its cycle and safety inventories by approximately 2.7%.

There exists a rich stream of literature regarding inventory management for dual-channel supply chain. Typically, game theoretical models are developed to examine the retailer's optimal order quantity decision in the retail channel and the manufacturer's optimal inventory level decision in online direct channel (Boyaci, 2005; Chiang and Monahan, 2005; Khouja and Stylianou, 2009; Modak and Kelle, 2019; Takahashi et al., 2011; Yang et al., 2017; Yao et al., 2009; Zhao et al., 2016). Since we direct our attention to store-based retailers handling demand from both online and offline channels, we do not review these papers in detail.

Customers from online and offline channels may be considered as different demand classes with heterogeneous delivery lead time, different delay or lost sales costs, and varying attitudes towards factors such as product price and waiting times. The inventory rationing problem seeks to maximise the expected profit when retailers fulfil both online and offline demand from integrated but limited inventory stockpiles. There are an extensive number of studies in the Operations Research literature on inventory rationing policies in various contexts, such as manufacturers dealing with several demand classes with different lead time or late penalty cost, and businesses in the airline industry handling demand associated with different prices (Ha, 1997; Zhao et al., 2005; Möllering and Thonemann, 2008; Hung et al., 2012). However, studies focusing specifically on different demand classes from different sales channels are very limited. Agatz et al. (2008) have reviewed some models in this field, including Cattani and Souza (2002), Ayanso et al. (2006), Ding et al. (2006). We review the latest models here for inventory rationing in the context of multiple channel retailing.

Chen et al. (2011) study the optimal inventory and admission policies for a scenario when two retailers not only serve their own walk-in customers but also fulfil online demand for an e-tailer who has zero inventory. The retailers have to decide whether to use their inventories to fulfil online demand now for a lower revenue or reserve stocks for their in-store customers in the future for a higher revenue. The authors examine the threshold policies for retailers in two scenarios: when the primary retailer (first option for the e-tailer) shares its inventory information with the secondary retailer and when the primary retailer does not share its inventory information. Alishah et al. (2015) also examine the rationing decision on whether to use limited store inventories to fulfil online demand. They approach the problem with a continuous-time dynamic program and present two heuristics for the optimal rationing policy. Their analyses show that the rationing threshold depends mainly on the cross-channel fulfilment cost and the mix of demand between online and offline channels.

2.5 Warehousing and fulfilment

2.5.1 Warehousing design and processes

The design and processes of warehouses need to be adapted in the transformation towards omni-channel retailing. When inventories are integrated for both online and traditional channels within the same warehouse, one major challenge is to process low-volume and highly unpredictable online orders with high-volume and less fluctuating store replenishment orders in a time- and cost-efficient manner. Customers' intolerance on delivery waiting time also prompts retailers to update their warehouse location criteria. Below we discuss the

adaptations needed in the realm of warehouse management to deliver omni-channel performance.

Warehouse location. When retailers integrate inventories across channels in their warehouses, the pressing delivery timeline for online demand propels omni-channel warehouses to be located closer to cities (e.g., urban logistics parks) where land is limited and expensive (Financial Times, 2016b; Kembro et al., 2018). The real estate investment sector has witnessed strong demand in recent years for smaller urban warehouses (Deloitte, 2014; Financial Times, 2018). There is extensive literature on facility location models for traditional BM retail (Melo et al., 2009). Since omni-channel warehouses process both store replenishment and online demand, the imminent delivery timeline and the more dispersed delivery destinations from online channel add complexity in the trade-offs to optimise the location of warehouse facilities. Therefore, model extensions or new models are needed to guide retailers to optimise the location of warehouses within their networks.

Warehouse layout. The expanding responsibility for online fulfilment has implications on the warehouse layout as well. Online channels generally have a wider range of product assortments, which requires more storage space (Hübner et al., 2016a; Kembro et al., 2018). In addition, online orders have scattered shipment destinations with different delivery time windows, which leads to more complex sorting activities and a potential need for bigger sorting areas (Kembro et al., 2018). Storage allocation has been extensively researched in traditional warehouse management literature but mainly from a single channel perspective (de Koster et al., 2007; Gu et al., 2010). It is interesting to consider whether and how traditional storage allocation models can be adapted for omni-channel purposes.

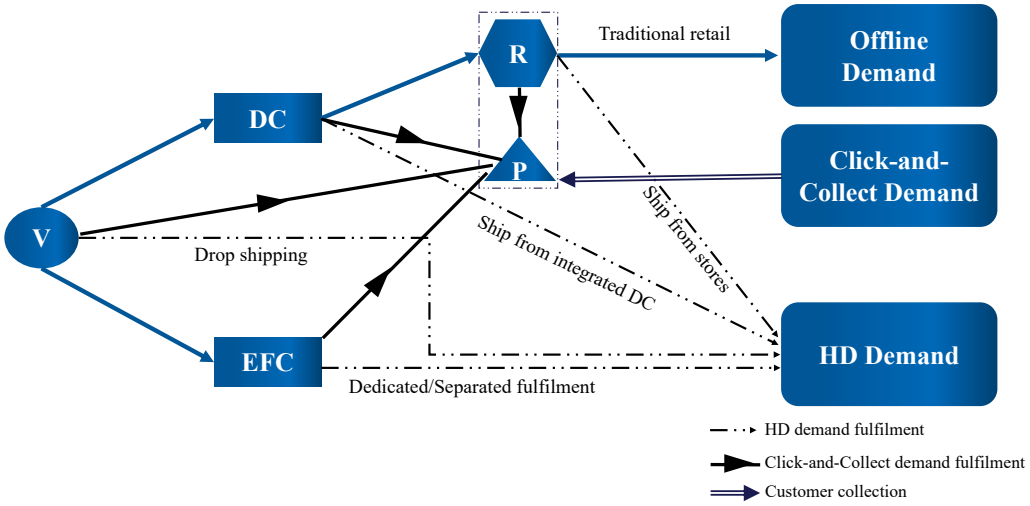
Warehouse activities. Retailers need to decide whether to integrate picking operations for demand from online and offline channels. Integrating picking operations for store replenishment and online fulfilment can reap synergy effects by pooling inventories and capacities. However, the operational efficiency gained from integrated picking operations depend on contextual factors such as store replenishment quantity and frequency (Hübner et al., 2016a; 2016c). Meanwhile, integrating picking for demand across different channels leads to more intricate and challenging picking operations (Hübner et al., 2016a; 2016c). If retailers integrate picking operations, the next question is to carry out fulfilment activities in a time and cost efficient way. Each fulfilment activity, from picking, routing, to sorting and packing needs to be investigated and potentially re-optimised in response to the implications from fulfilling orders for multiple channels. Questions such as the optimal picking routes to traverse the warehouse in order to pick both online and offline orders should be answered.

Warehouse workforce planning. One of the most important implications from integrating picking operations for online and offline channels is the workforce planning. First, performing order fulfilment for multiple channels requires retailers to rethink the workforce level for the increased workload in their warehouses. Second, whereas store replenishment pattern is relatively fixed, online demand may be much more unpredictable (Hübner et al., 2015). Additionally, demand from different channels can have spikes and dips at different times, which requires retailers to distribute and schedule warehouse workers flexibly (Kembro et al., 2018). Note that retailers may adjust the automation levels within their warehouses, which will also alter the economies of current workforce planning. There are many studies on workforce planning decisions in the Operations Research literature. Decisions on the number of workers required and/or the assignment of shifts to workers are examined by formulating mathematical models, typically mixed-integer linear programming (MILP) models. These models often aim at minimizing the staffing costs while subjecting to two types of constraints, namely a predefined service level and conformity with national labour legislations and regulations (Talarico and Duque, 2015). Workforce management studies in the retail sector include Chapados et al. (2014), Kabak et al. (2008), Mirrazavi and Beringer (2007), Pastor and Olivella (2008). Another well-researched field for workforce management is nurse rostering in hospitals (Burke et al., 2004; Ernst et al., 2004).

It is clear that there are more empirical studies than models for warehouse management in the domain of multiple channel retailing. The majority of the warehouse management literature focuses on either a traditional retail setting or a pure online retail context. It remains to be seen how warehousing related decisions such as workforce planning, storage allocation, picking method (e.g., batch picking, zone picking, etc), and picking routing should be adjusted to the omni-channel retail context.

2.5.2 Order fulfilment

Seamless shopping experience with flexible delivery/collection options is viable only when retailers are capable of fulfilling online orders from multiple stocking locations in their networks. Here we divide online demand into click-and-collect and HD demand, where click-and-collect includes BOPS or pick-up in 3rd party locations. If the pickup point is within retail stores, it is the well-known BOPS demand. Otherwise, customers collect online orders from 3rd party locations, such as convenience stores, lockers in a gas station, etc. Both HD and click-and-collect demand can be fulfilled from various locations, including retail stores, integrated DCs, dedicated EFCs and suppliers' warehouses (i.e., drop-shipping) (See Figure 7).



Note: Vendors (V), Distribution Centre (DC), Ecommerce fulfilment centre (EFC), Retail Stores (R), Pickup Point (P), Buy-Online-Pickup-in-Store (BOPS), Home Delivery (HD)

Figure 7. An online fulfilment network with four fulfilment options, adapted from Ishfaq and Raja (2018).

Key issues and related research

Online order allocation. Many researchers have proposed mathematical models to determine how to allocate online orders across various fulfilment locations, i.e., from where retailers should fulfil their online orders. Studies often consider a specific supply chain structure in which the locations of potential online fulfilment facilities are given. The nature of allocation can be static or dynamic. Static order allocation means that once an online customer is assigned to a fulfilment location based on certain policies, orders from the customer will always be routed to the assigned location, regardless of the real-time inventory position across different fulfilment locations. In contrast, dynamic allocation takes into account the real-time inventory level and future demand (e.g. based on information such as promotion planning) across the retailer's systems.

Many studies focus on the static order allocation for retailers who serve both online and offline customers. However, they often implicitly treat online demand as equivalent to HD demand without taking account of the increasingly prevalent BOPS demand (Alptekinoğlu and Tang, 2005; Mahar et al., 2009a; Bretthauer et al., 2010; Liu et al., 2010). These studies have formulated mathematical models to determine how to assign online demand to various potential online fulfilment locations, with the objective to minimise total operating costs including the fixed investment cost, inventory holding and backordering cost, order handling cost and transportation costs. Typically, this research consider a two-echelon distribution system, which comprises a central or regional warehouse in the upper echelon and a number of retail/e-tail locations such as DCs or stores in the lower echelon. Each lower echelon

facility receives in-store and online demand, which could be examined as correlated (Alptekinoğlu and Tang, 2005) or independent (Mahar et al., 2009a; Bretthauer et al., 2010; Liu et al., 2010). The following assumptions are often imposed:

- The two-echelon distribution system is under periodic review. In each period, the inventory status of the system is observed and any orders are placed. Any orders scheduled to arrive in the period arrive and the upper echelon facility (e.g., the central warehouse) allocates stocks to the lower echelon facilities (e.g., stores) according to the equal fractile allocation policy (Eppen and Schrage, 1981).
- Replenishment lead times are deterministic and known.
- Fulfilment locations are capacitated, i.e., there is an upper limit on the number of orders that can be processed in each fulfilment location. Additionally, in-store demand is satisfied from local retail locations and unmet online demand is backordered.
- Both online and in-store orders contain a single item and therefore no order split is involved.

The expected inventory holding and backorder cost per period in their models are computed using the same approach as in Eppen and Schrage (1981) and the transportation cost is often assumed to be linear to the travel distance. These models generally highlight the trade-offs between the inventory holding and backordering cost from pooling against the transportation cost. Additionally, numerical experiments are conducted to investigate the impact of problem parameters such as the demand variance, online demand percentage and unit transportation costs on the order allocation decisions.

Assuming deterministic demand, Ishfaq and Raja (2018) use proprietary data to evaluate four fulfilment options: route online orders for drop-shipping, online fulfilment in retail stores, in integrated DCs and in dedicated EFCs. Notably, their work is the first to incorporate extensive online fulfilment options and to take into acount the size of products by setting a constraint on the storage capacity in each fulfilment facility. As the unit fulfilment cost, holding cost and delivery cost for each type of facilty vary, their results unveil the scenarios under which one fulfilment option outperforms the others. An extension from Ishfaq and Raja (2018) is the joint optimisation of online product pricing and order fulfilment problems to maximise the expected profit for retailers with multiple channels (Ishfaq and Bajwa, 2019). In this case, demand is price sensitive and there is a limit on financial investment in the system-level inventory in each planning period. The mathematical model in Ishfaq and Bajwa (2019) also provides insight on how much online demand the retailer should service in each period to maximise its profit.

When fulfilling online orders in retailer stores, the implications of walk-in customers should be considered. To this end, Scott and Scott (2008) develop a non-linear programming model to minimise the average store congestion, which is approximated by the waiting time in stores. Assuming known walk-in demand in each store and online demand in a region, the authors model the waiting time in stores (proxy of store congestion) using a M/M/1 queuing model with store capacity being the service rate and workload (walk-in demand and allocated online demand) being the arrival rate. There is a given budget for the expenditure on fulfilling online orders in each store. Their analyses provide insights on when retailers should enhance store capacity with a higher budget and when retailers should resort to DCs for fulfilment when demand grows.

The above papers often implicitly assume that there is no inventory replenishment during the planning horizon and no transshipment between fulfilment locations. Torabi et al. (2015) allow for inventory transshipment among fulfilment facilities for a certain unit transshipment cost. They assume that online demand is deterministic and online orders are initially allocated to the nearest fulfilment centre. However, orders can be transferred to another location if the designated fulfilment centre has no inventory. The order allocation and transshipment decisions examined in their paper are mainly for online retailers without traditional in-store shoppers.

Since static assignment policies do not consider real-time inventory positions across online fulfilment locations within a retailer system, stores with a low inventory level may still be assigned for online fulfilment while another location could have excessive inventories. Such static policies can be myopic. To this end, researchers have investigated how to use information about current inventory positions and future demand distribution to dynamically assign online orders across fulfilment locations.

Xu et al. (2009) examine the benefits of periodically reevaluating and reassigning orders that have been assigned to warehouses but have not yet been picked in order to reduce the number of order shipments. They formulate the reevaluation problem as a fixed-charge multiproduct network flow problem with a time dimension to minimise the total number of shipments and develop a heuristic solution approach. They test their heuristics on several data sets from a major online retailer and show that their proposed model could almost halve the number of split orders. Mahar and Wright (2009) propose a ‘quasi-dynamic’ allocation policy to postpone online fulfilment decisions for a single product so that more updated inventory position information can be leveraged. Using such a ‘quasi-dynamic’ policy, they evaluate the total cost (including the inventory holding and backordering, order handling and

transportation costs) and the customer waiting cost separately. Customer waiting penalties are incurred for each time unit of waiting between order placement and receipt. Their computational results show that the 'quasi-dynamic' policy can reduce the inventory cost across fulfilment locations by 14% over the instantaneous order allocation approach. Mahar et al. (2009b) propose two dynamic online order assignment rules by investing in the continous monitoring of online demands and inventory positions in the retail system. The first rule utilises the information on the current inventory levels at different fulfilment locations and assigns an online order to the location which leads to the least total cost between the time lapse of order allocation and the end of the planning period. The second rule takes into account in-transit inventory replenishement quantities for each location. Their computational results show such simple dynamic assignment rules can reduce the total cost by approximately 8.2% over the optimal static policy.

Acimovic and Graves (2015) develop a heuristic for online fulfilment allocation decisions to minimise the immediate plus estimated future expected total outbound shipping costs. They consider different delivery-time options desired by customers, with faster delivery services associated with more expensive transportation modes (e.g., Air for same day). In this case, savings in the shipping cost can be achieved not only by reducing the travel distance but also by using cheaper transportation modes. They propose a linear programming-based heuristic with account for current inventory levels and future demand to decide which fulfilment centre should be chosen for fulfiling an online order using which transportation mode. Their numerical experiments based on industry data report a reduction of 1% on the total outbound shipping cost. Additionally, they show that their heuristic is most valuable for high sales volume SKUs (stock keeping units).

Jasin and Sinha (2015) investigate the more complex situation when each online order consists of multiple items. They develop two fulfilment heuristics to allocate online orders across fulfilment faclities to minimise the total shipping cost with incorporation of demand forecast. The shipping cost function in their model is assumed to be linear in the size of package and transportation distance from designated fulfilment facilties to customers. This function is estimated using linear regression with data on UPS rates for 891 combinations of the package size and the location of facilities and customers. Ardjmand et al. (2018) study the online order cartonization and allocation problem simultaneously by developing a multi-objective model to minimise both the shipping cost and delivery time. They argue that cartonizing the right products together and shipping from the right fulfilment centres for online demand can play a decisive role in the shipping cost and delivery time. Each item and each carton is assumed to be three-dimensional object with a specified length, width and

height and the order cartonization therefore is essentially a bin packing problem. Their numerical results show that the integration of order cartonization and fulfilment allocation is more beneficial for retailers when the online order volume is high, when inventory levels in fulfilment centres are low (and more orders are outsourced to a wholesaler), and when the variability of the size of items is high.

Among the papers mentioned above, two studies focus on retailers with both online and offline channels (Mahar and Wright (2009) and Mahar et al. (2009b)), while other papers concentrate on online retailing context. However, the dynamic order assignment policies for the online retail can still be used to guide omni-channel retailers. Online order assignment problem is not entirely new to the Operations Research community. An analogy can be drawn with the classic Location and Allocation Problem (LAP), which involves the location of facilities and the allocation of customer demand to facilities in the design of a distribution system. Readers are referred to (Brimberg et al., 2008; Klose and Drexl, 2005; Manzini and Gebennini, 2008; Owen and Daskin, 1998) for more details about the LAP.

To reiterate, studies on the order allocation problem often assume a specific supply chain structure with potential fulfilment locations a priori. To the best of our knowledge, only two papers have examined the fulfilment structure for retailers serving both online and offline customers. Mahar et al. (2014) examine the optimal set of retail stores to be deployed as pickup points for BOPS services. They classify online demand further into HD demand and BOPS demand, where HD orders can be handled only in EFCs, and BOPS orders are satisfied from store inventories unless stores experience a stock out. There are four alternatives for BOPS customers when in-store pickup is not offered in their local stores: 1) customers switch from BOPS service to HD service, 2) customers switch from BOPS to traditional in-store shopping, 3) customers satisfy their BOPS demand by going to another pickup point within the retailer's network, 4) customers abandon their BOPS orders. Their numerical experiments explore the impact of customer switching behaviours, various cost parameters and the ratio of HD and BOPS demand to the total online demand. Their results suggest that an 'Allsite' policy (all retail stores are used as pickup points for BOPS demand) can be suboptimal.

Aksen and Altinkemer (2008) present a location-routing model to jointly optimise the location of 'click-and-mortar' (CM) stores (serving customers from both online and offline channels) and the routing of delivery vehicles from CM stores to online customers. Note that CM stores are the only sources for online fulfilment in their paper. The model presented in Aksen and Altinkemer (2008) is featured as a single period, two-stage, deterministic demand,

capacitated vehicles, and hard time windows for deliveries. Additionally, they assume that the travel time between locations is known and vehicle capacity of trucks is uniform for all CM stores. Their model results provide insights on which BM stores should be converted as CM stores, which BM stores should be closed, whether and where to locate new CM stores, and from which CM stores to ship an online order with a hard delivery time window.

2.6 Last mile delivery

Last mile delivery has long been regarded as the Achilles' heel for retailers, since it is often formidably expensive (Devari et al., 2017). It has arisen with the boom in e-commerce. It is concerned mainly with the delivery planning decisions to ship online orders to customers. The difficulty and operational inefficiency of last mile delivery is attributed to the distinctive features of online orders, which are unpredictable, low in volume, high in variety and more scattered geographically with the various delivery destinations and faster delivery requirements (Metters and Walton, 2007). Delivery options, velocities and fee structures offered to customers have a significant impact on both delivery cost and customer satisfaction (Hübner et al., 2016c; Marchet et al., 2018). In today's retail world, there are mainly two delivery options - home delivery and click-and-collect. It is common practice for retailers to offer click-and-collect services, which reflect cross-channel integration (Brynjolfsson et al., 2013). There are two dimensions of the delivery velocity, including timelines (e.g., same or next day) and time slot (e.g., 4-5pm or 4-6pm). Based on the delivery options and velocity, we categorise delivery services into nine types (see Figure 8). In practice, many retailers offer an express delivery service (e.g., same/next day) and a standard delivery service (e.g., 3-5 days) with different delivery charges.

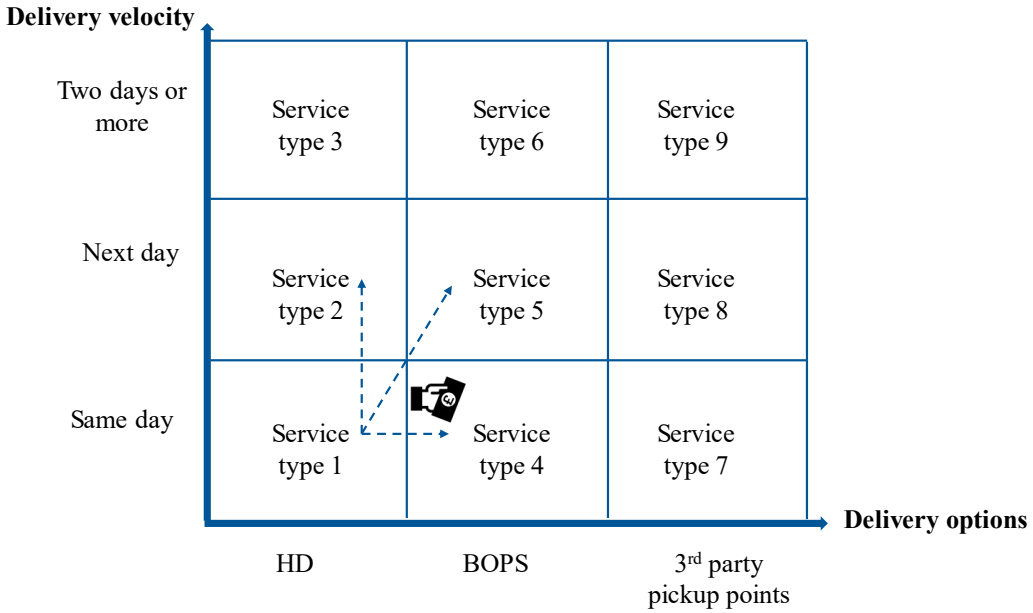


Figure 8. Classification of delivery services in omni-channel retailing.

Compared to internet retailing, the most significant novelty of last mile delivery in omni-channel retailing is the addition of the click-and-collect option. The implication of such extra flexibility is the rise of several new issues in last mile delivery. One issue is the potential integration of shipments between store replenishment orders and online orders for BOPS customers. Another issue is related to the pricing for delivery services to manage customers' demand for different delivery options and velocity. The following sections discuss key decisions in last mile delivery relevant specifically for retailers moving to omni-channel offerings.

Key issues and related literature

Order acceptance mechanism. Given the high level of last mile delivery costs, retailers need to decide whether they would provide delivery services for all incoming orders or design a mechanism to reject certain orders that lead to unprofitable fulfilment. Several studies have applied revenue management to decide how many delivery requests should be accepted to maximise total profits for directly delivering online orders to customers. Campbell and Savelsbergh (2005) emphasise the importance of order capture when planning the routing and scheduling of online deliveries. Following the revenue management idea in the airline industry, they propose a mathematical model and algorithms (based on insertion heuristics) to dynamically determine whether to accept or reject a delivery request and the assignment of delivery time slots for the accepted requests. In the same spirit, Ehmke and Campbell (2014) examine how to maximise profits by maximising the number of delivery requests to be accepted in metropolitan areas, while ensuring feasible delivery routing in compliance with promised service time windows.

Regarding the click-and-collection option, researchers have argued that not all products are suitable for BOPS service. Instead of designing some order acceptance mechanisms, they develop analytical models to show that retailers can reap higher profit by limiting BOPS services to specific product types and service areas (Bhatnagar and Syam, 2014; Cao et al., 2016; Gao and Su, 2017; Jin et al., 2018). It is worth pointing out that the decision with regards to accepting HD orders or providing BOPS services for certain types of products is driven by profit maximisation over a short term. It will be valuable to examine such decisions taking account of the long-term impact of not offering the desired delivery/collection services to customers.

Joint pricing and delivery optimisation. Generally, demand for different types of delivery service is not well balanced, with certain types extremely popular and others less desired. This motivates retailers to design a menu of different delivery services (i.e., different delivery

modes or velocities or time slots) and the corresponding delivery charges to maximise revenue. Many studies have examined pricing decisions problems in the context of online retailing such as the threshold needed to offer free-shipping (Leng and Becerril-Arreola, 2010; Becerril-Arreola et al., 2013) and the choice between two price partitioning strategies (i.e., stipulated shipping fee and free shipping with higher product price) (Gümüş et al., 2013).

In addition, there are many studies focusing on the joint optimisation of pricing and time slot management for HD services, i.e., the assignment of orders to different delivery slots. The objective is to maximise profit or minimizing delivery costs by reserving limited delivery capacity of popular time slots for the most profitable orders while steering less lucrative orders to other slots with sufficient capacities (Asdemir et al., 2009; Campbell and Savelsbergh, 2006; Yang and Strauss, 2017; Yang et al., 2014). Customer demand for each delivery slot is often estimated with utility theory. In this way, retailers can dynamically adjust the price for different delivery slots so that customers can be enticed to select a particular delivery slot leading to the most cost-efficient delivery schedules for retailers. The delivery cost is often estimated based on classical vehicle-routing techniques. Agatz et al.(2011) also investigated how to assign each online customer to the set of time slots in a service region. However, they only consider the allocation of time slots to customers to minimise the total delivery cost without considering the possibility of managing demand for different delivery time slots by price differentiation.

Apart from enticing customers to choose one slot over another, retailers can also steer customers to switch from one delivery option to another. Rabinovich et al. (2018) develop dynamic pricing solutions to incentivize customers to choose BOPS services over HD services, since the former services avoid the formidably expensive last mile delivery costs. To encourage customers to choose BOPS services, retailers often reward customers by waiving the associated in-store fulfilment fees indiscriminately. Rabinovich et al. (2018) use field data and econometric models and evaluate the economic value of the strategy they develop to induce customers to switch from HD to BOPS service, compared to the commonly used indiscriminate pricing strategy. They show that retailers adopting the indiscriminate pricing strategy can be worse off. The loss of the delivery charges for BOPS service and the additional operating cost at stores to perform order fulfilment can be higher than the savings in the last mile delivery cost. Conversely, using a targeted incentive they propose incurs much lower costs (approximately three times) than the savings in last mile delivery cost.

The key idea of applying revenue management to manage demand for different delivery services is service differentiation. Figure 8 depicts nine possible different services ignoring the time slot dimension. Reviewing the extant literature in this stream shows a clear focus on time slot management in HD services. Studies on managing demand for different options, i.e., BOPS and HD, have just started to emerge. This is an important avenue for future work on providing differentiated delivery/collection options with carefully crafted prices for each option in order to align delivery/collection requests with retailers' limited resources (e.g., delivery vehicles).

Vehicle routing problem (VRP). Research on VRP related issues has a long-standing history with a rich, varied and still growing body of literature. Different variants of VRP problems have been examined, such as VRP with time windows, with simultaneous delivery and pick-up, capacitated VRP and dynamic VRP. For an overview of this topic, readers are referred to Eksioglu et al. (2009), Golden et al. (2008), Laporte (1992), Laporte (2009) and Pillac et al. (2013). Comparatively, only a few researchers have studied the VRP in the omni-channel retailing domain. One new factor here is that resources, specifically delivery trucks, can be used for delivering orders from different channels. Abdulkader et al. (2018) propose a model that uses a fleet of homogenous vehicles to replenish stores from warehouse and to ship online orders from store inventories to customers (i.e., SFS). Their results provide a number of vehicle routes to visit both retail stores and online customers with the same fleet of vehicles, so that the total delivery cost can be reduced. Another variant is to combine the traditional store replenishment with online orders in the same delivery truck to minimise the transportation cost (Nekoiemehr et al., 2018; Paul et al., 2019). Some retailers fulfil BOPS orders in warehouses or DCs before shipping them to stores for customer collection (i.e., STS). As a result, some of these BOPS orders can be bundled with the traditional store replenishment orders to reduce the transportation cost. The proposed models in these studies often determine which online orders should be combined with store replenishment orders and which orders should be delivered directly with separate vehicles to minimise the total cost. The quoted lead time to online customers is also considered in the model (Nekoiemehr et al., 2018). It is common practice for omni-channel retailers to leverage stores for both BOPS and BORS services. Therefore, the problem of combining online and store replenishment orders can be extended to a delivery and pick-up VRP problem using a same truck to deliver online and store replenishment orders and also pick up returned online orders from the same store for further processing.

In addition, the timing of shipment has also received some attention with the growth of online retailing. The 'First-Come-First-Serve' (FCFS) time-based shipment practice has been

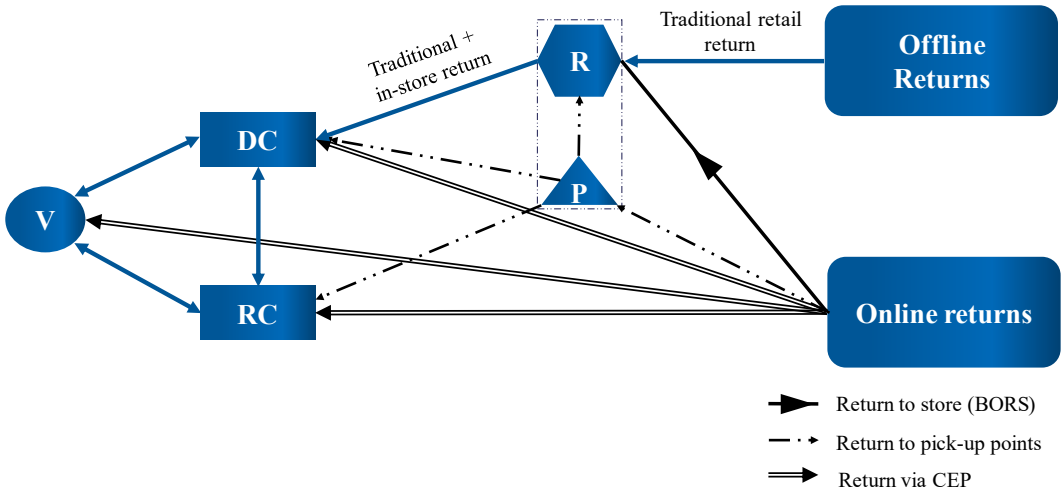
shown to be non-optimal - cost reduction benefits may be achieved in optimising the time to dispatch last mile deliveries. Wei et al. (2017) investigate the optimal time to ship online orders to balance the benefits of order consolidation and the cost for expedited shipping. With given deadlines for online orders and available shipping options, they develop a dynamic programming model to examine from which warehouse orders should be fulfilled, when orders should be shipped and what orders should be included in the shipment . Zhang et al. (2019b) examine the order split problem in last mile delivery where multiple parcels are shipped to the same customer at different times. They also study the benefits of delaying some orders for consolidation to reduce the number of shipments. However, they focus on the trade-off between reduced shipping cost and increased inventory holding cost by postponing orders for consolidation. They present an integer programming (IP) model to determine what orders should be postponed for consolidation and when to dispatch online orders from the delivery station to customers. The above-mentioned studies mainly focus on handling online orders. It is worth examining how to apply these studies to guide omni-channel retailers to determine when to dispatch shipments with the combination of online orders and store replenishment orders.

2.7 Returns management

Managing product returns involves consideration from both the marketing and the operational sides of a retail business. Poor returns management not only imposes direct financial costs on retailers but also jeopardizes retailers' brand due to deteriorating customer service. One dimension of returns management from a marketing perspective is customer management. To this end, many studies have examined the key drivers to consumer returns (Asdecker et al., 2017) and the impact of different factors on customers' returning behaviour. These factors include returns policies such as returns period (e.g., time window for returns to be accepted), the convenience of returns options (e.g., BORS option), and the cost to return (e.g., no refund, partial refund or full refund) (Bower and Maxham, 2012; Griffis et al., 2012; Minnema et al., 2016; Shang et al., 2017). The identification and mitigation of fraudulent online consumer returns has also attracted research attention (Harris, 2010; Hjort and Lantz, 2012; IBM, 2001). From the operations perspective, retailers should efficiently manage the physical reverse logistics from receiving, sorting, processing, analysing (the value of returned items) to disposing (e.g., remanufacturing, donate, route to discount outlets, etc.), so that inventory cycle time and returns handling costs can be reduced (Stock et al., 2006). Many problems with physical reverse logistics, such as the reverse network design, lend themselves to modelling techniques. Readers are referred to Fleischmann et al. (1997), Rogers et al. (2012) and Govindan et al. (2015) for reviews on quantitative models in the domain of reverse logistic in traditional retailing.

One inexorable feature for omni-channel retailers is the high return rate, as online sales capture a growing share of total retail sales. Online return rates are typically much higher than that in traditional retail. Approximately one third of online orders are returned, compared to the 9% return rate in the traditional channel (Bloomberg, 2017). Bernon et al. (2016) also provide empirical evidence to show that online-based return rates for clothing and home product categories on average double the store-based return rates. The significant volume of returned products calls for more attention to returns management.

Another novelty of returns in the omni-channel era is the propagation of cross-channel returns. To facilitate customer returns, it is common practice for retailers to offer BORS options for online purchases (Bernon et al., 2016). Hübner et al.(2016a) is one of the first studies to present a holistic view on the return network in omni-channel retail. They identify two main return options: in-store returns and home collection through CEP. Building on Hübner et al.(2016a), we present a general return network configuration based on returns entry points and returns processing facilities (see Figure 9). Apart from in-store returns and home collection, returning to 3rd party locations (e.g., parcel lockers) is also possible. For example, CollectPlus, a UK-based courier service expert, has a wide network of collection and return points to support its retail clients (e.g., John Lewis) to offer flexible click-and-collect and cross-channel returns options (www.collectplus.co.uk). Returned products could be handled directly in a store, a centralised returns centres (RC), an integrated DC, or even be routed to the suppliers/manufacturers for disposition. The increased volume of returned products and the increased variety of returns options in omni-channel retail add further complexity to design the returns network and manage the returns handling processes.



Note: Vendors (V), Distribution Centre (DC), Return centre (RC), Retail Stores (R), Return Point (P), Courier, Express, and Parcel (CEP)

Figure 9. Product returns options in omni-channel retailing.

Key issues and related literature

Returns network design. The burgeoning volume of online returns puts a strain on retailers' DCs and/or RCs, urging them to consider expanding returns handling capacities. This concerns strategic decisions on the number, location and capacity of returns processing facilities. Another problem emerging in the omni-channel retail is the decision on the number of retail stores to be used to accept returns of online purchases. Customers can enjoy great convenience with BORS options, and retailers can potentially exploit cross-sell benefits from the increased store footfall. When in-store returned items are processed and restocked in stores for resale, transportation for shipping returns back to DCs/RCs is avoided and inventory cycle time is shortened. However, as online channels tend to have a broader product range, online orders could be returned to stores that do not sell the items, which necessitates further shipment back to the returns handling facilities. Therefore, when offering BORS options, retailers must decide which stores can accept returns from online purchases. They need to determine whether to allow all stores to accept online returns or only those that stock and sell the returned items. Naturally, such decisions are not isolated from other supply chain and inventory placement decisions.

The design of the returns network defines the possible routes for the flow of returned products. With the introduction of BORS, retailers need to decide which returned items should be kept in stores and which should be sent backward to DCs or RCs at what cadence (Radhi and Zhang, 2018). Some retailers with less store presence may seek partners (e.g., sister retail group or parcel service providers such as CollectPlus) to provide more extensive in-store returns options (Bernon et al., 2016). In this case, retailers need to arrange the flow of returned products from locations outside their distribution networks to the corresponding processing facilities.

Compared to the rich literature on reverse logistics network design in traditional retail, studies on returns management with multiple retail channels are remarkably scarce (Agatz et al., 2008). Studies such as Min et al. (2006) have focused on the reverse network design problem for e-tailers who collect returns from customers to centralised RCs via CEP. However, they do not account for BORS options, which reflect cross-channel integration and have become increasingly prevalent. To the best of our knowledge, Mahar et al.(2014) is the first to develop a mathematical model to examine the best set of retail stores that should be used as BOPS and/or online returns points. In their study, undesired products purchased online can be returned either to retail stores or to a central warehouse (via CEP). The former option avoids the cost of shipping online returns to the central warehouse but incurs fixed operating costs to maintain in-store returns handling capabilities. There are three streams of

demand to be served, namely HD demand, BOPS demand and online returns demand. The authors develop a mixed integer non-linear program (MINLP) model to handle these three demand streams in a network comprising both retail stores and the central warehouse. It is worth pointing out that their model implicitly assumes that items returned to stores will be processed and restocked in stores, i.e., stores keep all returned items.

There is a lack of studies on reverse network design for retailers offering cross-channel returns options. Additionally, the existing literature rarely differentiates the optimal network design for different types of products with different lifecycles. It is important to compare the cost, speed and profit margin of BORS and returns via CEP for online purchases. Such comparisons could be used to manage demand for different returns options with the application of revenue management, so that customers could be enticed with appropriate incentives (e.g., vouchers) to choose the least costly return options for retailers.

Returns processes. As discussed, retailers need to decide whether to keep returns via BORS options as on-hand stocks at stores or send them back to facilities such as DCs. Dijkstra et al. (2019) investigate dynamic transshipment policies for returned products via BORS. The commonly used static policies for BORS are 'Ship None' (all cross-returned products are kept at stores) and 'Ship All'(all cross-returned products are shipped back to DCs of the online channel). The authors argue that such static policies are not optimal and may exacebate the inventory unbalance issue. They propose dynamic transshipment policies using MDP. The MDP model optimises the number of returned products to be transshiped from physical stores to DCs for the online channel in each planning period in a single selling season, with the objective to minimise the total cost (including inventory holding, transshipment and penalty cost for each unsold stock by the end of the period).

Muir et al. (2019) study the interaction between a retailer's product return policy (i.e., same channel or cross-channel returns options) and processing structure (centralised or decentralised) on inventory performance, which is measured by inventory availability and timeliness. They develop two research hypotheses based on contingency theory and test them using discrete-event simulation experiments. The simulation model parameters include customer demand characteristics (e.g. order inter-arrivals and quantity), (Q, r) inventory policies, transportation velocity (e.g. replenishment lead-time and returns transport time), the return rate and the probability of return recovery. The model outputs include the on-time fill rate to capture inventory availability and the average order cycle time (i.e., average time elapsed between order placement and order shipment) to represent timeliness. Their results show that decentralised returns processing in retail stores improves inventory performance

as products can be processed, re-shelved and resold more quickly. Additionally, when demand seasonality is not extremely high, a cross-channel returns policy together with a decentralised processing structure may mitigate the adverse consequences of product seasonality.

Incorporating product returns into inventory and pricing policies. When offering BORS, it is likely that online returns could end up at stores that do not sell the returned products. Stores may also end up with excessive stocks (own inventories and returned products) that cannot be absorbed by customers. As a result, offering BORS options requires inventory re-balancing in retailers' distribution networks. It also necessitates adjustment of ordering quantity decisions considering the increased amount of resalable products from online returns. Such consideration could change the equilibrium of retailers' profit functions and therefore yield new optimal product prices (Bernon et al., 2016). Generally, research in this field uses a benchmark case in which BORS is not offered (i.e., single-channel for online returns). Quantitative models are then developed to optimise the inventory ordering decisions for different facilities in different channels and/or product pricing decisions. As a result, comparisons can be made between the results in the case of BORS options and in a benchmark case. Such comparisons generate insights on whether retailers should offer BORS and how retailers should adjust their ordering and pricing decisions when allowing for BORS.

Radhi and Zhang (2019) examine the inventory ordering decisions for both online and offline channels when considering returned products. Assuming a full refund policy and exogenous unified price in both channels, they formulate mathematical models to derive the optimal ordering quantities for both online and physical stores under four policies. These four policies differ regarding whether BORS is available and whether the transfer of ownership for returns via BORS incurs payment (i.e., centralised or decentralised returns management). Any returned product can only be resold once in a selling season if it is resalable. A returned resalable product is associated with a return collection cost and a returned but not resalable product is redeemed with a salvage value. Their results show that under the inventory shift policy for BORS, the order quantity in physical (online) stores is no higher (lower) than that under any no-inventory shift policy. When the ownership of returns shifts to the store that accepts the returned products, the returned resalable products are regarded as replenishment for the store and therefore lowers its ordering quantities. They also conduct numerical analyses on the impact of various factors, such as the shipping cost (for both fulfilment and returns), customers' channel preferences and the resalable probability, on the ordering quantity decisions for each channel.

He et al. (2020) utilise game theory to jointly optimise the pricing decisions for both new and refurbished products (after returns) and the ordering decisions for both online and offline channels. A benchmark case without the BORS option is used (i.e., single-channel for online return). They assume that the price for refurbished products is always lower than that for new products and all returned products can be refurbished for resale. Their results show that by allowing for BORS, the retailer can charge a higher price for refurbished products, decrease the order quantity for the store, and increase the order quantity for the online channel. In most cases, the BORS option can achieve a win-win outcome, which provides retailers with higher profits and customers with better service without paying extra fees. Such benefits are more significant when the proportion of customers choosing the BORS option is high. Only when the price for new products is sufficiently higher than that for the refurbished products will the BORS option yields lower profit for the retailer, compared to the benchmark case.

Radhi and Zhang (2018) examine the pricing strategy for a retailer with a physical store and an online store under same-channel and cross-channel returns policies. Sales in both stores (online or physical) are assumed to be linear with product prices. The order quantity in each store takes into account the possibility of a product returned to that store and the probability of the product being resalable in that store. The profit function in each store is the result of revenue from sales minus the purchase costs from inventory ordering. Both cases of centralisation and decentralisation have been examined. Their results suggest the existence of a threshold value for customers' preference for the physical store, beyond which the price in the physical store should be higher than that in the online store.

Zhang et al. (2018) study the value of the omni-channel strategy and its optimal pricing and inventory decisions, compared to the benchmark online-only strategy. Under the omni-channel strategy, a consumer can choose the 'reserve-online-pick-up-and-pay-in-store' (ROPS) option, as well as the common online direct option (i.e., pay online and wait for shipments). The omni-channel strategy provides customers with more purchasing options but also helps retailers to expand their customer base. The cancellation cost from ROPS and the return cost from online direct option are incorporated in the customer utility expression, which affects the demand function and therefore the revenue obtained. They also consider the cross-selling benefits when customers visit retail stores for ROPS purpose. Using a newsvendor model, the authors demonstrate how to obtain the corresponding optimal pricing and ordering quantity to maximise profit under the omni-channel strategy and the online-only strategy respectively. They show that the return rate from online direct option decreases while the cancellation rate from ROPS option increases under the omni-channel strategy.

Ofek et al. (2011) investigate the impact of product returns on deciding the pricing and store assistance (SA) level for BC retailers. Retailers can invest in a higher SA level (e.g. more-knowledgeable salespeople and greater shelf display) to reduce the return rate in the offline channel. There is a return cost for both retailers and customers. Return costs for customers capture the disutility, insufficient refund and the opportunity cost of time related to returns processes. Customer utility in their study is dependent on the product price, the degree of product differentiation between two competing retailers, the return cost and the shopping trip cost. The authors first examine a benchmark scenario of a monopolist retailer that owns both products and operates both online and offline channels. They later study the duopoly case with two competing retailers where each retailer owns one of the products and non-cooperatively sets the SA level and then product price. Though their analyses mainly focus on whether BM retailers should introduce the online channel (i.e., channel strategies), they show how the product return rate can affect marketing (i.e., price) and store operations decisions (i.e. SA level).

The abovementioned research is concerned with the inventory and product pricing decisions whilst assuming that return costs for online purchases are given. Naturally, both the product pricing and the returns fee will affect customer utility and therefore sales, which we elaborate in the following section on returns policies. It is value-adding for future work to extend the research discussed above and examine decisions on both product pricing and returns fee together with inventory ordering decisions taking into account anticipated returns. Comparison between two price partitioning strategies (stipulated returns fee vs. free returns with higher product price) that has been discussed in deliveries can also be explored in the case of returns.

Returns policies. Optimisation of returns policies concentrates mainly on determining the charges for online returns (i.e., restocking fee) and the returns period. Existing quantitative models often jointly optimise the restocking fee for online returns and the product price in order to maximise the expected profit. Product pricing is usually assumed to be identical across online and offline channels, as suggested from the empirical evidence in Cavallo (2017). Generally, the BORS option is free of charge if provided, whilst customers might be charged for returns via CEP as per retailer's return policies. Additionally, this research often segments customers into different groups based on customers' returns costs for different options, such as the hassle cost or inconvenience of travelling to stores for BORS. Jin et al. (2019) present a multi-stage, non-cooperative game framework in a duopoly competition to study when a retailer should offer BORS and how to set the product price and online return fees. The two retailers sell horizontally differentiated products through online and offline

channels. Customer heterogeneity is captured by the differences in physical shopping costs, return costs, and retailer preferences (customers are assumed to be uniformly distributed on the Hotelling line (Hotelling, 1929)). The two retailers decide whether to offer BORS independently at stage 1 and decide their selling prices and online return fees at stage 2 after observing one another's BORS offering. At stage 3, customers make their purchase decisions by taking into account the likelihood of product mismatch and return. They show that the adoption of BORS can increase or decrease profit under different conditions. At least one retailer will offer BORS in equilibrium when their products are sufficiently differentiated and when the salvage value of returns is significantly higher via offline (BORS) than online (returns via CEP). Conversely, none of the retailers will offer BORS if their products are highly substitutable or if the salvage value of returns via BORS is comparable to that via online.

Nageswaran et al. (2019) develop analytical models for an omni-channel retailer to determine the product price (uniform across channels) and the charge for online returns via CEP (BORS option is free). Customers could shop online and potentially return the product afterwards (i.e., the online type). Alternatively, customers could visit the store for product inspection before purchasing and therefore do not initiate returns (i.e., the store type). Their model considers the ratio of customers choosing to shop online and the level of convenience of BORS. Such consideration allows for exploring the impact of improving customers' accessibility to stores (e.g., more store openings). They show that the optimal amount of refund for returns in an omni-channel firm exceeds its salvage value, which explains why most retailers with multiple channels offer generous refund policies in practice. Their analyses suggest that retailers should offer full refunds when returned products can be salvaged at mild discounts and when most customers choose to visit the physical stores before purchasing. They also recommend that retailers should charge a fee for returns via CEP when retailers have a wide store network and better in-store salvage opportunities.

Chen and Chen (2017) present analytical models to examine a retailer's decisions on channel selection, customer returns policy, and pricing for both online and offline channels. Customer demand is dependent on factors including the product price, average customer satisfaction rate and average unit returns cost (e.g., travel to stores for BORS). The last two factors capture the impact of returns policy on demand. They first derive the profit function and the optimal product price for the benchmark case when the retailer operates a single channel. Their results show that if a retailer is restricted to adopt a single channel, the online channel is less attractive under the 'no returns allowed' policy, since the customer dissatisfaction rate will be very high. They later investigate the dual-channel case and study

decisions on pricing and returns policies. Considering the ease of providing personalised pricing via online channels (i.e., discriminating pricing across channels), the authors also examine the impact of personalised pricing for the online channel on the optimal pricing and returns policies. Their results identify conditions under which a retailer should choose a single online channel, single traditional channel and the dual-channel (with both online and offline channels). They also suggest that the retailer should offer a 'Money-Back-Guarantee' in a channel when the net salvage value of the returns is positive. Further, it is more beneficial for the retailer to improve the net salvage value of the returns in the online channel than in the traditional offline channel. They show that using the personalised pricing for the online channel makes the online channel and the 'Money-Back-Guarantee' policy more attractive. Compared to a uniform pricing across channels, a personalised pricing for the online channel can help the dual-channel retailer increase its market share and profit (up to double).

The returns period defines the time window during which returned products can be accepted and refunded. Difrancesco et al. (2018) develop a non-linear programming model to jointly optimise the returns period (strategic) and the service rates at returns handling facilities (operational), with the objective to maximise total profit. They model a closed-loop fashion supply chain in which returns can flow from customers to the following destinations: 1) to warehouses to be refurbished or remanufactured so that products can be re-introduced into the primary market; 2) to suppliers to be recycled and then re-introduced into the primary market in the form of new raw material; and 3) to outside of the supply chain to be disposed or sold in the secondary market (e.g., outlet stores with lower product price). Their model factors in the effects of product age and operational delays through the supply chain as well as delay due to customer sojourn time, expressed as the expected throughput time of an M/M/1 queue. They derive a closed-form expression for the optimal return window and for the optimal service rates at the sorting/testing centres and refurbishment facilities. Their analytical results are cross-validated with data from Zalando, a German online fashion retailer. Numerical experiments are conducted to provide managerial insights, such as the effects of return window, customer sojourn time, product aging, and return rate.

Though some research has shed light on when retailers should offer BORS services, such guidance mainly considers inventory related costs and product prices and does not incorporate into logistics costs to handle returns that are received from different channels (i.e., via BORS or CEP). Therefore, it is vital to examine the cost to serve for different returns options and understand under what conditions BORS options are beneficial for retailers. Questions surrounding BOPS options such as the ideal product types and service areas for

BOPS services can also be asked in the case of BORS services. In this way, retailers can make better informed decisions on BORS service offerings.

2.8 A conceptual map of omni-channel SCM

By synthesizing and critically reviewing all of the relevant literature discussed above, we propose a general conceptual map for key supply chain and operations management decisions that retailers must consider in the transition to omni-channel (see Figure 10). The upper right boxes in Figure 10 list four areas (order information exchange, channel assortment, delivery options and returns policies) of channel integration that can be triggered by customers, thereby influencing the front-end customer experience. The lower right boxes present four areas (inventory management, warehousing & fulfilment, last mile delivery and returns management) of channel integration that can be controlled by retailers to achieve back-end supply chain and operational efficiency. Both front-end customer experiences and back-end SCM are shaped by the channel strategies (the number and role of multiple channels) and contextual factors including the characteristics of retailer, market, product and supplier.

Although we have discussed the key issues and decisions under the four broad areas separately, these decisions are necessarily inter-related. For example, the fulfilment decision on from which facility to satisfy online demand depends on the inventory placement decision. It also affects the last mile delivery planning since the fulfilment location defines the origin of delivery routes. Therefore, it may be myopic in some cases to optimise decisions in one area alone without taking account of its implications in other areas. It is necessary to balance the comprehensiveness and tractability of the problems under investigation.

Figure 10. The conceptual map of key supply chain and operations management decisions for omni-channel retailing.

The studies discussed in this chapter often conduct numerical experiments and sensitivity analyses for model parameters such as demand volatility, online order volume, product size/weight, supplier lead-time, and so forth. These parameters are essentially indicators of contextual factors related to retailer characteristics, market characteristics, product

characteristics and supplier characteristics. Such contextual factors affect the critical supply chain and operations decisions for omni-channel performance. For example, household supermarkets such as Tesco often use a combination of retail stores and dedicated EFC(s) to fulfil online groceries orders. Considering the product perishability and high customer intolerance on waiting, a drop-shipping option with long supplier lead time is unlikely to be used. Store-based fulfilment is more attractive, as the wide store network allows for express deliveries. Besides, an online grocery order generally consists of approximately 40 lines of items and groceries are fast moving consumer goods with high order volumes in a period (e.g., a week or a month) (Fernie and Spark, 2014). Such product features justify the utilisation of dedicated EFCs in regions with high demand level, because retailers can obtain significant savings on the fulfilment cost owing to higher picking efficiency. We have therefore highlighted the impacts of these contextual factors in our conceptual map.

Based on the discussion from section 2.4 to section 2.7, we summarise the major contributions from existing studies and highlight promising research agendas and questions in contemporary retailing for each of the four planning domains identified earlier (see Table 2). These should help to guide retailers to reconfigure their supply chain in their journeys to omni-channel.

Table 2 Research agendas for key SCM decisions in transition to omni-channel retail.

Planning areas	Research themes	Contributions from existing literature	Research agendas for future work
Inventory management	Inventory ownership	Decision on the optimal mix of in-house inventory and drop-shipping (Khouja, 2001; Bailey and Rabinovich, 2005; Netessine and Rudi, 2006; Chiang and Feng, 2010).	- How should the optimal mix of in-house inventory and drop-shipping differ across different types of retailer and across different product types?
	Inventory integration level	Explore the advantages of using inventories previously reserved for offline demand to fulfil online demand. Specifically, compare the strategy of partially integrating inventories at the store level against separated inventories for online and offline demand (Seifert et al., 2006; Bendoly et al., 2007; Hovelaque et al., 2007; Alishah et al., 2015; Govindarajan et al., 2017).	- What are the benefits of partially integrating inventories at the store level when transshipment between stores is possible? - Where should integrated inventories be located? Should inventories be integrated centrally at the DC level or at the store level, or single stock at all levels? - Examine the benefits of using online inventories to fulfil unsatisfied offline demand.
	Inventory replenishment policy	Decisions on how retailers should adjust the inventory control policies when: - drop-shipping is considered (Khouja and Stylianou, 2009; Ma et al., 2017)	- Examine the joint optimisation of inventory replenishment and online order allocation policies.

Planning areas	Research themes	Contributions from existing literature	Research agendas for future work
		- resalable returns are considered (He et al., 2020; Zhang et al., 2018) - cross-channel fulfilment is considered, i.e., integrating inventories for offline and offline demand (Schneider and Klabjan, 2013; Xu and Cao, 2019; Xu et al., 2017; Alawneh and Zhang, 2018).	
Warehousing & Fulfilment	Warehousing management	Empirical studies on the necessary adaptations in warehouse management (e.g., locations, layout design, picking and sorting activities, workforce planning) for retailers moving to omni-channel retailing (Hübner et al., 2015; Hübner et al., 2016a; Kembro et al., 2018).	- How can traditional warehousing management decisions, such as pick routing and storage allocation, be adapted to the omni-channel context? - How do the necessary adaptations vary among different types of retailer and across different product types?
	Order allocation	Optimal policies to allocate online orders across a set of known retail facilities (e.g., stores, DCs) to minimise total operating costs when: - order allocation decisions are static (Alptekinoğlu and Tang, 2005; Mahar et al., 2009a; Bretthauer et al., 2010; Liu et al., 2010; Ishfaq and Raja, 2018; Ishfaq and Bajwa, 2019; Torabi et al., 2015) - order allocation decisions are dynamic (Xu et al., 2009; Mahar and Wright, 2009; Mahar et al., 2009b; Acimovic and Graves, 2015; Jasin and Sinha, 2015).	- How to design appropriate performance metrics and incentive schemes to facilitate cross-channel fulfilment? - How can online order fulfilment operations be organised and planned in retail facilities (e.g., stores and DCs) to balance costs and service?
	Facility location for online order fulfilment	Decisions on the set of retail stores that should offer BOPS and BORS services (Mahar et al., 2004). Decisions on what stores to be converted as CM stores to serve walk-in customers and HD customers, what stores to be closed, and whether and where to locate new stores (Aksen and Altinkemer, 2008).	- What are the types and numbers of retail facilities (e.g., stores and DCs) that should be deployed to fulfil BOPS and HD orders as well as walk-in customers? - How should retailers reconfigure their store network (e.g., store closures) taking into account the shift of customer spending to online channels and the need to maintain some stores for omni-channel purposes (e.g., showrooming, BOPS/BORS points)
Last mile delivery	Order acceptance mechanisms	Decisions on whether to accept or reject an online delivery request and subsequently how to plan deliveries for the accepted requests (Campbell and Savelsbergh, 2005; Ehmke and Campbell, 2014).	- With account for the long-term effect on customer behaviour, should retailers accept all requests or allow certain order rejection for HD and BOPS services?

Planning areas	Research themes	Contributions from existing literature	Research agendas for future work
		Decisions on the products types and service areas that should be offered BOPS services (Bhatnagar and Syam, 2014; Cao et al., 2016; Gao and Su, 2017; Jin et al., 2018).	
	Delivery services optimisation	Joint optimisation of pricing and time slot management for HD services (Asdemir et al., 2009; Campbell and Savelsbergh, 2006; Yang and Strauss, 2017; Yang et al., 2014) Design of incentives (e,g., differentiated pricing) to steer customers to switch from HD services to BOPS services (Rabinovich et al., 2018).	- How to decide the various delivery/collection service offerings (i.e., order receipt window)? - How to design a menu of different delivery/collection services (i.e., choice of HD or BOPS services for different speed with different time slots) and the corresponding prices to maximise revenue? - Should retailers make all delivery/collection services available uniformly or discriminately across product ranges? - Under what conditions should retailers steer customers to switch from HD services to BOPS services and vice versa?
	Vehicle routing problems	Decisions on how to combine traditional store replenishment with online orders for BOPS services in the same delivery truck to minimise the transportation costs (Nekoiemehr et al., 2018; Paul et al., 2019). Decisions on when to dispatch online orders to balance the benefits from order consolidation and the potential cost from order expedition (Wei et al., 2017; Zhang et al., 2019b).	- When to dispatch shipments of the combination of online orders and store replenishment orders with various delivery deadlines? - Explore the benefits of using the same delivery truck to deliver online and store replenishment orders and to pick up returned online orders from the same store for further processing, which is a delivery and pick-up VRP problem.
Returns management	Returns configuration	Decision on the set of retail stores that should offer BORS services (Mahar et al., 2914). Design of returns network for online retailers that collect returns from customers to centralised RCs via CEP (Min et al., 2006).	- How should returns network be designed taking account of different product types with different lifecycle? - Should all stores accept online returns or only those that stock and sell the returned items? - If BORS orders can be processed in stores, what is the optimal number and location of RCs? What set of retail stores should be utilised to process the BORS products?
	Returns processes	Decision on the number of BORS products to be shipped from stores to DCs for further	- How should the flow of returned products be optimised within the network? For example, the assignment of

Planning areas	Research themes	Contributions from existing literature	Research agendas for future work
		processing (Dijkstra et al., 2019). Decision on which returned items should be kept in stores and which should be sent back to DCs or RCs at what cadence when offering BORS services (Radhi and Zhang, 2018). Investigate the impacts of returns policy (i.e., with or without BORS options) and returns processing structure (i.e., centralised or decentralised) on inventory performance (Muir et al., 2019).	returned products across various returns processing facilities, such as DCs and RCs or even suppliers for remanufacturing.
	Joint pricing and inventory optimisation taking account of returns	Decisions on how retailers should adjust their ordering and product pricing decisions when BORS option is offered (He et al., 2020; Radhi and Zhang, 2018; Radhi and Zhang, 2019; Zhang et al., 2018).	- What are the optimal prices for the product and returns option and how much inventories to order when considering anticipated returns?
	Returns policies	Decision on the charges for online returns via CEP (BORS option is often assumed to be free) (Chen and Chen, 2017; Jin et al., 2019; Nageswaran et al., 2019). Decision on the optimal returns period, i.e., time window during which returned products can be accepted and refunded (Difrancesco et al., 2018).	- When should retailers offer BORS options? - How should decisions on the returns fee for different returns options be made dynamically to maximise profit?

2.9 Justification of research questions in the thesis

Having reviewed the extant literature in the domain of omni-channel retail supply chain, we have highlighted a wide range of problems that could be investigated for future work (see Table 2). However, given limited research time, we focus our attention on the three main problems presented in section 1.4. The rationale of choosing these research problems is three-fold: their relevance to current challenges in omni-channel, our own interests, and our ability to make progress in developing relevant and insightful models given the time and resources available.

BOPS services have been promoted as the antidote for retailers to fight against pure online retailers such as the behemoth Amazon (CNN, 2019). BOPS service has continued to grow in popularity as a win-win solution for both retailers and customers (GlobalData, 2018b). To fully exploit the benefits of BOPS services, it is imperative for retailers to contemplate the details of BOPS order fulfilment operations and strike the right balance between cost and service. BOPS services are of interest to retailers in today's world and have also sparked

interests among researchers. However, due to this being an emergent phenomenon, existing studies on BOPS services are limited to strategic decisions such as the types of products suitable for BOPS services, the radius of service area for BOPS services, and the number of retail stores needed to provide BOPS services (as highlighted in the 'Facility location for online order fulfilment' row and 'Order acceptance mechanisms' row in Table 2). No work has been done for the strategic decision on BOPS services order receipt window (as highlighted in the 'Delivery services optimisation' row in Table 2) and the operational decisions on planning and organising fulfilment activities (e.g., the timing to start picking and the level of picking rate needed to process BOPS orders) (as highlighted in the 'Order allocation' row in Table 2). These decisions are crucial for retailers to unlock the benefits of BOPS services and to maintain and improve competence in the era of omni-channel. The importance and the lack of understanding of these decisions have motivated us to investigate the first and second of our research problems. Specifically, our first research problem fills the gap on the lack of decision models for the operational decisions on BOPS order fulfilment (i.e., the schedule of picking operations and the level of picking rate) and the second research problem sheds light on the strategic decision on the BOPS service offerings (i.e., order receipt window).

Online order fulfilment has triggered many conversations in the research community. It is critical since it affects customers' perception on service quality and accounts for a large share of the total operating cost for retailers (Boyer and Hult, 2006; Xing et al., 2010). Existing studies have examined both static and dynamic order fulfilment policies and the problem contexts include both pure online retailing and multiple channel retailing, in which conventional retail stores can also handle online orders. However, our literature review in section 2.5.2 has revealed that most existing studies assume that all stores are readily used for online fulfilment. Additionally, these studies mainly focus on HD demand without incorporation of BOPS demand (as highlighted in the 'Facility location for online order fulfilment' row in Table 2). In practice, retailers rarely adopt a complete decentralised fulfilment strategy to use all retail stores for handling online orders. The decision on the number of stores used for online fulfilment reflects the centralisation level of fulfilment operations and hinges on the trade-offs between different types of costs. For example, high inventory costs prefer less stocking locations (e.g., a higher centralisation level) according to the square root law (Maister, 1976) and high transportation costs desire more stocking locations (e.g., a lower centralised level) with shorter average distance to customers. Given the importance of the online fulfilment problem, the potential contribution of exploring the strategic decision on the number of stores deployed for online fulfilment with consideration for both HD and BOPS demand justifies the third problem in the thesis.

Chapter 3

Best Performance Frontier for Buy-Online-Pickup-In-Store order fulfilment

In this chapter, we examine store-picking operations for same day BOPS services by developing analytical models and deriving Best Performance Frontiers (BPFs) for single wave and multi-wave order picking scenarios. New relationships, propositions, and results are presented to determine the minimum picking rate needed in stores to guarantee a target service level, the number of picking waves a retailer should launch in an ordering cycle, and the timing of picking waves. We also examine demand surge scenarios with different order arrival rates in an ordering cycle. Insights from, and implications of the results are discussed for retailers that seek to benchmark their current BOPS performances and understand how to schedule and improve the picking of online orders in conventional retail stores and the picking rates needed to guarantee a desired service level for online customers.

3.1 Introduction

Retailing has entered the omni-channel era in which customers demand a channel-agnostic and seamless shopping experience across physical stores, mobile, online and other platforms. Both practitioners and academic researchers acknowledge that omni-channel is the future of retailing (Deloitte, 2017; EY, 2015; Saghiri et al., 2017; Sopadjieva et al., 2017).

One of the strongest omni-channel trends is the provision of BOPS services, which integrate online and offline operations by allowing customers to place orders online and collect them in their chosen stores (Chen et al., 2016). For retailers, BOPS services relieve them of the responsibility for managing last mile deliveries. It also allows them to use inventories inside stores within their networks for online fulfilment (Ishfaq and Raja, 2018). Additionally, it provides an opportunity for retailers to cross-sell and cross-promote products, which leads to revenue growth (Cao and Li 2015). Importantly, traditional retailers can leverage the capabilities of their existing store networks and distribution systems to benefit their customers. As noted by Forbes (2019b), retailers such as Walmart are using their vast network of physical stores to provide BOPS services to compete with the pure online retailers such as Amazon.

For customers, BOPS services are more flexible than home deliveries as they do not involve having to wait at home. They may also be more economical as most of the BOPS services offered by retailers are provided free (Forrester Research, 2018). A study by Gibson et al. (2016) showed that 67% of customers surveyed had used a BOPS service. BOPS services

are particularly popular during holiday shopping seasons when deliveries may be problematic. US shoppers increased their click-and-collect purchases by 50% on Black Friday and by 65% on Cyber Monday in 2018 compared to the previous year (McKinsey, 2019). BOPS services are offered by many retailers (Forrester, 2014) and are likely to become more prevalent as the landscape for omni-channel retailing is highly competitive (Ishfaq and Raja, 2018). Amazon now provides same or next day delivery, which raises the bar for online fulfilment speed. Omni-channel retailers can potentially fight back with same or next day BOPS fulfilment services or even fulfilment within a few hours using their well-established store networks (The Guardian, 2016).

Notwithstanding its popularity, fulfilling BOPS orders in stores is very challenging for retailers. Inaccurate store inventories hinder cross-channel fulfilment and also increase stock-out possibilities for walk-in customers (Forrester Research, 2014). In addition, in-store picking is a manual operation. Insufficient store staffing levels can lead to a low order fill rate for online customers (Mahar and Wright, 2017), while overstaffing burdens the retailer with higher costs. It is challenging for omni-channel retailers to leverage their conventional stores for online fulfilment while balancing the need to also serve walk-in customers. Retailers need to consider the timing of fulfilment activities without clogging the shopping store aisles and disrupting the shopping experience of walk-in customers (Enders and Jelasssi, 2009; JDA, 2016). Retailers need to optimise in-store picking operations to fulfil BOPS orders effectively and profitably (JDA 2016). Only a limited amount of research has been done to address problems associated with fulfilling BOPS orders within retail stores. Here we investigate and develop models for the in-store picking of BOPS orders.

Many retailers use the approach of picking online orders from a retail store to fulfil orders received online, particularly when same or next day collection is offered. Much of the literature on picking operations in warehouses or DCs does not apply directly to in-store picking due to the distinctive features of the store picking setting. In-store order fulfilment activities are mainly manual with low fulfilment efficiencies (Hübner et al., 2016b) and typically each order is assigned to a specific picker without order splitting. For example, each picker is typically responsible for less than 6 orders in the stores of Tesco before receiving the next picking list (Boyer et al., 2003). Orders are picked one by one or in small batches to be ready for collection for the advertised collection time in stores. In contrast, in warehouses and DCs, a multi-item order may be picked by several pickers simultaneously with each picker picking a specific item for a group of orders in batch picking (Parikh and Meller, 2008). Orders are then sorted and packed, ready for delivery. A further difference between in-store and warehouse or DC picking operations lies in the fluctuation of picking

rate. Picking rate is more unstable in stores with walk-in customers, as pickers may have other jobs within the store and picking activities may be halted during peak hours to avoid congestion. Thus, there are many open questions in relation to optimal in-store picking operations for online orders.

A specific attribute of the BOPS problem is the presence of pre-specified hard deadlines for the ordering window and the advertised pick up time at which an ordered item will be available. Although there is a significant number of studies on minimizing travel distance and throughput time in the warehouse fulfilment literature, only a very limited amount of work has been carried out on optimising fulfilment performance in the presence of specific delivery deadlines. In fact, there are very few studies on order fulfilment activities in general with a specified due date (Yan et al., 2010; Nekoiemehr et al., 2018). Doerr and Gue (2013) proposed and justified a performance metric they termed 'Next Schedule Deadline' (NSD) for deadline-oriented picking in the context of warehouse fulfilment systems. NSD measures the fraction of online orders arriving between two consecutive cut-off times that are fulfilled before a pre-defined deadline. A picking wave is defined as the release of a batch of online orders for fulfilment (Çeven and Gue, 2017). Çeven and Gue (2017) determine the optimal release times and the number of waves to fulfil online orders against daily deadlines treating the picking rate as a known parameter and focusing on the optimal wave release time to maximise NSD. Their research is based on a warehouse picking setting where the order cut-off time is assumed to be the same as the deadline. We use the NSD metric to answer different questions in the context of store picking with an advertised order cut-off time and an advertised pickup time, which provides a deadline for the completion of order fulfilment operations. For BOPS services we consider both the picking wave release time(s) and the picking rate to be unknown variables to be determined.

From the reviewed related literature, it is clear there is a paucity of studies examining in-store picking activities for the BOPS mode of retail fulfilment. To fill this gap, we develop analytical models to meet a target service level for in-store BOPS demand fulfilment whilst seeking to minimise the picking rate. Retailers seek to minimise the required picking rate because in-store picking activities are predominantly manual and the picking rate is directly related to the staffing level, which translates into store labor costs. We derive BPFs to attain the best combination of the time to start picking and the minimum picking rate required for any pre-defined service level. We investigate single and multiple wave scenarios. We determine the number of picking waves required in a day and their timings to meet BOPS demand through in-store fulfilment. The analysis also determines the time for a retailer to commence order picking and investigates whether a retailer should pick orders once per cycle

or multiple times throughout the cycle. Our research is one of the first attempts to study operational level decisions for in-store picking.

3.2 Best Performance Frontiers (BPFs) for BOPS order fulfilment

We first describe the in-store picking problem for a BOPS service and then identify a BPF for BOPS order fulfilment through single and multi-wave picking.

3.2.1 Problem description

We consider the setting where an omni-channel retailer promises customers to have their orders ready for collection after a specific deadline if they place their orders before the cut-off time. For example, the British supermarket, Sainsbury's, currently offers customers a same day BOPS service: if you order before 12am, you can collect after 4pm. This cycle repeats daily so that the orders that should be available for collection by today's deadline (4pm today) were placed sometime between yesterday's cut-off time (12am yesterday) and today's cut-off time (12am today) (see Figure 11).

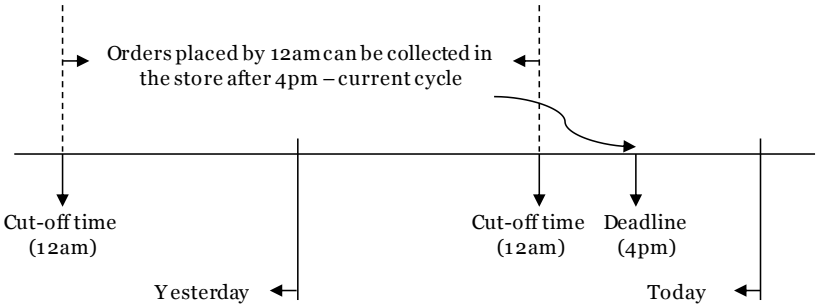


Figure 11. Illustration of the same day BOPS service promise by Sainsbury's.

The retailer has to pick all orders that have accumulated for the store during a cycle of length T. The order cut-off time and the promised collection deadline for the current cycle of orders are denoted by t_c and t_d. Hence, the current cycle of orders is defined by the time that elapses between two consecutive order cut-off times, $t_c - T$ and t_c.

Most retailers will aspire to meet their BOPS promise with a high level of service reliability. We therefore use the NSD metric to measure service level for BOPS demand (Doerr and Gue, 2013):

$$NSD = \frac{Orders\ that\ arrive\ between\ t_c - T\ and\ t_c\ and\ have\ been\ picked\ by\ t_d}{Orders\ that\ arrive\ between\ t_c - T\ and\ t_c} \tag{3.1}$$

As retailers seek to guarantee online fulfilment service reliability, the minimum acceptable value of NSD will be high, typically in excess of 0.9 (Ocado, 2019).

We assume the daily order demand rate in any location, λ, to be constant. This is a reasonable initial assumption given that BOPS systems are typically high volume with incoming online demand accumulating over time. We consider different levels of demand in different periods later in the analysis (Section 3.3). We assume the in-store picking rate μ to be constant and that demand can be met, i.e., $\mu \geq \lambda$ always holds. Retailers are interested in keeping a constant staffing level and therefore constant picking rate, as hiring and firing employees will incur costs. With a specific picking rate, the retailer may or may not finish all the accumulated orders in the current cycle before the deadline.

Once picking commences, the retailer may seek to pick orders in a single wave or may perform multiple waves of picking. When orders are picked in multiple waves, we assume that these are immediately sequential (see Section 3.4), i.e., the next picking wave starts immediately after the end of the current wave. We assume that there are no restrictions on store operations hours and that picking activities can occur round the clock. We also assume that orders are picked on a FCFS basis and the processing time for each order is the same.

Figure 12 illustrates three different response policies to a constant order arrival process (demand rate λ): the dashed orders curve shows the cumulative demand over time. Three picking waves are released at times $t_w - T$, t_w and $t_w + T$. Response 1 processes online orders with a rate equal to the demand rate ($\mu = \lambda$); response 2 picks orders at a faster rate than the demand (the picking duration $T_{\mu 2}$ is shorter than cycle T), and response 3 picks orders at an infinite rate ($\mu = \infty$). Figure 12 does not incorporate cut-off times and deadlines. Our analysis in the next sections aims at quantifying how different response policies (in terms of picking rate μ and wave release time t_w) impact on service level for given cut-off times t_c, deadlines t_d and demand rate λ. Table 3 lists the notation we use for this in-store picking problem.

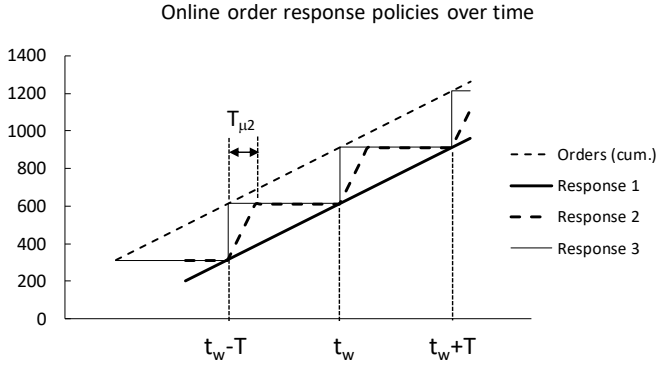


Figure 12. Different response policies to a constant order arrival process.

Table 3 Notation for the single and multi-wave store picking problem.

Notation	Explanation
N	The number of picking waves for a cycle of orders
T	The cycle length (one day, i.e., $T = 1$ for BOPS)
t_c	The order cut-off time
t_d	The deadline when BOPS orders should be ready for customers to collect
t_f	The finishing time of the (last) picking wave for a cycle of orders
t_w	The starting time (i.e., release time) for picking using a single wave
t_w^*	The starting time for picking using a single wave on the BPF
$t_{w,j}$	The starting time for picking the *j*-th wave (j= 2, … N)
$t_{w,j}^*$	The starting time for picking the *j*-th wave (j= 2, … N) on the BPF
$t_{w,1}'$	The starting time for picking the first wave with *N-1* picking waves
t^*	Demand rate switch point in the case with two different arrival rates
λ	BOPS demand arrival rate
μ	The picking rate in stores using a single wave
μ^*	The minimum picking rate in stores using a single wave
μ_N	The picking rate in stores with *N* picking waves for a cycle of orders
μ_N^*	The minimum picking rate in stores with *N* picking waves

3.2.2 Single wave picking

In single wave picking, a picking wave is released at time t_w to process all λT orders that have accumulated in (t_w-T, t_w] with a picking rate $\mu \geq \lambda$. The orders that contribute to *NSD* are those that arrive in order cycle (t_c-T, t_c] and that are picked before the deadline t_d with $t_c \leq t_d < t_c$+T. For a constant picking rate μ, picking activities will finish at time $t_f = t_w + \lambda T/\mu$. Either one picking wave or two picking waves from successive cycles contribute to the *NSD*. The number of orders from the current cycle that have been picked before deadline t_d are attributed to one picking wave only, starting in the current cycle at t_w (e.g. see Figure 13(a) and 13(b)). Alternatively, the orders are attributed to two picking waves: one picking wave starting at t_w and another picking wave starting either at $t_w - T$ or at $t_w + T$. This can only happen if the retailer starts picking for the current order cycle too late, i.e., $t_c < t_w \leq t_d$, therefore the picking wave starting at $t_w - T$ contributes to NSD, see Figure 13(c) and 13(d)) or too early, $t_c - T < t_w \leq t_d - T$, and therefore the picking wave starting at $t_w + T$ contributes to NSD. Due to the cyclical nature of the picking problem, the scenario with picking waves at t_w and t_w+T is equivalent to the scenario with picking waves at t_w-T and t_w and does not require a separate analysis. We discuss the remaining two cases separately.

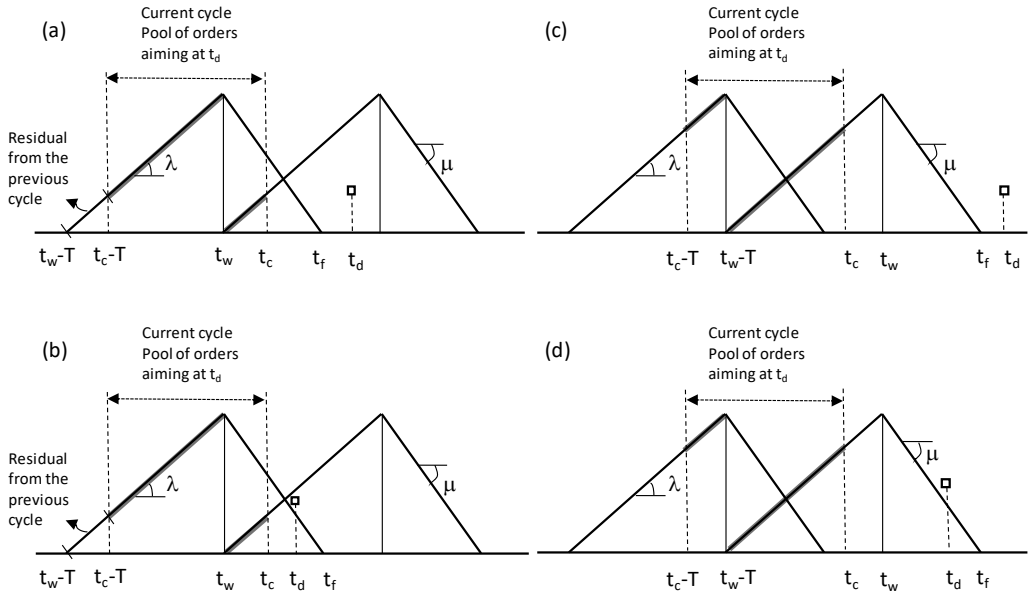


Figure 13. Single wave picking: only the picking wave in the current cycle contributes to NSD in (a) and (b), and both the picking waves in the current and previous cycle contribute to NSD in (c) and (d).

Case 1: Only the picking wave in the current cycle contributes to NSD, i.e., $t_d - T < t_w \leq t_c$.

In this case, if $\mu \geq \lambda T/(t_d - t_w)$, then $t_f = t_w + \lambda T/\mu \leq t_d$, which means picking activities finish before the deadline (see Figure 13(a)). The number of orders arriving in the current cycle is λT. Among them, those arriving from $t_c - T$ to t_w have been fulfilled, while those arriving from t_w to t_c will not be picked before the deadline. Thus, for this case:

$$NSD = \frac{(t_w - (t_c - T))\lambda}{\lambda T} = 1 - \frac{t_c - t_w}{T}$$

If $\mu < \lambda T/(t_d - t_w)$, then $t_f > t_d$ and picking activities finish after the deadline (see Figure 13(b)). By the deadline, $(t_d - t_w)\mu$ orders will have been fulfilled. Among these picked orders, $((t_c - T) - (t_w - T))\lambda$ orders are 'residual' from the previous cycle. Therefore, the number of orders that arrive in the current cycle and are picked before the deadline is $(t_d - t_w)\mu - ((t_c - T) - (t_w - T))\lambda$. Thus:

$$NSD = \frac{(t_d - t_w)\mu - ((t_c - T) - (t_w - T))\lambda}{\lambda T} = \frac{(t_d - t_w)}{\lambda T}\mu - \frac{t_c - t_w}{T}$$

In summary, when only the current picking wave contributes to NSD (i.e., $t_d - T < t_w \leq t_c$):

$$NSD = \begin{cases} \frac{(t_d - t_w)}{\lambda T}\mu - \frac{t_c - t_w}{T} & \text{for} \quad \lambda \leq \mu < \frac{\lambda T}{(t_d - t_w)} \\ 1 - \frac{t_c - t_w}{T} & \text{for} \quad \mu \geq \frac{\lambda T}{(t_d - t_w)} \end{cases} \tag{3.2}$$

Based on Eq. (3.2), NSD increases linearly in picking rate μ until a certain point (i.e., $\mu = \frac{\lambda T}{(t_d - t_w)}$) after which it remains at a constant level equal to $1 - \frac{t_c - t_w}{T}$. Since $\frac{t_c - t_w}{T} \geq 0$, $NSD \leq 1$. Clearly, the retailer can only achieve a 100% service level ($NSD = 1$) when picking starts at the order cut-off time ($t_w = t_c$) with a minimum critical picking rate $\mu = \frac{\lambda T}{(t_d - t_c)}$. If the picking rate is not sufficiently high, the retailer can never achieve a 100% service level ($NSD = \frac{(t_d - t_w)}{\lambda T}\mu - \frac{t_c - t_w}{T} < 1$). At the lowest picking rate, i.e., $\mu = \lambda$, $NSD = \frac{(t_d - t_c)}{T}$ and is independent of the wave release time t_w.

Case 2: Both the picking waves in the current cycle and previous cycle contribute to NSD, i.e., $t_c < t_w \leq t_d$.

In this case, if $\mu \geq \lambda T/(t_d - t_w)$, then $t_f \leq t_d$ and picking activities finish before the deadline (see Figure 13(c)). Since all orders arriving from $t_c - T$ to t_c will have been picked by the deadline t_d, $NSD = 1$.

For $\mu < \lambda T/(t_d - t_w)$, picking activities finish after the deadline (see Figure 13(d)). Orders arriving from $(t_c - T)$ to $(t_w - T)$ have been picked by the wave from the previous cycle, while the wave in the current cycle has processed $(t_d - t_w)\mu$ orders. Therefore, if $\lambda(t_c - (t_w - T))/(t_d - t_w) \leq \mu < \lambda T/(t_d - t_w)$, the picking wave in the current cycle will fulfil all orders arriving from $(t_w - T)$ to t_c by the deadline and therefore $NSD = 1$.

Otherwise, the number of orders arriving in the current cycle and that have been picked by the deadline is $(t_d - t_w)\mu + ((t_w - T) - (t_c - T))\lambda$ and

$$NSD = \frac{(t_d - t_w)\mu + ((t_w - T) - (t_c - T))\lambda}{\lambda T} = \frac{(t_d - t_w)}{\lambda T}\mu + \frac{t_w - t_c}{T}$$

In summary, when both the picking waves in the current and previous cycle contribute to NSD (i.e., $t_c < t_w \leq t_d$):

$$NSD = \begin{cases} \frac{(t_d - t_w)}{\lambda T}\mu + \frac{t_w - t_c}{T} & \text{for} \quad \lambda \leq \mu < \frac{\lambda(t_c - t_w + T)}{(t_d - t_w)} \\ 1 & \text{for} \quad \mu \geq \frac{\lambda(t_c - t_w + T)}{(t_d - t_w)} \end{cases} \tag{3.3}$$

From Eq. (3.3) it is clear that NSD increases linearly in picking rate μ until a certain point (i.e., $\mu = \frac{\lambda(t_c - t_w + T)}{(t_d - t_w)}$) after which it becomes constant at a level equal to 1. At the lowest picking rate, i.e., $\mu = \lambda$, $NSD = \frac{(t_d - t_c)}{T}$ and is independent of the wave release time t_w.

3.2.3 Best Performance Frontier (BPF) for BOPS with single wave picking

We use a numerical example to demonstrate how the results from Section 3.2.2 can be used to derive the BPF for BOPS services with a specified order cut-off time and deadline. Consider Tesco's current service promise for its same day BOPS offering: orders placed before 9am will be ready for collection after 12noon. For convenience, we consider a two-day planning horizon [0, 2] with day 1 (yesterday 0am-12pm) = [0, 1] and today (0am-12pm) = [1, 2]. Today's order cut-off time and deadline are then $t_c = 1 + 9/24 = 1.375$ and $t_d = 1 + 12/24 = 1.5$. Assume the demand rate at a particular location is λ = 300 orders per day. In Case 1 when $t_d - T < t_w \leq t_c$, we investigate the minimum picking rate needed and the maximum service level that can be achieved when the retailer decides to pick today at 6am ($t_w = 1 + 6/24$), 8am ($t_w = 1 + 8/24$) and 9am ($t_w = 1 + 9/24$), respectively. The computation is based on Eq. (3.2) with $\lambda = 300$ and $T = 1$ and the results are illustrated in Figure 14.

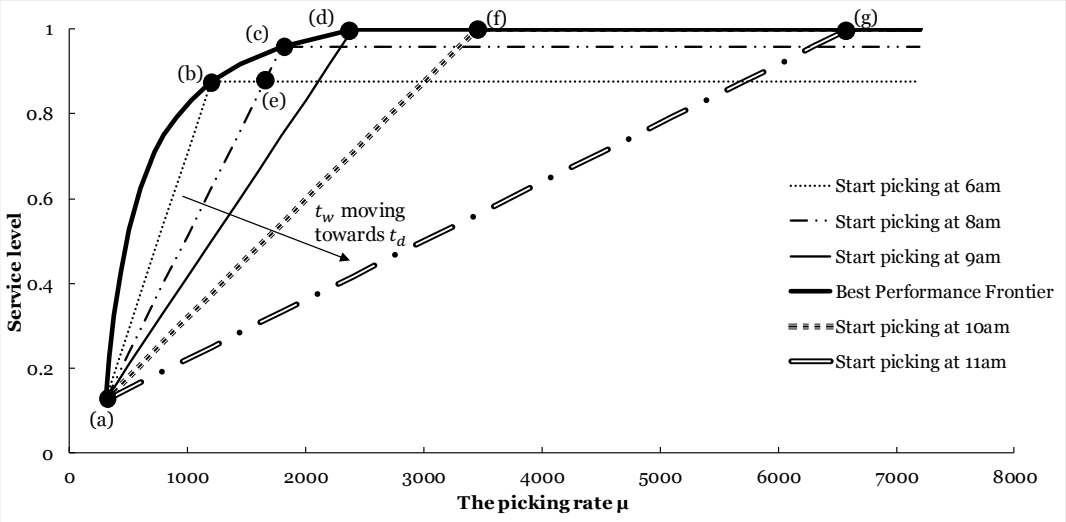


Figure 14. BPF for single wave picking with pre-specified order cut-off and deadline times – a same day BOPS example.

Point (a) in Figure 14 shows that the retailer can achieve a service level of $(t_d - t_c)/T = (1.5 - 1.375)/1 = 0.125$ at the lower boundary value for the picking rate (i.e.,$\mu = \lambda = 300$), independent of the wave release time. The picking rate cannot be less than the demand rate as backlogs would build up constantly. When the retailer increases the picking rate from its lower boundary value, the service level increases linearly up to a critical point ($\mu =$

$\lambda T/(t_d - t_w)$) at which the service level reaches a maximum value $(1 - (t_c - t_w)/T)$ and remains constant for any further increase in the picking rate. For example, in the case where the retailer starts picking exactly at the cut-off time of 9am, i.e., $t_w = t_c$, any increase in the picking rate below a level of $\mu = \lambda T/(t_d - t_w) = 300/(1.5 - 1.375) = 2400$ will increase the service level. A picking rate of 2400 is the critical point where the service level reaches its highest point, shown by point (d) in Figure 14, where service level is 1. This critical point is the minimum picking rate μ required to achieve the maximum service level. Clearly, if the retailer starts picking earlier than 9am, the lower the maximum service level that can be achieved, but also the lower the critical picking rate needed to achieve that maximum service level. For example, if the retailer starts picking at 6am, the maximum service level attainable is 87.5% $(1 - (1.375 - 1.25)/1 = 0.875)$ and the corresponding minimum picking rate needed is only $\mu = 1200$ (point (b) in Figure 14). When aiming at high service levels, retailers should wait and schedule the wave release time t_w close to the cut-off time t_c and process orders with a higher minimum picking rate.

Connecting all the critical points for different wave release times yields the curve shown in Figure 14 (solid line), which is the BPF. In a BOPS system operating with pre-defined cycle length T, order cut-off time t_c, deadline t_d and demand rate λ, any point on the BPF provides the retailer with the minimum picking rate μ and the corresponding wave release time t_w for a given service level. For example, if the retailer sets a target service level of 95.8% (i.e., $NSD = 0.958$), Eq. (3.2) yields a wave release time of 8am $(t_w = t_c - (1 - NSD)T = 1.375 - (1 - 0.958)1 = 1.333)$ and a corresponding minimum picking rate of $\mu = \lambda T/(t_d - t_w) = 300/(1.5 - 1.333) = 1800$, shown by point (c) on the BPF in Figure 14.

The BPF informs the retailer about the best picking rate and wave release time combination for any desired service level. Retailers can use the BPF as a diagnostic tool and benchmark their current operations against the best performance that can be attained. For example, if a retailer releases the picking wave at 8am and operates with a picking rate of 1650, a service level of 87.5% will be achieved, shown by point (e) in Figure 14. Releasing the wave earlier (e.g. 6am) will achieve the same service level but will require a picking rate of only 1200, as shown by point (b) in Figure 14. Keeping the wave release time at 8am but increasing the picking rate to 1800 would increase the service level to 95.8% (point (c) in Figure 14). On the other hand, the BPF shows that for a picking rate of 1650 a wave release time of $t_w = t_d - \lambda T/\mu = 1.5 - 300/1650 = 1.318$ (=7.38am) will achieve a service level of 94.3% $(NSD = 1 - (t_c - t_w)/T = 1 - (1.375 - 1.318)/1 = 0.943)$. This example shows that a small (22 minutes) change in wave release time could result in a rather large improvement

in service performance. More generally, retailers should moderate their service level expectations when operating well outside the BPF.

Based on Eqs. (3.2) and (3.3), the wave release time can be expressed by the picking rate at the critical point. Therefore, the service level NSD along the BPF can be expressed as a univariate function of the picking rate:

$$NSD = \begin{cases} 1 + \frac{t_d - t_c}{T} - \frac{\lambda}{\mu} & \text{for} \quad \lambda \leq \mu < \frac{\lambda T}{(t_d - t_c)} \\ 1 & \text{for} \quad \mu \geq \frac{\lambda T}{(t_d - t_c)} \end{cases}.$$

Clearly, NSD increases in the picking rate before it peaks at 100% and levels off. The gradient of the BPF provides further insights on the trade-offs between the service level and the management of picking operations. From Eq. (3.2), we can see that $NSD = 1 - (t_c - t_w)/T$ and $\mu = \lambda T/(t_d - t_w)$ on the BPF. The first derivative of NSD with respect to t_w is $\frac{dNSD}{dt_w} = \frac{1}{T}$, meaning that if the retailer wants to improve the service level by Δ, picking activities should be scheduled ΔT time units later (i.e., increase t_w by ΔT). This in turn leads to an increase in the picking rate of $\frac{\lambda T^2}{(t_d - t_w)(t_d - (t_w + \Delta T))}\Delta$ $(= \frac{\lambda T}{(t_d - (t_w + \Delta T)} - \frac{\lambda T}{(t_d - t_w)})$. Generally, $\frac{\lambda T^2}{(t_d - t_w)(t_d - (t_w + \Delta T))} \gg 1$ holds since demand rate λ is very large compared to time related parameters such as t_d. That is to say, the retailer can enhance the service level by commencing picking activities later. However, a much higher picking rate is needed to process orders in a narrower window. These relationships can be used to quantify the adverse impact on the picking rate and therefore on labor costs when a retailer seeks to improve the service level.

Figure 14 also illustrates the relationship between the picking rate and service level for case 2 when the retailer starts picking at 10am and 11am, respectively. If the retailer moves the wave release time away from the order cut-off time towards the deadline, the order processing window becomes narrower and the picking rate required to achieve a target service level increases strongly. For example, to guarantee a 100% service level, if the retailer starts picking at 10am, a picking rate of 3450 is required, shown by point (f) in Figure 14 and if picking is postponed until 11am, a minimum picking rate of 6600 is needed, shown by point (g) in Figure 14. Likewise, the t_w=10am and t_w=11am performance curves also start at the service level $(t_d - t_c)/T = 0.125$ for $\mu = \lambda = 300$. Although a 100% service level is achievable when picking starts after the cut-off time, this requires very high picking rates, which may not be a comfortable or recommended operational situation for many

retailers. In the remainder of this chapter, we therefore focus on situations where the wave release time(s) for the current order cycle is (are) in the interval $(t_d - T, t_c]$.

3.2.4 Key relationships for single wave in-store picking

In the following two propositions, we formalize the relationship between service level, picking rate, and the wave release time along the BPF for single wave BOPS in-store picking.

Proposition 1: If a retailer sets a service level β for same day BOPS service, the minimum picking rate needed and the corresponding wave release time in single wave picking are on the BPF, satisfying $t_w{}^* = t_c - (1-\beta)T$ and $\mu^* = \lambda T/(t_d - t_c + (1-\beta)T)$.

Proof:

The results follow immediately from Eq.(3.2): when $1 - \frac{t_c - t_w}{T} = \beta$, the wave release time on the BPF is $t_w{}^* = t_c - (1-\beta)T$. This can be achieved as long as $\mu \geq \lambda T/(t_d - t_w)$. Since the retailer seeks to minimise the picking rate, the minimum picking rate $\mu^* = \lambda T/(t_d - t_w{}^*) = \lambda T/(t_d - t_c + (1-\beta)T)$. This minimum picking rate is the critical point to achieve a pre-defined service level. Any picking rate μ lower than μ^* only achieves a service level $\frac{(t_d - t_w)}{\lambda T}\mu - \frac{t_c - t_w}{T} < \beta$, whereas for the same $t_w{}^*$ any picking rate higher than μ^* only achieves service level β. Similarly, releasing the wave before $t_w{}^*$ can only result in a lower maximum service level, whereas releasing the wave after $t_w{}^*$ can only achieve the same service level β with a higher picking rate.

Note that if a retailer aspires to achieve a 100% service level, i.e., $\beta = 1$, picking should be commenced at the exact order cut-off time, i.e., $t_w{}^* = t_c$. Otherwise, the retailer can start picking before the order cut-off time for a lower service level.

Proposition 2: The retailer will always finish picking activities in a cycle at the deadline if operating on the BPF.

Proof:

Substituting $t_w{}^* = t_c - (1-\beta)T$ into $\mu^* = \lambda T/(t_d - t_c + (1-\beta)T)$, yields $t_d = t_w{}^* + \lambda T/\mu^*$, which is equal to the finishing time t_f.

3.3 Single wave picking with two different arrival rates

Here we consider the scenario where a retailer experiences two different arrival rates, λ_1 and λ_2, in the current cycle of orders. Specifically, we assume that there is a specific point t^*, at which the demand arrival rate λ_1 experienced in $[t_c - T, t^*]$ shifts to λ_2 for the remainder of the cycle $[t^*, t_c]$. In some retail contexts, many customers will seek to place their orders well before the order cut-off time, i.e., there will be an early demand surge ($\lambda_1 > \lambda_2$), to reduce their risk of missing the cut-off time and failing to qualify for a same day BOPS service. In other retail contexts, many customers may procrastinate and place orders just before the order cut-off time, i.e., a late demand surge ($\lambda_1 < \lambda_2$) (see Figure 15). Keeping other model assumptions the same, we show how the picking rate μ and wave release time t_w on the BPF are affected by the timing and duration of demand surges in a cycle for the single wave case.

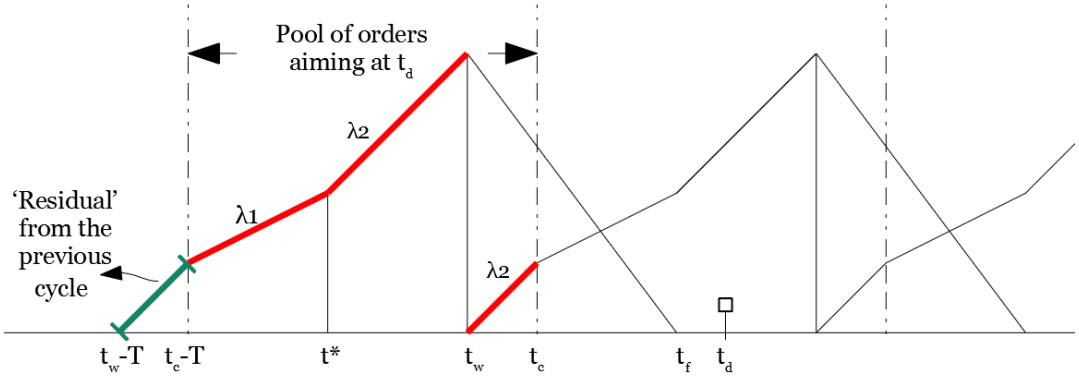


Figure 15. Single wave picking model with a late demand surge.

Using the same analytical approach, with a given service level β, the expressions for the wave release time $t_w{}^*$ and picking rate μ^* are as follows:

$$t_w{}^* = \begin{cases} \beta\left(1-\frac{\lambda_2}{\lambda_1}\right)(t^*-t_c)+\beta T+t_c-T & \text{for } t_w \le t^* \le t_c \\ t_c-(1-\beta)\left(\frac{\lambda_1}{\lambda_2}(t^*+T-t_c)+t_c-t^*\right) & \text{for } t^* < t_w \le t_c \end{cases} ; \quad \mu^* = \frac{(\lambda_1-\lambda_2)(t^*-t_c)+\lambda_1 T}{(t_d-t_w)} \quad (3.4)$$

Readers are referred to Appendix A.1 for the derivation of the service level and critical points for two different demand rates within a cycle. When comparing Eq. (3.4) with Eq. (3.2), the difference in the critical points for this scenario and the constant demand rate scenario is not straightforward. Specifically, the difference in the critical points depends on the value of t^* (i.e. the demand rate switch point) and the difference between two arrival rates λ_1, λ_2. We conduct numerical analyses below to show how the relationships between the two arrival rates λ_1, λ_2 and the change in demand rate switch point t^* impact on the minimum picking rate.

3.3.1 The timing of demand surge

We examine whether an early or a late demand surge is advantageous for the retailer, i.e., which situation requires a lower picking rate. As before, our example is based on Tesco's current service promise for its same day BOPS offering: orders placed before 9am will be ready for collection after 12 noon, and consider a two-day planning horizon [0, 2] with day 1 (yesterday 0am-12pm) = [0, 1] and today (0am-12pm) = [1, 2]. Today's order cut-off time and deadline are $t_c = 1.375$ (9am today) and $t_d = 1.5$ (12noon today). We fix the service level $\beta = 0.95$. We set $\lambda_1 = 600, \lambda_2 = 300$ in an early surge scenario and $\lambda_1 = 300, \lambda_2 = 600$ in a late surge scenario. The values set for the demand rate switch point t^* are {0.5, 0.6, 0.7, 0.8, 0.875, 0.9, 1, 1.1, 1.2, 1.3, and 1.35}. For example, $t^* = 0.875$ (i.e. 9pm yesterday) means demand rate is 300 orders/day from 9am to 9pm yesterday and 600 orders/day from 9pm yesterday to 9am today in a late surge scenario. For each t^*, we obtain and summarise the results (see Figure 16) for both $t_w{}^*$ and μ^* based on Eq. (3.4).

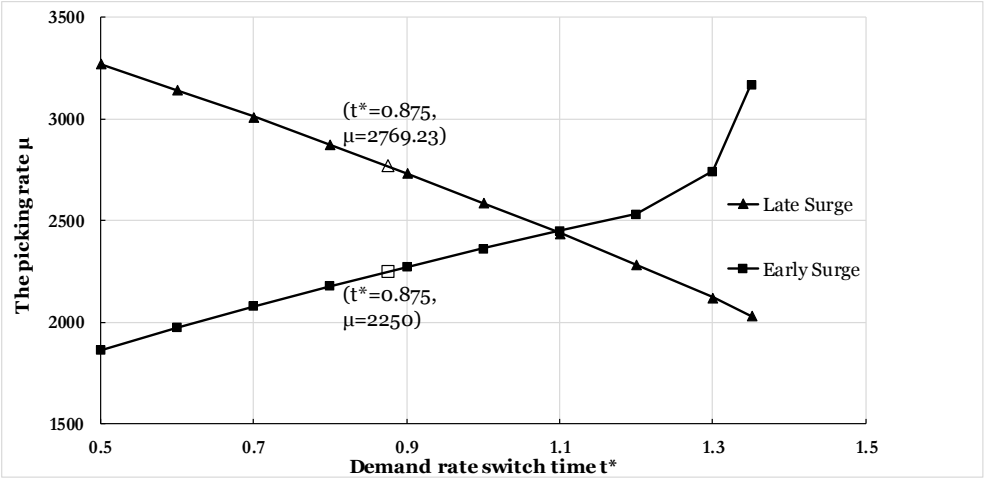


Figure 16. The minimum required picking rate in an early and a late surge scenario for single wave picking.

All other conditions being the same, when t^* is small, an early surge is more advantageous for the retailer as it requires a lower picking rate compared to a late surge scenario. However, as t^* increases, a late surge becomes more favorable for the retailer. This is explained by the fact that when t^* is small, there is a shorter window $[t_c - T, t^*]$ for orders to arrive at rate λ_1 and a longer window $[t^*, t_c]$ for orders to arrive at rate λ_2. Therefore, the late surge scenario ($\lambda_1 < \lambda_2$) will have more orders in the current cycle and will require a higher picking rate to finish the workload. However, as t^* increases, fewer orders will build up in the late surge scenario and therefore a lower picking rate is required. Hence, a late surge becomes more appealing to the retailer after a certain point t^*.

When $t^* = 0.875$ (i.e., 9pm yesterday) the current cycle is split into 12 hours of arrival rate λ_1 and 12 hours of arrival rate λ_2. Therefore, the number of orders that accumulate in the current cycle is the same for the early and late surge scenario. However, the picking rate required in the early surge (2250) is lower than that in the late surge scenario (2769.23). In other words, although the same number of online orders have accumulated in the system, early and late surges require different picking rates and wave release times. These results indicate that retailers should consider carefully the impact of the timings and durations of different demand rates experienced in a BOPS cycle.

3.3.2 The difference between the two demand rates

In addition to the timing of demand surges, we are also interested in the impact of the relative levels in the two demand rates. For the numerical experiments, we fix $t_c = 1.375, t_d = 1.5, \beta = 0.95$ and the values set for t^* are $\{0.5, 0.6, 0.7, 0.8, 0.875, 0.9, 1, 1.1, 1.2\}$. Based on Eq. (3.4), we investigate the picking rate in three scenarios when the demand rate difference is 300, 500, and 900, respectively (see Figure 17).

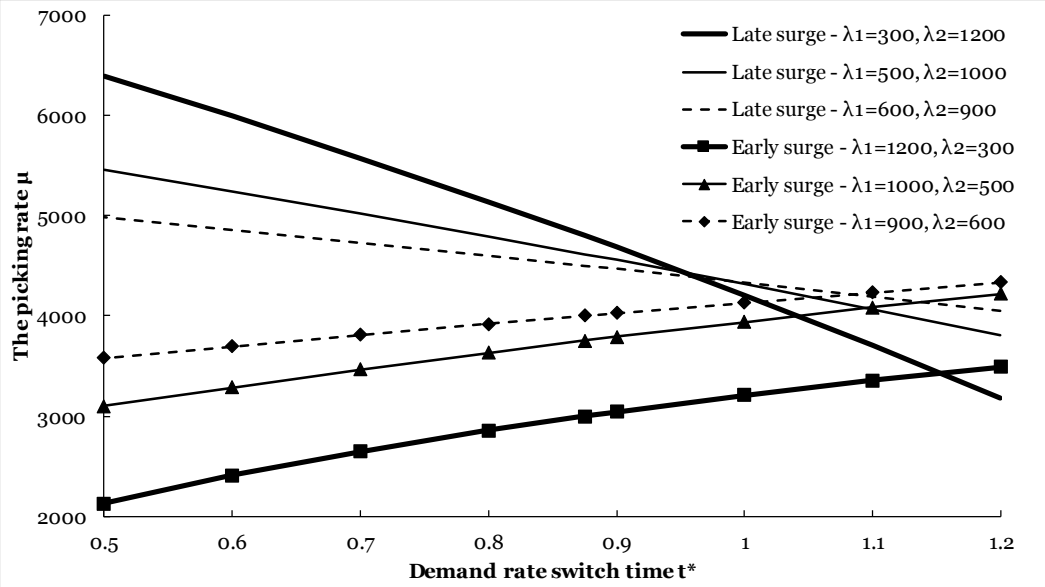


Figure 17. The minimum picking rate when the arrival rate difference is 300, 500 and 900.

Clearly, when the demand rate discrepancy is relatively small, the difference in the minimum picking rate required for a late or an early surge is also small. However, when the retailer encounters two substantially different arrival rates in a day, the difference in picking rate required for an early or a late surge is considerable. In other words, when the difference in the relative magnitude of the two arrival rates is small, the timing of a demand surge has limited impact on the retailer. However, when the difference in the two arrival rates is large, the picking rates required in an early surge and a late surge are very different.

3.4 A generalised store wave picking model

We have shown the relationship between the picking rate and wave release time for single wave picking for a BOPS service with a given service level. A natural extension therefore is what is the minimum picking rate when a retailer picks with more than one wave in a day? Additionally, how many picking waves should a retailer launch in a day? To this end, we present a generalised model where the service level is a function of both the number of picking waves and the picking rate.

In multiple wave picking, N picking waves are released at times $t_{w,1}$, $t_{w,2}$, …, $t_{w,N}$ to process all λT orders that have accumulated in $(t_{w,1}-T, t_{w,1}]$ with a picking rate $\mu_N \geq \lambda$. The orders that contribute to NSD are those that arrive in order cycle $(t_c-T, t_c]$ and that are picked before the deadline t_d with $t_c \leq t_d < t_c+T$. We assume that picking waves in a cycle are immediately sequential, i.e., the next picking wave starts immediately after the end of the current wave (see Figure 18 for a 3-wave model).

Workload is the number of orders accumulated for a pick wave to process. We first summarise the expressions for the workload handled in each wave in a multi-wave picking model in Table 4 and show the derivation of the service level. We then discuss the two cases where the picking rate equals the arrival rate and where the picking rate is greater than the arrival rate. In each case, we reveal the intertwined relationship between the service level, the picking rate, the number of picking waves and wave release times in a cycle.

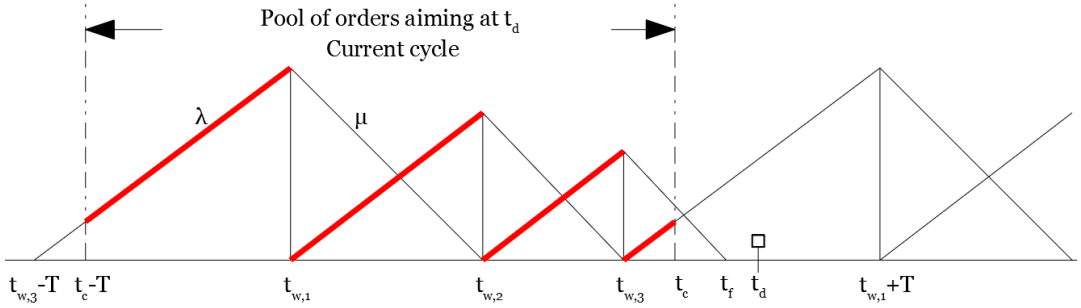


Figure 18. A generalised store wave picking model - illustration of 3 waves per cycle.

Table 4 Workload for each picking wave in the generalised store picking model.

Waves	Workload
First wave	$(t_{w,1} - (t_{w,N} - T))\lambda = (t_{w,2} - t_{w,1})\mu_N$
Second wave	$(t_{w,2} - t_{w,1})\lambda = (t_{w,3} - t_{w,2})\mu_N$
.	
.	
.	

i-th wave	$((t_{w,i} - t_{w,i-1}))\lambda = (t_{w,i+1} - t_{w,i})\mu_N$
.	
.	
.	
Second last wave (N-1)	$(t_{w,N-1} - t_{w,N-2})\lambda = (t_{w,N} - t_{w,N-1})\mu_N$
Last wave (N)	$(t_{w,N} - t_{w,N-1})\lambda = (t_f - t_{w,N})\mu_N$

Solving the equations in Table 4, we find that the daily picking activities finish at:

$$t_f = t_{w,N} + \frac{(\frac{\lambda}{\mu_N})^N}{(\frac{\lambda}{\mu_N})^0 + (\frac{\lambda}{\mu_N})^1 + (\frac{\lambda}{\mu_N})^2 + \cdots + (\frac{\lambda}{\mu_N})^{N-1}} T = \begin{cases} t_{w,N} + T/N & \text{for} \quad \mu_N = \lambda \\ t_{w,N} + \frac{\lambda^N(\mu_N - \lambda)}{\mu_N(\mu_N{}^N - \lambda^N)} T & \text{for} \quad \mu_N > \lambda \end{cases}. \quad (3.5)$$

If picking finishes before the deadline (i.e., $t_f \leq t_d$), all orders arriving from $t_c - T$ to $t_{w,N}$ will have been picked by the deadline. Orders arriving from $t_{w,N}$ to t_c will fail to be picked by the deadline. Thus:

$$NSD = \frac{(t_{w,N} - (t_c - T))\lambda}{\lambda T} = 1 - \frac{t_c - t_{w,N}}{T} \quad \text{for} \quad t_f \leq t_d$$

If picking finishes after the deadline (i.e., $t_f > t_d$), $(t_d - t_{w,1})\mu_N$ orders will have been fulfilled by the deadline. However, among these picked orders, $((t_c - T) - (t_{w,N} - T))\lambda$ orders are 'residual' from the previous cycle. Thus:

$$NSD = \frac{(t_d - t_{w,1})\mu_N - ((t_c - T) - (t_{w,N} - T))\lambda}{\lambda T} = \frac{(t_d - t_{w,1})\mu_N}{\lambda T} - \frac{t_c - t_{w,N}}{T} \quad \text{for} \quad t_f > t_d$$

In summary, the service level NSD for N picking waves can be expressed as follows:

$$NSD = \begin{cases} \frac{(t_d - t_{w,1})\mu_N}{\lambda T} - \frac{t_c - t_{w,N}}{T} & \text{for} \quad t_f > t_d \\ 1 - \frac{t_c - t_{w,N}}{T} & \text{for} \quad t_f \leq t_d \end{cases}. \quad (3.6)$$

Evidently, if $\beta = 1$, then $t_{w,N} = t_c$. As a result, we present the following proposition:

Proposition 3: If a retailer strives for a 100% service level, i.e., $\beta = 1$, picking activities should be scheduled in a way such that the last wave starts exactly at the order cut-off time to guarantee service with a minimum picking rate.

Based on Eq. (3.5) and Eq. (3.6), the exact solutions for $t_{w,N}$, μ_N and N depend on the expression of finish time t_f, which hinges on the comparison of the arrival rate and the picking rate. Therefore, we discuss two separate cases.

Case 1: Picking rate is equal to arrival rate, i.e., $\mu_N = \lambda$.

In this case, orders are evenly distributed among picking waves so that each wave has a duration of T/N. Therefore, $t_{w,N} = t_{w,1} + (N-1)T/N$ and $t_f = t_{w,N} + T/N$. Based on Eq. (3.6), if $t_f > t_d$ and $\mu_N = \lambda$, $NSD = \frac{(t_d - t_{w,1})\mu_N}{\lambda T} - \frac{t_c - t_{w,N}}{T} = 1 - \frac{1}{N} + \frac{t_d - t_c}{T}$. If $t_f \leq t_d$, NSD is a constant which can be derived by using the point when $t_f = t_d$. Replacing $t_{w,N}$ with $t_f - T/N = t_d - T/N$ in $NSD = 1 - \frac{t_c - t_{w,N}}{T}$ gives $NSD = 1 - \frac{1}{N} + \frac{t_d - t_c}{T}$. Thus, when $\mu_N = \lambda$, service level $NSD = 1 - \frac{t_c - t_{w,N}}{T} = 1 - \frac{1}{N} + \frac{t_d - t_c}{T}$.

An interesting observation is that when $\mu_N = \lambda$, the service level only depends on the number of picking waves, regardless of wave release times. Specifically, the service level increases as the number of picking waves launched per cycle increases but with a diminishing gradient (i.e., $\frac{dNSD}{dN} = \frac{1}{N^2} > 0$ and $\frac{d^2NSD}{dN^2} = -\frac{2}{N^3} < 0$). In this case, with a given *NSD* service level β, the retailer can be informed with the minimum number of picking waves needed: $N \geq \left\lceil \frac{T}{(1-\beta)T + (t_d - t_c)} \right\rceil$. For example, the retailer needs to launch at least $\left\lceil \frac{T}{(t_d - t_c)} \right\rceil$ waves for a 100% service level. An intrinsic quality of this case is that the retailer has to pick continuously, i.e., the order picking window equals to cycle length T, which could be unrealistic for retail operations in practice.

Case 2: Picking rate is greater than arrival rate (see Figure 18**),** i.e., $\mu_N > \lambda$.

Based on Table 4, we have the following expressions:

$$
\begin{aligned}
t_{w,i} &= t_{w,1} + \frac{\lambda {\mu_N}^{N-1} - \lambda^i {\mu_N}^{N-i}}{({\mu_N}^N - \lambda^N)} T, \\
t_{w,N} &= t_{w,1} + \frac{\lambda({\mu_N}^{N-1} - \lambda^{N-1})}{({\mu_N}^N - \lambda^N)} T, \\
t_f &= t_{w,N} + \frac{\lambda^N(\mu_N - \lambda)}{\mu_N({\mu_N}^N - \lambda^N)} T \ . \qquad (3.7)
\end{aligned}
$$

Based on Eq. (3.6), NSD remains constant at the level of $1 - \frac{t_c - t_{w,N}}{T}$ as long as $t_{w,N} + \frac{\lambda^N(\mu_N - \lambda)}{\mu_N({\mu_N}^N - \lambda^N)} T \leq t_d$. Denote $g(\mu_N) = \frac{\lambda^N(\mu_N - \lambda)}{\mu_N({\mu_N}^N - \lambda^N)} T$ and clearly:

$$\frac{dg(\mu_N)}{d\mu_N} = -\frac{\lambda^N[N{\mu_N}^{N-1}+(N-1){\mu_N}^{N-2}\lambda+\cdots+2\mu_N\lambda^{N-2}+\lambda^{N-1}]T}{[\mu_N({\mu_N}^{N-1}+{\mu_N}^{N-2}\lambda+{\mu_N}^{N-3}\lambda^2+\cdots+\mu_N\lambda^{N-2}+\lambda^{N-1})]^2} < 0.$$

That is to say, picking rate μ_N can only be minimised when $g(\mu_N) = \frac{\lambda^N(\mu_N-\lambda)}{\mu_N({\mu_N}^N-\lambda^N)}T$ is maximised, which means that $t_{w,N} + \frac{\lambda^N(\mu_N-\lambda)}{\mu_N({\mu_N}^N-\lambda^N)}T \leq t_d$ is binding. This reinforces proposition 2 that the last picking wave always finishes exactly at the deadline if the retailer operates on the BPF. Additionally, with a given service level β, the last wave commences at $t_{w,N} = t_c - (1-\beta)T$ and the relationship between the number of picking waves and picking rate on the BPF satisfies

$$\frac{\lambda^N(\mu_N-\lambda)}{\mu_N({\mu_N}^N-\lambda^N)} = \frac{t_d-t_{w,N}}{T} = \frac{t_d-t_c}{T} + (1-\beta). \tag{3.8}$$

This leads to the following proposition delineating the relationships between the number of picking waves, the minimum picking rate, and the wave release times on the BPF for the generalised model.

Proposition 4: When the picking rate is greater than the demand arrival rate, for a given service level, the more picking waves a retailer launches per cycle, the lower the picking rate required but the earlier the first wave must commence.

The proof of proposition 4 is presented in Appendix A.2. We use the same example of Tesco's same day BOPS offering to illustrate proposition 4. The picking rate μ_N and the wave release times $t_{w,i}$ are computed based on Eq. (3.7) and (3.8) for each N (see Table 5). Tesco advertises that all orders placed before 9am today will be ready for collection after midday. Therefore, the current cycle ($T = 1$) starts from 9am yesterday ($t_c - T = 0.375$) to 9am today ($t_c = 1.375$), during which all orders accumulated are due at midday today ($t_d = 1.5$). To achieve a 95% service level, the retailer should start picking at 7:48am ($t_{w,1} = 1.325$) today with a picking rate of 1714.29 if operating one picking wave per cycle. However, if the retailer decides to pick an extra wave for the same service level, picking should start at 11:38pm yesterday ($t_{w,1} = 0.985$) but with a picking rate approximately three times smaller ($\mu_N = 582.657$). To reduce the picking rate to its boundary value, i.e., $\mu_N = \lambda = 300$, the retailer should launch at least 6 picking waves ($N \geq \left\lceil \frac{T}{(1-\beta)T+(t_d-t_c)} \right\rceil = \left\lceil \frac{1}{(1-0.95)+(1.5-1.375)} \right\rceil = 6$) for the target service level and pick continuously.

To visualise how the picking rate needed for a guaranteed service level can be lowered with more picking waves, we generate the BPF for each N and for desired service levels between 93% and 100% based on Eq. (3.8) (See Figure 19). By launching more picking waves per cycle, the retailer could push the picking rate needed for a guaranteed service level closer to the boundary level (i.e., $\mu_N = \lambda = 300$). However, Figure 19 also reveals the diminishing effect of reducing picking rate by initiating more picking waves per cycle. For example, with a target service level of 95%, the retailer could decrease the picking rate needed by 1131.633 (1714.29-582.657) from single wave to two waves per cycle. However, if the retailer picks an extra wave from two-wave picking, the picking rate can only be reduced by 175.006 (582.657-407.651). Essentially, Proposition 4 means that if the retailer launches more picking waves per cycle, picking will start earlier but will still finish the last wave at the deadline. As a result, the order processing window is prolonged and therefore the retailer could lower the picking rate needed to attain the target service level. However, the picking rate cannot be lowered infinitely since it needs to be no less than the demand rate to ensure feasibility. Therefore, the impact of increasing the number of picking waves on reducing the picking rate diminishes as the picking rate approaches the demand rate. In general, retailers have a budget for the number of pickers and therefore the picking rate for online fulfilment services. Eq. (3.8) can be used by retailers to decide the number of picking waves per cycle and the wave release times flexibly based on the budget or resources they have for BOPS order picking operations.

Table 5 Generalised store wave picking model - a numerical illustration.

The number of picking wave launched per cycle (N)	The minimum picking rate (μ_N)	The 1st wave release time $t_{w,1}$	The 2nd wave release time $t_{w,2}$	The 3rd wave release time $t_{w,3}$	The 4th wave release time $t_{w,4}$	The 5th wave release time $t_{w,5}$	The picking duration ($t_d - t_{w,1}$)
1	1714.29	1.325					0.175
2	582.657	0.985	1.325				0.515
3	407.651	0.764	1.087	1.325			0.736
4	344.367	0.629	0.894	1.124	1.325		0.871
5	313.454	0.543	0.751	0.951	1.142	1.325	0.957
6	300	-	-	-	-	-	1.0

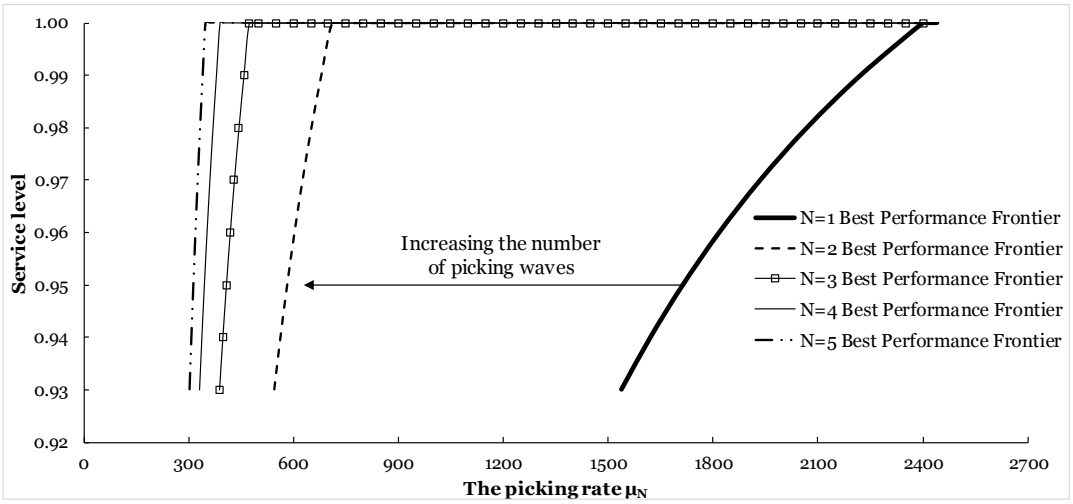


Figure 19. The shift in BPF as the number of picking waves per cycle increases.

3.5 Managerial insights and implications

The landscape for omni-channel retailing and specifically for BOPS fulfilment services is highly competitive. It is vital for retailers to understand how to schedule and improve the picking of online orders in conventional retail stores (JDA, 2016) and to understand the minimum staffing levels needed to guarantee a desired or advertised service level (Mahar and Wright, 2017). For example, the retailer 'PrettyLittleThing' had to suspend its advertised next day delivery service for weeks as its fulfilment capacity could not match the very high demand (Retail Gazette, 2018). To demonstrate the contribution of our work, we discuss the managerial insights and implications of the analytical models and results presented from Section 3.2 to Section 3.4 for omni-channel retailers who provide BOPS services.

When designing BOPS service offerings, retailers may be tempted to set the order cut-off time and deadline by rules of thumb considering factors such as business opening and closing hours or customer lifestyles in the operating regions (Ko et al., 2007). For example, the deadline for same day BOPS services is typically 12 noon, 2pm or 4pm in the UK when customers finish work and then visit stores to collect their orders. However, there are no guidelines to help retailers to decide when to start picking BOPS orders. Clearly, if retailers wait for more orders to accumulate, fewer orders arriving late in the cycle would be missed. However, this shortens the order processing window and therefore a higher picking rate is required, which incurs higher labour costs. Our analytical models assist retailers in deciding when to start picking, the number of picking waves to deploy, and to understand the impact of the wave release time on the required picking rate.

In single wave BOPS picking, we have presented the BPF. For a specific service level, the corresponding point on the BPF curve shows the minimum picking rate needed and allows

the required wave release time to be calculated. The BPF can assist retailers in several ways. First, the results presented allow the lowest staffing level and the schedule of picking activities for a given BOPS service level to be determined. Second, retailers can use the BPF to identify the maximum service level attainable for a given picking rate with the known parameters of demand rate, order cut-off time, and deadline. Third, the BPF serves as a diagnostic tool for retailers to benchmark their current operations against the best that could be attained. The positive but diminishing gradient of the BPF means that an increase in the picking rate improves the service level but that its impact becomes weaker. In other words, a large increase in the picking rate is needed when retailers aim at achieving service level close to 100%. Recall that if the service level increases by Δ, the picking rate required has to increase by $\frac{\lambda T^2}{(t_d - t_w)(t_d - (t_w + \Delta T))}\Delta$. This finding provides a way to quantitatively measure the impact of the increase in the service level on the picking rate, which can be used by retailers to consider trade-offs between service level and costs.

For retailers that aspire to achieve a 100% service level, picking activities should be scheduled in a way to ensure that the last picking wave is released exactly at the order cut-off time. In this way, all orders arriving in the current cycle can be fulfilled by the promised deadline. If a less ambitious service level is set, i.e., $0 < \beta < 1$, the retailer can schedule all the picking waves before the order cut-off time. For retailers interested in reducing the picking rate, they should never start a picking wave after the order cut-off time. Otherwise, a higher picking rate associated with higher costs, is needed to maintain the target service level.

In the generalised model, we show that the more picking waves a retailer launches per cycle, the lower the picking rate needed for a guaranteed service level. Specifically, the required picking rate can be pushed closer to its boundary value (i.e. $\mu_N = \lambda$) when launching more picking waves per cycle. However, the picking rate reduction benefit becomes less significant as the number of picking wave per cycle increases. Meanwhile, the retailer should commence the first picking wave earlier while still finishing picking activities at the deadline (i.e., $t_f \leq t_d$ is binding). Therefore, the overal picking duration is longer when the retailer picks more waves per cycle with a lower picking rate. Hence, our results show there is a trade-off between lowering the picking rate and prolonging the overall picking duration. This observation can assist retailers to make more sensible planning decisions on picking resources, particularly staffing level in stores. For example, instead of hiring more pickers to boost the picking rate for a target service level for BOPS offerings, retailers could avoid such a hiring cost by asking current pickers to work extra time, which can be quantified by our

analytical models. With the longest possible picking duration and a target service level given retailers' specific business operating environment (e.g. regulations on working hour limits and policies on payment for working overtime), our results can be used to determine the lowest staff level for BOPS picking. More generally, the results can help retailers to schedule their picking activities flexibly for a target service level depending on the picking resources determined by workforce budgets at their disposal.

In the case when a retailer experiences two different arrival rates per cycle, before the demand rate switch point, a lower picking rate is needed when more customers place their orders well before the order cut-off time, i.e., when there is an early surge in demand. Moreoever, the shorter the duration of the early surge, i.e., a smaller value for parameter t^*, the lower the picking rate needed compared to a late surge situation and therefore the more beneficial this business environment for the retailer. However, if most customers place their BOPS orders well before the cut-off time and this high demand rate continues until a few hours before the order cut-off time when the demand rate drops (i.e., higher t^* and therefore longer duration of the early surge), it is less appealing for the retailer as the picking rate required could be higher than that in the late surge situation.

When leveraging a conventional retail store network for online fulfilment, retailers need to set appropriate picking rates to balance the advertised service level and the operating costs. The results indicate that the decision on the level of picking rate should be made taking account of the timing and duration of the demand changes (i.e., the value of parameter t^*). The analysis shows that when a BOPS ordering cycle is evenly split into two periods with different demand rates in each period, an early or late surge will lead to different picking rates even though the number of accumulated orders is the same. If the difference between the two arrival rates is relatively small, the difference in the minimum picking rate required between early and late surge is also small. However, if the two arrival rates are considerably different, the BOPS demand pattern (i.e., whether it is an early or a late surge) is critical for the retailer as the minimum picking rate will be very different for different order arrival patterns. Futhermore, by investigating the impacts of the timing and the magnitudes of demand rate flunctuations on the required picking rate and therefore the labour costs, retailers can decide whether it is appropriate to introduce incentive schemes to influence the demand rate. Specifically, if the current customer ordering behaviour requires a high picking rate, retailers may seek to steer customer demand to a pattern requiring a lower picking rate by offering price discounts or other incentives, so long as the extra costs for incentives are lower than the labor cost savings.

3.6 Concluding remarks

The rise and popularity of BOPS services has brought both blessings and challenges for omni-channel retailers. Without considering how to carry out in-store picking operations, retailers may fulfil BOPS orders in an unprofitable way while jeopardizing store services for walk-in customers. To provide insights on the operational decisions for in-store picking activities, we have developed the BPF to determine the minimum picking rate, the timing to start picking activities and the required number of picking waves to fulfil online orders to achieve a pre-defined service level. The BPF reveals the best combination of wave release time(s) and the picking rate for a given service level. A key finding is that each additional picking wave leads to an earlier start for the daily picking activities and therefore a longer picking duration but with a lower picking rate needed. We also investigate the case with different demand arrival rates in an ordering cycle and highlight the impacts of the both the timing and duration of demand surges on the required minimum picking rate.

In this study, we have derived the solutions based on a given service level. A natural extension is to examine the optimal service level of BOPS offerings for retailers. Enhancing the service level could lead to increased sales and therefore revenue growth while imposing higher operating costs to retailers. As a result, research on determining the optimal service level to balance the trade-off between service and cost could be very valuable. Future research could also consider the case when retailers adjust the picking rate in each wave according to the workload. The combinatorial optimisation of picking and delivery in ‘ship from store’ fulfilment is also an interesting area for further investigation.

Chapter 4

Deadline-driven fulfilment: the optimal online ordering window for retail store collection

In this chapter, we present cost minimisation models to determine the optimal online order receipt window for BOPS services, the time to commence order processing, and the minimum required processing rate. We first analyse the case in which the retailer fulfils BOPS orders from decentralised stores where customer collection takes place. We then analyse the case in which the retailer fulfils BOPS orders from centralised hub stores or warehouses and then ships them with a fixed transportation window to other stores for customer collection. The analysis shows the trade-offs that arise in optimal BOPS fulfilment and its sensitivity to the principal model parameters, such as BOPS order arrival rate, unit inventory holding cost and unit delay penalty cost. We apply the models to explain differences observed in contemporary retail operating strategies and present frameworks to show the impact of both product and retailer characteristics on the optimal service window offering for store collection. The impact of centralisation on total fulfilment costs is also highlighted. The two normative frameworks show that same-day collection is more appropriate for fast moving products with high value and/or perishability, whilst less ambitious next day or longer collection offerings are more appropriate for slow moving products with low value and/or long shelf life.

4.1 Introduction

Omni-channel retailers offer more order fulfilment choices to customers than traditional bricks-and-mortar retailers. In particular, the store collection fulfilment model is growing worldwide (Hübner et al., 2016a). Forbes (2019c) note that 68% of a sample of US shoppers used such services repeatedly. 64% of UK retailers now offer store collection services to online customers (OrderDynamic, 2018). Store collection services are frequently labelled as BOPS or click-and-collect services. We use the former term here for consistency. The benefits of BOPS services from the viewpoint of both customers and retailers have been discussed in Chapter 3.

Given the increasing popularity and importance of BOPS services, it is in the retailer's interest to provide competitive BOPS services in omni-channel retailing. Typically, the retailer promises an earliest collection time for customer orders placed online by an advertised cut-off time, hence setting a deadline by which BOPS orders should be ready. Table 6 shows current BOPS service promises offered by a selection of major UK retailers. Less than a third of UK retailers commit to have BOPS orders ready within 24 hours. The

majority take more than 2 days (OrderDynamic, 2018). KPMG warned that UK retail could be losing as much as £4.9 billion a year in not providing next day delivery and collection options (KPMG, 2018). The stakes for not offering competitively fast BOPS services are therefore very high.

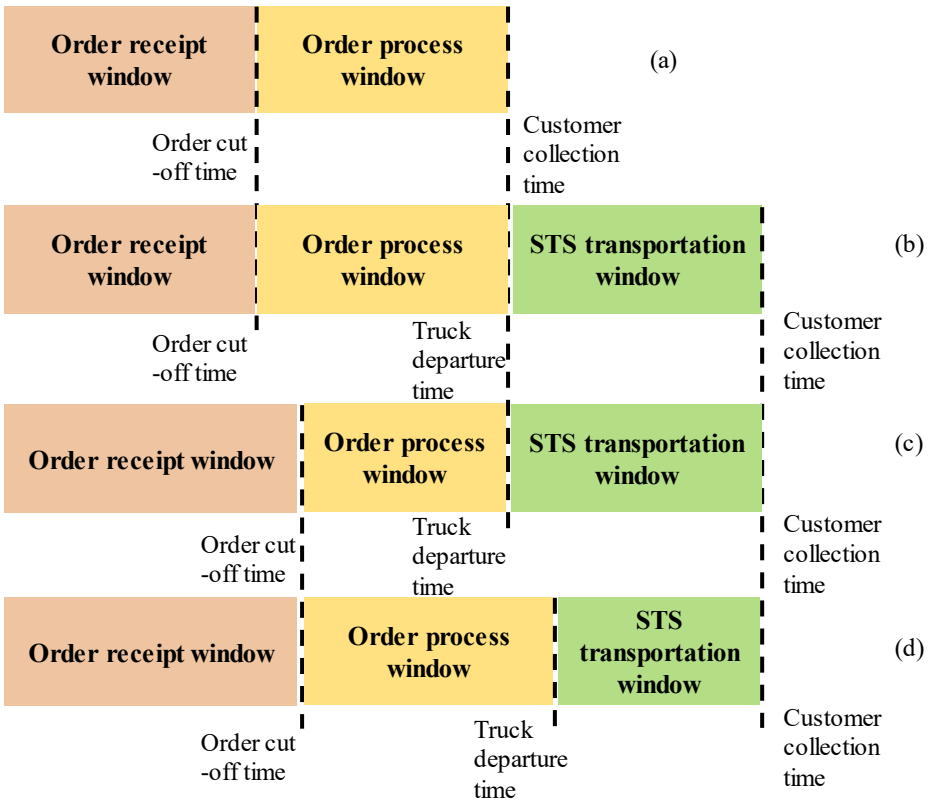
Table 6 Examples of BOPS service offerings from British retailers.

Retailer	Sector	Cut-off time	Earliest collection time
Tesco	Groceries	9am	1pm same day
		1pm	4pm same day
Sainsbury's	Groceries	12 noon	4pm same day
John Lewis	Department store	8pm	2pm next day
Marks & Spencer	Department store	10pm	12 noon next day
Next	Fashion	Midnight	12 noon next day
Boots	Health & beauty	8pm	12 noon next day

It is worth pointing out that a small number of leading retailers also provide continuous ordering services such as the 'in-store collection after 1 hour' promise by Currys PC World. Such continuous ordering service is only feasible when items requested online are stocked in the collection stores and when it is possible to finish order picking as customers arrive at stores. For example, it would be very problematic to offer continuous ordering service for orders requiring more picking effort due to the number of items requested (e.g., groceries). Therefore, the continuous ordering service is typically restricted to a small range of products at selected stores. Here, we focus our analyses on the more commonly used BOPS service form with a specified order cut-off time and a corresponding deadline (i.e., earliest collection time) to ensure wide application of the results.

BOPS orders can be fulfilled in two ways - from the local collection store or from another fulfilment location within a retailer's store and distribution network and then shipped to the relevant branch for customer collection (i.e., STS). We use the term BOPS in this study to include both the local store fulfilment and the STS approach. Faster BOPS order fulfilment is essentially a trade-off between three components: the order receipt window, the order processing window, and the STS transportation window (see Figure 20). The interaction of different time windows in satisfying online demand has been emphasised in the practitioner press in the context of home delivery (Supply Chain 24/7, 2015). In this study, we focus on store collection and do not consider home delivery. The order receipt window is the time that elapses between two consecutive order cut-off times. The order processing window is the time needed to fulfil online orders from picking to packing. The STS transportation window

is the time needed to move BOPS orders from the order processing location to the collection point when orders are not fulfilled locally.



Note: Ship-To-Store (STS) transportation window is the time needed to move BOPS orders from order processing facilities to collection points.

Figure 20. Trade-offs in BOPS order fulfilment (adapted from Supply Chain 24/7 (2015)).

If BOPS orders are fulfilled from stores where customers collect their orders, the STS transportation window is zero (see Figure 20(a)). All things being equal, if a retailer uses the STS approach, customers will be promised a later collection time (see Figure 20(b)). Given an advertised deadline to get orders ready for collection, a retailer may aspire to provide a longer order receipt window with a later cut-off time to capture last-minute shoppers and potentially increase sales. However, this may require a shorter order processing window (see Figure 20(c)), necessitating a higher order processing rate. A more compressed order processing window can result in service deterioration with more late orders, which may have an adverse effect on a retailer's bottom line in terms of penalty costs and on its reputation. Retailers can also offer a longer receipt window and meet the collection deadline by shortening the STS transportation window (see Figure 20(d)). However, a short transportation window may impede cost-efficient vehicle route planning (Boyer et al., 2009). To meet the BOPS service promises, retailers must carefully balance the lengths of these three intertwined time windows.

Retailers may set a single fixed daily order cut-off time but variations exist. Tesco currently offers two slots for same day BOPS orders (see Table 6): the first order receipt window is 4 hours (9am to 1pm) and the second is 20 hours (1pm to 9am next day). Some retailers specify an order receipt window longer than one day. The UK book and stationery retailer, WH Smith, offers two delivery deadlines for BOPS services per week, namely Wednesdays and Fridays (www.whsmith.com.uk). In Superdrug, a leading UK beauty products retailer, standard BOPS orders are delivered to stores for collection each Wednesday, offering an order receipt window of 7 days for free BOPS service (www.superdrug.com). Retailers also compete on BOPS services at holiday times. Major differences were reported in the online cut-off times for Christmas 2018 in the UK, with some retailers offering a receipt window until 8pm on the 23rd of December whilst others terminated their order receipt windows as early as the 20th of December for collection and delivery services (The Independent, 2018; The Telegraph, 2018). The potential impact on a retailer's competitiveness, the customer perception of a retailer's services, the lack of consistency in practice, and the ramifications for other decisions such as the required processing rate, make it both critical and challenging for retailers to determine the optimal length of the order receipt window.

Retailers often tailor their online delivery and collection services to product types. UK customers can collect their groceries in 3 hours in a Tesco store after the cut-off time (www.tesco.com), but have to wait for 3-5 working days to collect a pair of shoes ordered online from Clarks (www.clarks.co.uk). In Marks & Spencer (M&S), next day BOPS is available for clothing and beauty products, while only home delivery with a waiting time from one to six weeks is offered for furniture (www.marksandspencer.com). The willingness of customers to wait for online customers also differs across product types (Digitalcommerce360, 2017). For example, customers are more tolerant of waiting for houseware products than consumer electronics (Digitalcommerce360, 2017). Thirumalai and Sinha (2005) show that customer satisfaction with order fulfilment decreases from convenience goods (e.g. home and office supplies) to shopping goods (e.g. apparel), and to specialty goods (e.g. laptops). Hence, it is important to understand which product types should be offered with same-day collection and which types should be fulfilled later.

The optimal order receipt window for BOPS services is a critical decision at the interface of marketing and operations but has not been examined previously in the literature. In this chapter, we develop analytical models to provide retailers with guidelines to specify the optimal front-end BOPS service offering and the required back-end operational decisions to process BOPS orders to meet a promised service deadline. Specifically, we answer three questions:

1. How long should the order receipt window be for a BOPS service?
2. When should the retailer commence BOPS order processing for a given collection deadline and what is the minimum order processing rate required?
3. How do product and retailer characteristics impact the decisions on BOPS service offerings?

We develop an analytical model to jointly optimise the length of the order receipt window, the time to start order processing and the order processing rate needed, with a given STS transportation window. We start our analysis with the important case when retailers fulfil all BOPS orders in local decentralised stores (and thus a zero STS transportation window). We then extend the model for a known STS transportation window, which indicates a more centralised order processing strategy where BOPS orders are fulfilled from a few facilities and then delivered to other stores for customer collection.

We make four main contributions in this chapter. This is the first work to present an optimisation model for BOPS fulfilment operations in a deadline-driven system. We develop cost minimisation models to help retailers determine the length of the ordering window and collection speed offered to customers. Second, we provide optimal solutions for the retailer's back-end operations, specifically the minimum order processing rate needed to ensure BOPS orders are ready for a promised deadline. Third, we cast light on the impact of the level of fulfilment centralisation on each of these decisions and the total fulfilment costs. Our analysis reveals that the cost impact from picking centralisation may depend on the relative importance of inventory costs vs. store congestion cost. Fourth, our work is the first to investigate the impact of product types and retailer profiles on BOPS service offerings. We present two frameworks to illustrate how product and retailer characteristics can affect the design of BOPS service speed offerings, highlighting the conditions under which same day BOPS services might be appropriate and the conditions when a less ambitious BOPS service offering can be sufficient.

4.2 Model formulation and analysis

4.2.1. Problem description and assumptions

We consider a retailer offering with a pre-specified collection service deadline for orders placed online before a cut-off time, which is a common BOPS service offering in today's retail world (see Table 6). The order cut-off time and the corresponding deadline are denoted as t_c and t_d, respectively.

We call the order receipt window a cycle, T, defined as the time that elapses between two consecutive order cut-off times. T is a decision variable in our model. The orders accumulated during a particular cycle $(t_c - T,\ t_c]$ should be processed, transported and ready for collection by the deadline, t_d. The deadline t_d can be sometime in the next cycle (or day, if T = 1 day) $t_c < t_d < t_c + T$, or can be several cycles later. When the deadline is in the next cycle, we denote the BOPS service speed $t_d - t_c$ as fT, where f is a given fraction within (0, 1).

We assume that the demand is received continuously over time and that the demand rate λ is constant. This is a reasonable assumption in large-scale retailing with centralised online order management systems. Retailers distribute online demand across multiple fulfilment facilities including retail stores, which will receive allocated online demand at a relatively stable rate.

After accumulating orders during the receipt window T, the retailer decides when to start picking operations and the picking rate required to meet the promised deadline. The time to start picking is denoted as t_w, which is a decision variable in the model. We assume that a single picking wave is released at t_w in each cycle. The wave release time t_w could be before, at, or after the order cut-off time t_c. If $t_w < t_c$, orders arriving after t_w can only be processed in the next cycle and will miss their designated deadline.

We assume that the same amount of effort is needed to pick each order and that sufficient inventory is available at the picking facility to fulfil all customer orders. The order processing rate is denoted as μ, and is also a decision variable in the model. We assume that μ cannot be less than the demand rate (i.e., $\mu \geqslant \lambda$), otherwise the system becomes congested. We also assume that the picking rate is constant within the order processing window. Varying the picking rate through adjusting staffing levels implies higher costs, which could be explored in extensions of the model. Order picking is driven by the promised deadline, which affects the time to start picking and the picking rate needed. Thus, for a given t_d (or f), when to start picking is equivalent to determining the length of the fulfilment service window $T_e = t_d - t_w$.

Figure 21 illustrates the operational logic in our model when the STS transportation window, denoted as L, is zero. The top diagram shows the evolution of the system orders over time: online orders accumulate in the system at a constant rate λ, and are processed starting at the wave release time t_w, at a faster rate μ. The bottom diagram illustrates the evolution of the online order work-in-process (WIP) over time. For every cycle, starting at t_w, WIP builds up at a constant rate μ. At the deadline t_d, orders that were picked before t_d leave the system, i.e.,

are collected by customers. All remaining unprocessed orders are designated as being late. After t_d, the late orders WIP decreases to zero at rate μ.

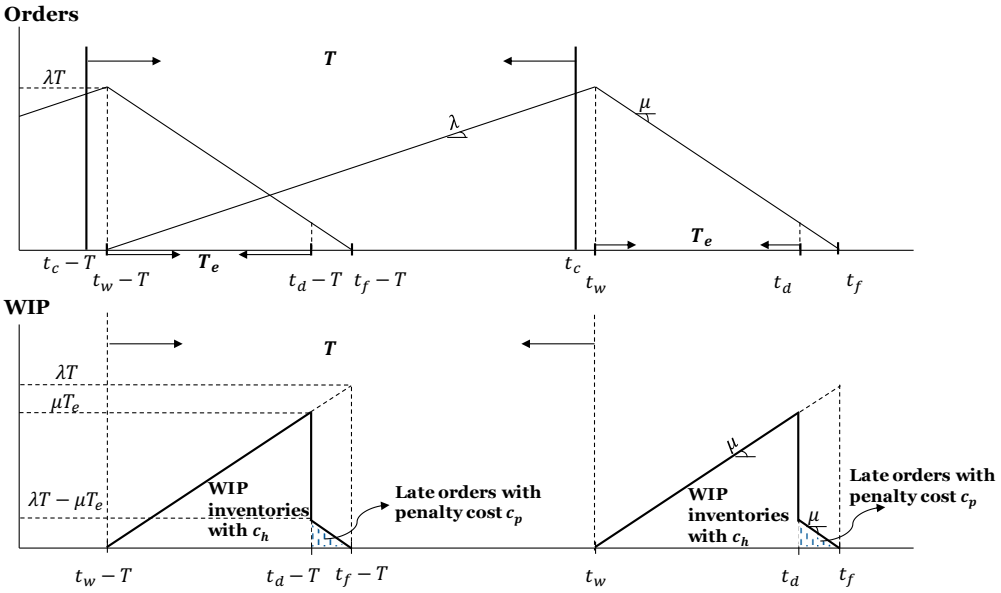


Figure 21. The evolution of order accumulation, WIP inventories, and late orders over time when $t_w > t_c$ and $L = 0$.

We now discuss the different costs in the model. Every cycle, the retailer generates picking lists for the accumulated online orders (λT in total) in the system so that they are ready for processing. We assume that this incurs a fixed setup cost, K. This is consistent with the operations management literature (Rao et al., 2005) and analogous to the set-up cost for generating picking lists and initiating picking routes in the warehouse picking literature (de Koster et al., 2012). Some of the orders will be processed and ready by the promised deadline, whereas others may miss the deadline. If an online order is prepared before the deadline, it will have to be stored (possibly in a protected or temperature controlled environment) before it is collected by customers at time t_d. These orders can be considered as WIP inventory, which is not available to other customers and has to be securely stored and handled appropriately, until collected by the customer. We associate with these WIP inventories an inventory holding cost c_h in a similar way as in the classical inventory management literature. The cost c_h (inventory holding cost per order per time unit) incorporates the cost of storage, and cost of spoilage, breakage and deterioration (Palaka et al., 1998). Orders that are not prepared by the deadline incur a deadline delay cost, c_p, which penalises each late order for each time unit it is delayed after the promised deadline. This is similar to penalty costs for stock outs in the traditional inventory management literature, although in our context the penalty cost is associated with exceeding the promised deadline. If the guaranteed deadline is not met, the retailer may offer customers a discount, or a partial or full refund to placate

them. For example, Domino's Pizza advertised '30 minutes or your pizza free' (Financial Times, 2013) and Amazon offered a $20 account credit for future purchase in addition to refunding delivery fees for late Christmas deliveries in 2017 (Sakraida, 2017). The penalty cost c_p has the same dimension as c_h, but is much larger than c_h.

For each order processed, the retailer incurs a variable picking cost. The unit variable picking cost, denoted as $c_\mu(\mu)$, represents all costs proportional to the volume of orders processed, including package material, employees, and equipment. We assume that the unit variable picking cost $c_\mu(\mu)$ is continuously differentiable and non-decreasing convex in μ, i.e., $c_\mu'(\mu) > 0$ and $c_\mu''(\mu) \geq 0$. The higher the picking rate μ, the lower the picking efficiency due to factors such as in-store congestion, and therefore the higher the unit variable picking cost $c_\mu(\mu)$. The warehouse picking literature also acknowledges that congestion is unavoidable when multiple pickers work concurrently within the same area, which leads to the erosion of picking efficiency (Gue et al., 2006; Pan and Wu, 2012). In the capacity management literature, researchers generally assume that successive investments to increase capacity incur higher marginal costs (Boyaci and Ray, 2003; Nguyen and Wright, 2015; Ray and Jewkes, 2004). In warehouse management, Gue et al. (2006) examined the impact of congestion on the time each worker is blocked, showing that the congestion is exacerbated when there are more workers. Our assumption that the unit variable picking cost increases at an non-decreasing rate when the picking rate increases ($c_\mu''(\mu) \geq 0$) is therefore well-established in the literature.

The notation used for the model formulation is summarised in Table 7. In the following section, we present cost minimisation models to optimise the length of the order receipt window T, the fulfilment service window T_e (and therefore the picking start time $t_w = t_d - T_e$), and the picking rate μ. Section 4.2.2 presents our analysis for a zero STS transportation window. Section 4.2.3 extends the model to consider both a non-zero short and long STS transportation window L. The model's objective function minimises the fixed setup cost, WIP inventory holding costs, lateness penalty costs and a variable picking cost per time unit.

Table 7 Notation for the cost minimisation models.

λ	Order arrival rate, i.e., the number of BOPS orders arrived per time unit
μ	Order picking rate, i.e., the number of BOPS orders picked per time unit
c_h	Inventory holding cost, £ per order per time unit
c_p	Penalty cost, £ per late order per lateness time
$c_\mu(\mu)$	Unit variable order picking cost, £ per order picked

$c_\mu'(\mu)$	Marginal variable order picking cost; the first derivative of $c_\mu(\mu)$
$c_\mu''(\mu)$	The second derivative of $c_\mu(\mu)$
K	Fixed cost for setting up order picking activities
t_c	The cut-off time for customers to place orders online
t_d	The deadline when orders placed by the cut-off time will be ready for collection
t_w	The time to start processing BOPS orders in a cycle
t_f	The finish time of overall picking activities in a cycle
t_p	The time to finish late orders from the previous cycle
L	The STS transportation window
T	The order receipt window, i.e., the length of a cycle
T_e	The fulfilment service window, i.e., $T_e = t_d - t_w$
f	The fulfilment speed as a fraction of a cycle, i.e., $f = (t_d - t_c)/T$

An asterisk * indicates optimality.

4.2.2 The order fulfilment model when $L = 0$

In a cycle $(t_c - T, t_c]$, the number of orders accumulated is λT and the number of orders picked by the deadline is $\mu(t_d - t_w)$. The retailer could start picking before the order cut-off time, i.e., $t_w < t_c$, striving for a lower picking rate needed with a longer processing window. However, orders arriving between $(t_w, t_c]$ will not be processed by t_d because each cycle releases only one picking wave, which means $(t_c - t_w)\lambda$ orders will be late. The retailer could also start picking at or after the order cut-off time, i.e., $t_w \geq t_c$, (see Figure 21) so that all orders that have accumulated in a cycle could be processed by t_d given a sufficient μ. We show in Appendix B.1 that from a cost minimisation perspective, the retailer should not initiate picking before the order cut-off time. Therefore, in the analysis below, we focus on the scenario in which the retailer starts picking at or after the order cut-off time, i.e., $t_w \geq t_c$.

Any orders fulfilled before the deadline will contribute to WIP inventories with a holding cost c_h and after the deadline will be regarded as late orders with penalty cost c_p. As shown in Figure 21, WIP inventories build up from zero to $\mu(t_d - t_w) = \mu T_e$ during the fulfilment service window $(t_w, t_d]$. Therefore, the average WIP level per time unit is $\frac{1}{T}(\frac{\mu T_e}{2} \times T_e + 0 \times (T - T_e)) = \frac{\mu {T_e}^2}{2T}$. For $L = 0$, all orders processed before the deadline will be satisfied on time. The number of late orders is $\lambda T - \mu T_e$ and these will all be processed $(\lambda T - \mu T_e)/\mu$ time units after the deadline. Hence, the average late orders per time unit is

$\frac{1}{T}\left(\frac{(\lambda T-\mu T_e)}{2}\times\frac{(\lambda T-\mu T_e)}{\mu}+0\times\left(T-\frac{(\lambda T-\mu T_e)}{\mu}\right)\right)=\frac{(\lambda T-\mu T_e)^2}{2T\mu}$. The fixed picking cost and variable picking cost per time unit are K/T and $c_\mu(\mu)\lambda$, respectively.

The problem of optimising BOPS fulfilment decisions in a deadline driven system with $L = 0$ can be formulated as follows:

(P1) $$\text{Min } TC_1(T_e,T,\mu)=c_h\frac{{T_e}^2\mu}{2T}+c_p\frac{(\lambda T-\mu T_e)^2}{2T\mu}+\frac{K}{T}+c_\mu(\mu)\lambda,$$

Subject to

$$\mu T_e-\lambda T\le 0, \tag{4.1}$$

$$T_e-fT\le 0, \tag{4.2}$$

$$\lambda-\mu\le 0, \tag{4.3}$$

$$-T_e\le 0. \tag{4.4}$$

Constraint (4.1) is non-linear and guarantees that at most λT orders can be processed by the deadline. Constraint (4.2) ensures compliance with the prerequisite that the wave release time is at or after the cut-off time $t_w \geq t_c$ and therefore, $T_e - fT \leq 0$. Constraint (4.3) ensures that the picking rate is no less than the demand rate and Constraint (4.4) guarantees the non-negativity of the decision variable T_e.

Problem (P1) has a non-linear objective function with both linear and non-linear constraints. We apply the Karush-Kuhn-Tucker (KKT) necessary conditions to find the critical point and the Linear Independence Constraint Qualification (LICQ) to show that the KKT conditions hold true at optimality. We then show that this critical point is indeed optimal using KKT sufficient conditions (Bazaraa et al., 2006). The proofs are presented in Appendix B.2. We summarise the optimal solution structures in Table 8. Wolfram Mathematica 12 has been used to obtain derivatives, solve equations, and conduct the numerical analysis carried out in this Chapter.

Our analysis in Appendix B.2 also shows that constraints (4.1) and (4.4) cannot be binding at optimality. This indicates that there is always a fraction of orders that will be late and that T_e must be strictly positive (T_e = 0 would mean that $\mu = \infty$ at an infinite cost). Table 8 presents four relevant optimal solution structures: no binding constraints, only constraint (4.2) or (4.3) is binding, both constraints (4.2) and (4.3) are binding. When constraint (4.3) is binding, orders are processed at the arrival rate $\mu = \lambda$.

Table 8 Optimal solution when $t_w \geq t_c$ and $L = 0$.

Decision variable	No binding constraints	Constraint (4.2) is binding	Constraint (4.3) is binding	Constraints (4.2) and (4.3) are binding
${T_e}^*$	$\left(\frac{4Kc_p c'_\mu(\mu^*)}{{c_h}^2(c_h+c_p)}\right)^{1/3}$	fT^*	$\sqrt{\frac{2Kc_p}{c_h(c_h+c_p)\lambda}}$	$\sqrt{\frac{2Kf^2}{c_p(1-f)^2\lambda+c_h f^2\lambda}}$
T^*	$\frac{1}{\lambda}\left(\frac{2K^2(c_h+c_p)}{c_h c_p c'_\mu(\mu^*)}\right)^{\frac{1}{3}}$	$\sqrt{\frac{2K\mu^*}{c_h f^2 {\mu^*}^2+c_p(\lambda-f\mu^*)^2}}$	$\sqrt{\frac{2K(c_h+c_p)}{c_h c_p \lambda}}$	$\sqrt{\frac{2K}{c_p(1-f)^2\lambda+c_h f^2\lambda}}$
μ^*	$\left(\frac{Kc_h c_p}{2c'_\mu(\mu^*)^2(c_h+c_p)}\right)^{\frac{1}{3}}$	$\sqrt{\frac{c_p\lambda^2 T^*}{(c_h+c_p)f^2T^*+2c'_\mu(\mu^*)\lambda}}$	λ	λ
${TC_1}^*$	$\lambda\left(\frac{4Kc_h c_p c'_\mu(\mu^*)}{c_h+c_p}\right)^{\frac{1}{3}}$ $+c_\mu(\mu^*)\lambda$	$\frac{(c_h+c_p)f^2T^*\mu^*}{2}+\frac{c_p\lambda^2T^*}{2\mu^*}$ $-c_p\lambda fT^*+\frac{K}{T^*}+c_\mu(\mu^*)\lambda$	$\sqrt{\frac{2Kc_h c_p\lambda}{(c_h+c_p)}}$ $+c_\mu(\mu^*)\lambda$	$\sqrt{2K\lambda(c_p(1-f)^2+c_h f^2)}$ $+c_\mu(\mu^*)\lambda$

Note that when only constraint (4.2) is binding, the expressions for the three decisions variables are coupled, and their optimal values have to be calculated iteratively for a pre-set precision ε. This situation recurs later in this chapter (and depending on the expression for $c_\mu(\mu)$ may also happen in other cases) and the same iterative approach can be used to find the solutions for coupled decision variables. The iterative approach to estimate μ and T starts by setting e.g., $\mu_0 = \lambda$ and follows steps 1 to 4 with $i = 1$ initially:

- Step 1: Update the picking rate $\mu_i = \mu_{i-1}$
- Step 2: Compute the order receipt window $T_i = \sqrt{\frac{2K\mu_i}{c_h f^2 {\mu_i}^2+c_p(\lambda-f\mu_i)^2}}$
- Step 3: Compute the picking rate $\mu_{i+1} = \sqrt{\frac{c_p\lambda^2 T_i}{(c_h+c_p)f^2T_i+2c'_\mu(\mu)\lambda}}$
- Step 4: if $\mu_{i+1} - \mu_i > \varepsilon$, go back to step 1 with $i = i + 1$; otherwise, stop.

The relationships presented in Table 8 can be used to investigate the sensitivity of the three decision variables to changes in problem parameters. We analyse the sensitivity at the optimal interior point (i.e., when no constraints are binding):

- When the demand rate λ increases, the order receipt window T decreases. Thus, one mechanism that can be deployed by the retailer to handle a rising demand rate is to adjust the order receipt window.
- When the unit inventory holding cost c_h increases, both T_e and T decrease and μ increases. For products of higher values or that are perishable (i.e., high c_h), the

retailer should provide a shorter order receipt window and process orders at a higher rate. As a result, the picking operation can start at a later time, i.e., a shorter T_e.

- When the lateness penalty cost c_p increases, T_e and μ increase while T decreases. If it is very costly to miss the delivery deadline (i.e., a high c_p), the retailer will be obliged to enhance the picking rate and commence picking operations earlier (i.e., longer fulfilment service window) to eliminate the number of late orders. Another mechanism to reduce late deliveries is to reduce the number of accumulated orders by shortening the order receipt window T.
- When the marginal variable picking cost $c'_\mu(\mu)$ increases, T_e increases while T and μ decrease. A higher $c'_\mu(\mu)$ implies a higher cost to increase the picking rate, which will motivate the retailer to adopt a lower picking rate. A lower picking rate means that fewer orders will be processed within a specific time frame. Therefore, to avoid more late orders, the retailer should shorten the order receipt window to reduce the number of orders accumulated in a cycle and also start picking earlier (i.e., longer T_e).
- When the fixed order picking cost K increases, all the decision variables increase. Prolonging the order receipt window T results in fewer cycles within a specific time frame, thus incurring lower fixed order picking costs. A longer receipt window leads to more online orders in each cycle, which requires the retailer to commence picking at an earlier time (i.e., a longer T_e) with a higher picking rate.

4.2.3 The order fulfilment model when $L > 0$

When a retailer fulfils BOPS demand from several centralised facilities such as warehouses and then ships to other stores for customers to collect, the STS transportation window $L > 0$. We do not consider the detail of transportation operations here. We modify our original model by incorporating a fixed STS transportation window L and investigate how this affects the optimal decisions. We further assume that the inventory holding cost applies to all WIP during the STS transportation window L. Only orders processed L time units before t_d will be ready for collection on time, and orders will only be late and therefore penalised if available after t_d. After t_d, late orders will be shipped to the designated collection stores once being picked in order to avoid further disappointment to customers. Therefore, the unit lateness penalty cost c_p also captures the expenses for potentially inefficient immediate transportation of late orders. Because retailers are generally interested in providing a cut-off time as late as possible to capture last-minute shoppers and enhance their competitiveness, our analysis focuses on the case where the retailer starts picking exactly at the order cut-off time, i.e., $t_c = t_w$. The case where $t_w > t_c$ can be explored in a similar fashion but is much

more lengthy and slightly more complex. The fulfilment service window T_e (i.e., $t_d - t_w = t_d - t_c$) is the collection speed, and cannot be shorter than L. T_e increases when L becomes longer. The WIP inventories and late orders depend on L and T_e. We discuss two cases below, corresponding to a short and long STS transportation window L. In Section 4.3 we elaborate on how we model the variable order processing cost $c_\mu(\mu)$ and we explain how a transportation cost per order can be incorporated.

Case 1: Short T_e given a short L (see Figure 22**),** i.e., $L \leq T_e \leq \lambda T/\mu$.

We use an example to explain. Assume a retailer receives $\lambda = 50$ orders/hour and processes orders with a rate $\mu = 100$ orders/hour. The retailer has $T = 1$ day and a daily order cut-off time t_c at 8am with a promised collection deadline t_d at 6pm same day ($t_d - t_c = 10/24$ day). Only orders picked L time units before the deadline can be prepared on time. For example, if the retailer has a 2-hour STS transportation window, orders picked before 4pm (8h×100 orders/hour = 800 orders) will be shipped to stores on time for collection. If L increases to 4 hours, orders picked before 2pm (600 orders) will be on time and the number of late orders increases (see Figure 23). Orders incur a lateness penalty if fulfilled after the deadline. Therefore, the average WIP inventories remain the same and the average lateness increases when L increases.

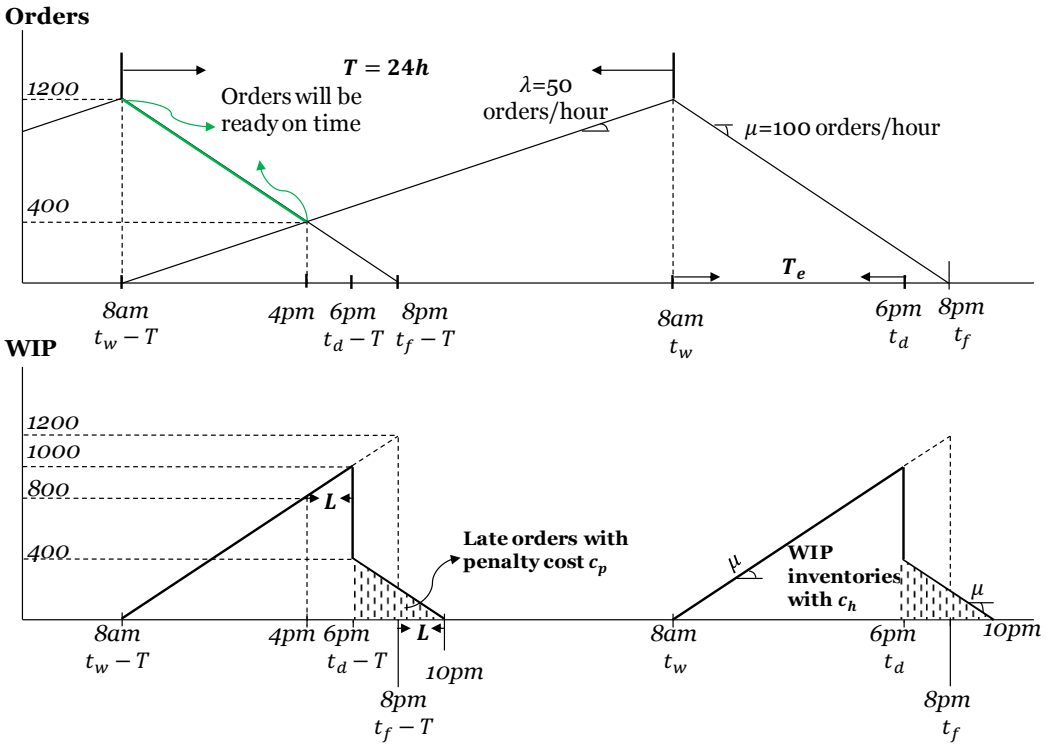


Figure 22. The evolution of order accumulation, WIP inventories and late orders over time when L = 2h.

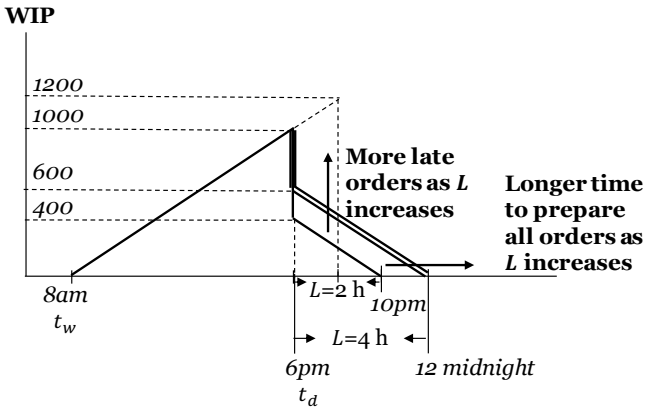


Figure 23. The impact of STS transportation window L on the number of late orders.

As shown in Figure 22, the average WIP inventories are $\frac{1}{T}\left(\frac{\mu T_e}{2} * T_e + 0 * (T - T_e)\right) = \frac{\mu {T_e}^2}{2T}$. The number of late orders are $\lambda T - \mu(T_e - L)$, which leads to the average late orders per time unit $\frac{1}{T}\left(\frac{(\lambda T - \mu(T_e - L))}{2} * \frac{(\lambda T - \mu(T_e - L))}{\mu}\right) = \frac{(\lambda T - \mu(T_e - L))^2}{2T\mu}$. The BOPS fulfilment problem for this case is

$$\text{(P2)} \qquad \text{Min } TC_2(T_e, T, \mu) = c_h \frac{\mu {T_e}^2}{2T} + c_p \frac{(\lambda T - \mu(T_e - L))^2}{2T\mu} + \frac{K}{T} + c_\mu(\mu)\lambda,$$

Subject to

$$\begin{aligned} \mu T_e - \lambda T &\le 0, \\ \lambda - \mu &\le 0, \\ L - T_e &\le 0. \end{aligned} \tag{4.5}$$

The new constraint (4.5) ensures that the service level is always higher than zero. We follow the same approach as Section 4.2.2 to obtain the optimal solutions (see Appendix B.3). The results for different relevant sets of binding constraints are summarised in Table 9.

Table 9 Optimal solution for $L > 0$ - case 1 ($L \le T_e \le \lambda T/\mu$).

Decision variable	No binding constraints	$\mu T_e - \lambda T \le 0$ is binding	$\lambda - \mu \le 0$ is binding	$\mu T_e - \lambda T \le 0$ and $\lambda - \mu \le 0$ are binding
${T_e}^*$	$\frac{c_p \lambda T^* + c_p L \mu^*}{(c_h + c_p)\mu^*}$	$\lambda T^*/\mu^*$	$\frac{c_p(T^* + L)}{(c_h + c_p)}$	$\sqrt{\frac{2K + c_p L^2 \lambda}{c_h \lambda}}$
T^*	$\sqrt{\frac{2K\mu^*(c_h + c_p) + c_h c_p L^2 {\mu^*}^2}{c_h c_p \lambda^2}}$	$\sqrt{\frac{2K\mu^* + c_p L^2 {\mu^*}^2}{c_h \lambda^2}}$	$\sqrt{\frac{2K(c_h + c_p) + c_h c_p L^2 \lambda}{c_h c_p \lambda}}$	$\sqrt{\frac{2K + c_p L^2 \lambda}{c_h \lambda}}$
μ^*	$\sqrt{\frac{c_h c_p \lambda^2 {T^*}^2}{c_h c_p L^2 + 2(c_h + c_p)c'_\mu(\mu^*)\lambda T^*}}$	$\sqrt{\frac{c_h \lambda^2 {T^*}^2}{c_p L^2 + 2c'_\mu(\mu^*)\lambda T^*}}$	λ	λ
${TC_2}^*$	$c_h \frac{\mu^* {T_e}^{*2}}{2T^*} + c_p \frac{(\lambda T^* - \mu^*({T_e}^* - L))^2}{2T^*\mu^*} + \frac{K}{T^*} + c_\mu(\mu^*)\lambda$			

Table 9 shows that when L increases, both T_e and T increase, whilst μ decreases or levels off at the level of demand rate. The expressions for the three decision variables are coupled and the optimal results can be calculated with same iterative procedure described in Section 4.2.2.

Case 2: Long T_e given a long L (see Figure 24**),** i.e., $\lambda T/\mu < T_e \leq \lambda T/\mu + L$.

If L is long, customers have to wait longer for collection, i.e., T_e is prolonged. The evolution of WIP orders shows a different pattern in this case (see Figure 24). We illustrate the evolution of WIP inventories using the same parameters as above but with $L = 2$ days and the retailer promises that orders placed by 9am today will be ready by 6pm two days later. In Figure 24, the retailer accumulates 1200 orders and finishes picking at 9pm each cycle. Given the 2 days STS transportation window, orders picked by 6pm today will be available on time by 6pm two days later as promised, and orders picked from 6pm to 9pm (300 orders) will be late. WIP inventories build up during picking operations and remain constant at a level of 1200 until the deadline when late orders are penalized. Because L is long, the total WIP may come from different cycles. In the example in Figure 24, in a the cycle between t_{w3} and t_{w4}, the WIP inventory level is comprised of three components from three different cycles:

- $\lambda T(t_{d1} - t_{w3}) = \lambda T(T_e - L)$ from picking operations between t_{w1} to t_{f1};
- $\lambda T(t_{w4} - t_{w3}) = \lambda T^2$ from picking operations between t_{w2} to t_{f2};
- $\lambda T\left(t_{w4} - t_{f3}\right) + \frac{\lambda T}{2}\left(t_{f3} - t_{w3}\right) = \lambda T\left(T - \frac{\lambda T}{\mu}\right) + \frac{\lambda^2 T^2}{2\mu}$ from picking operations between t_{w3} to t_{f3}.

Because $L = 2T$, the total WIP inventory level is $\lambda T T_e - \frac{\lambda^2 T^2}{2\mu}$ and the average WIP inventory level per time unit is $\lambda T_e - \lambda^2 T/2\mu$ (2550 orders/day in this example). When L and T_e become larger (smaller), more (less) cycles may contribute to WIP but the average WIP inventory level per time unit is calculated with the same expression: $\lambda T_e - \lambda^2 T/2\mu$.

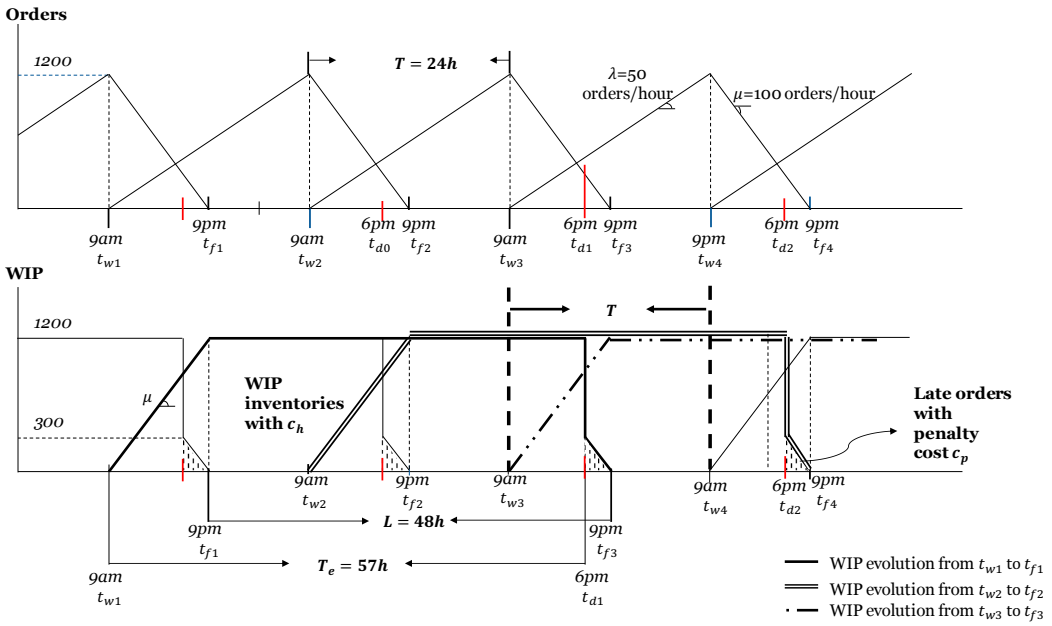


Figure 24. The evolution of order accumulation, WIP inventories and late orders over time when L = 2days.

The number of late orders in each cycle is $\lambda T - (T_e - L)\mu$ and the average late orders per time unit remain as $\frac{(\lambda T - \mu(T_e - L))^2}{2T\mu}$. The BOPS fulfilment problem in this case is

$$\text{(P3)} \quad \text{Min } TC_3(T_e, T, \mu) = c_h(\lambda T_e - \frac{\lambda^2 T}{2\mu}) + c_p \frac{(\lambda T - \mu(T_e - L))^2}{2T\mu} + \frac{K}{T} + c_\mu(\mu)\lambda,$$

Subject to

$$\begin{aligned} \lambda T - \mu T_e &\leq 0, \\ \mu(T_e - L) - \lambda T &\leq 0, \\ \lambda - \mu &\leq 0, \\ L - T_e &\leq 0. \end{aligned} \tag{4.6}$$

The new constraint (4.6) means that at most λT orders will be fulfilled on time. The same approach is used to derive the optimal solution (see Appendix B.4). Constraint (4.6) cannot be binding at optimality, reiterating that the retailer should not fulfil all accumulated orders on time from a cost minimisation perspective. The optimal solution is summarised in Table 10.

Table 10 Optimal solution for $L > 0$ - case 2 ($\lambda T/\mu < T_e \leq \lambda T/\mu + L$)

Decision variable	No binding constraints	$\lambda - \mu \leq 0$ is binding
${T_e}^*$	$(\frac{4Kc'_\mu(\mu^*)(c_p - c_h)}{c_p {c_h}^2})^{\frac{1}{3}} + L$	$\sqrt{\frac{2K(c_p - c_h)}{c_h c_p \lambda}} + L$

T^*	$\frac{1}{\lambda}\left(\frac{2K^2c_p}{c_h c'_\mu(\mu^*)(c_p-c_h)}\right)^{\frac{1}{3}}$	$\sqrt{\frac{2Kc_p}{c_h(c_p-c_h)\lambda}}$
μ^*	$\left(\frac{Kc_h(c_p-c_h)}{2c_p c'_\mu(\mu^*)^2}\right)^{\frac{1}{3}}$	λ
${TC_3}^*$	$\lambda\left(\frac{4Kc_h c'_\mu(\mu^*)(c_p-c_h)}{c_p}\right)^{\frac{1}{3}}$ $+c_h\lambda L + c_\mu(\mu^*)\lambda$	$\sqrt{\frac{2Kc_h(c_p-c_h)\lambda}{c_p}}$ $+c_h\lambda L + c_\mu(\mu^*)\lambda$

Table 10 shows that in case 2, the STS transportation window L directly affects the fulfilment service window T_e, whilst the order receipt window T and the picking rate μ are independent of L. It can be shown that the impacts on other problem parameters are similar to the $L = 0$ case.

As noted earlier, L is an indicator of the centralisation level of BOPS fulfilment operations. Although T is not directly affected by L, it is not entirely independent of the picking centralisation level. The more centralised the fulfilment operations, the higher the demand level but the longer the L. Therefore, a longer L is often associated with an increased λ, which in turn allows for a shorter T. The exact impact of the picking centralisation level on the optimal solutions needs further exploration of the relationship between L and λ, a topic which could be pursued in future work.

4.3 Retail operating strategies

We use publicly available data for 5 major UK-based retailers to apply our model and evaluate their BOPS service offerings. Specifically, we provide insights on why some retailers set a short order receipt window of only a few hours while others offer a much longer receipt window. In addition, we cast light on why the speed of store collection offered by retailers differs in different retail sectors. Building on these insights, we present two frameworks to show how the speed of BOPS offerings should be tailored based on the characteristics of products and of retailers.

4.3.1 Estimation of model parameters for major UK retailers

We consider the current BOPS service offerings of the following UK retailers - Tesco and Sainsbury's in the grocery sector, M&S and Tu clothing from the J Sainsbury's Group in clothing, Habitat from the J Sainsbury's Group in homeware, Boots in health and beauty and WH Smith in books and stationery. Their current BOPS practices are summarised in the first three rows of Table 11. We estimate the model parameters shown in the remaining rows from

publicly available data. We explain the rationale for the case of Tesco below. The details for the other retailers and the corresponding parameter estimations are given in Appendix B.5.

Table 11 Model parameters

	Tesco	Sainsbury's	Tu clothing	M&S Clothing	Boots	Habitat	WH Smith
Sector	Groceries	Groceries	Fashion	Fashion	Health & beauty	Homeware	Books & Stationery
Advertised Ordering window T (day)	0.5	1	1	1	1	1	3.5
Cut-off time	9am (1pm)	12noon	2pm	10pm	8pm	N/A	N/A
Earliest collection time	12noon (4pm) same day	4pm same day	12noon next day	12 noon next day	12 noon next day	3-5 working days	Every Wednesday and Friday
Annual online sales (£m)	3702.3	1722	100	535	171.3	19.5	59
BOPS percentage	30%	23%	80%	71%	75%	30%	12%
No. of items / order	60-80	60-80	1-3	1-3	1-3	1-3	1-3
Average Order Value (AOV) (£)	100	100	70	70	50	30	15
No. of BOPS orders per day (×1000)	30.43	10.851	3.13	13	7.04	0.534	1.293
No. of picking facilities	291	124	1	19	1	2	3
No. of collection points	320	124	1000+	1000+	2200	2000+	600+
L (day)	0	0	0.75	0.5	0.5	3	3
λ (orders per day per facility)	105	88	3132	783	7040	268	432
c_h	0.5	0.5	0.058	0.058	0.041	0.025	0.012
c_p	123.3	189.5	2.877	2.877	2.055	1.233	0.616
c_0	5	5	0.25	0.25	0.25	0.25	0.25
c_t	0	0	0.5	0.25	0.6	0.25	0.25
μ_0	400	400	480	480	480	480	480
α	0.2	0.2	0.015	0.015	0.015	0.015	0.015
K	20	20	100	100	100	50	50

The demand level for Tesco, i.e., the number of BOPS orders received per day, is approximated by dividing the daily BOPS sales by the Average Order Value (AOV). The British online grocery market size was approximately £12.3 billion in 2018 (Mintel, 2019). As the market leader, Tesco accounts for 30.1% of the sales or £3702.3 million. On average, 23% of online shoppers choose store collection for their groceries in the UK (Mintel, 2019). For Tesco, the number of customers collecting their online groceries in stores has increased to over 10% (Tesco, 2019). We assume that BOPS sales represent 30% of total online sales in Tesco. With an AOV in Tesco of £100 (Internet Retailing, 2010), the daily BOPS orders received in Tesco are estimated at 30,430 ($\frac{£12.3bn*30.1\%*30\%}{£100*365} \approx 30430$).

Tesco has rolled out BOPS services to 320 of its stores, among which 291 stores offer same-day collection (Mintel, 2019). Most of the stores selected for BOPS services have online picking operations but a few locations, including four express stores, do not process online orders. For these locations, groceries are fulfilled in a nearby picking store before being transported for customer collection (Tesco, 2014). Hence, we make the assumption that Tesco processes BOPS orders in these 291 stores where customers can pick up their groceries 3 hours after the cut-off time. With such a decentralised BOPS fulfilment strategy, the STS transportation window is negligible and L is assumed to be zero. The daily BOPS order rate in each of these 291 stores is therefore 105 orders/day (30430/291 ≈ 105).

For groceries with temperature-sensitive items, the unit inventory holding cost c_h is high. For example, Walmart has used giant self-service kiosks with refrigerators and freezers for customers to collect online groceries (Business Insider, 2017). We assume c_h = £0.5 per order per day for Tesco.

It is not uncommon for retailers to offer 10% discount if deliveries miss the promised due date (Marino et al., 2018), but we have calculated the penalty cost c_p for a pre-specified *NSD* service level. The *NSD* service level is the fraction of cycle orders that meet the deadline. It is calculated as $NSD = \mu^*(T_e^* - L)/\lambda T^*$. By substituting the optimal expressions for the three decision variables from Tables 8, 9 or 10, relationships can be obtained between c_p, c_h, *NSD* and possibly, other model parameters. For example, for the optimal interior point in Table 8, the service level expression reduces to $NSD = c_p/(c_p + c_h)$ from which c_p can be calculated: for Tesco with c_h = £0.5 and *NSD* = 0.98, we obtain a penalty cost c_p = £24.5 per order per day. The penalty cost c_p = £24.5 may not be high enough to reach the required service level when constraints are binding. Using the results from Table 8 when constraint (4.2) is binding, we obtain: $c_p = NSD^2(c_h + 2\lambda c'_\mu(\mu^*)/f^2T^*)/(1 - NSD^2)$. This

expression needs to be solved iteratively (together with μ^*, T^* and T_e^*), each time providing a new estimate for c_p, until the desired service level is reached. With this approach we obtain for Tesco, c_p = £123.3 for an *NSD* = 0.98 (with c_p = £24.5, a service level *NSD* = 0.90 is reached).

We model the variable picking (and transportation) cost through the function $c_\mu(\mu) = c_t + c_0(1+\alpha)^{\frac{\mu}{\mu_0}-1}$, where c_t is an optional transportation cost per order (for Tesco, c_t = 0 because L = 0), c_0 is the unit variable picking cost when $\mu = \mu_0$ and α is the discount factor for picking efficiency due to congestion ($0 < \alpha < 1$), i.e., α represents the dependence of the variable picking cost on congestion. When $\mu = \mu_0$, the system is least congested with minimum picking efficiency deterioration and the variable picking cost is the lowest, i.e., c_0. However, if a retailer doubles the picking rate to expedite fulfilment, the system becomes more congested and picking efficiency degenerates, so that the variable picking cost increases to $(1+\alpha)c_0$. c_0 can be interpreted as the sum of order handling costs within the picking facility. It is worth highlighting that the cost function for $c_\mu(\mu)$ in this study is valid for the practical operational range $\mu \geq \lambda$ considered here. Future work may explore different cost functions to consider the case when $\mu < \lambda$ as well (e.g., $\mu = 0$ as a special case). The unit order picking and handling for groceries is approximately £5 (McKinsey, 2013) and therefore $c_0 = £5$ for Tesco.

An online grocery order typically comprises 60-80 items across 3 temperature regimes (Fernie and Spark, 2014). Processing efficiency deteriorates significantly when stores are congested with walk-in customers. We assume α = 20% for Tesco and μ_0 = 400 orders/day (i.e., about 5% lower than the rate required (420 per day) to pick the daily orders in two 3-hour picking cycles). As a result, the variable picking cost function for Tesco is $c_\mu(\mu) = 5(1.2)^{\frac{\mu}{400}-1}$ .The fixed picking cost in Tesco is assumed to be £20 because the daily BOPS volume in each picking facility is relatively low. Same day groceries are often collected $T_e \approx$ 3 hours after the cut-off time and Tesco has two cycles per day (T = 12 hours), so we have set $f \approx T_e/T = 0.25$.

4.3.2 Theoretical optimality for retailers

With the input of problem parameters in Table 6, we apply the appropriate analytical relationships from Tables 8, 9 and 10 to compute the optimal solutions for the order receipt window T, the fulfilment service window T_e, the picking rate μ, and the total fulfilment costs *TC* (see Table 12) for each retailer. To highlight the cost reduction benefits of our models, we also compute *TC* given the current service offerings from these retailers. In the

computation we use the retailer's actual T and assume that the retailer starts order processing at the cut-off time ($T_e = t_d - t_c$). In the case of Tesco and Sainsbury's we use our optimal μ^* (which guarantees $NSD = 0.98$, because constraint (4.2) is binding in the optimum). For the current offerings of other retailers, we calculate μ so that 98% of the cycle orders λT are processed in $T_e - L$.

Table 12 Model results for retailer service offerings ($NSD = 0.98$)

Retailer		Tesco	Sainsbury's	Tu Clothing	M&S clothing	Boots	Habitat	WH Smith
Theoretical optimum	T^*(day)	1.679	2.193	1.108	2.811	0.840	10.966	7.789
	T_e^*(day)	0.420	0.365	1.752	1.290	1.323	4.380	5.411
	μ^*(orders /day)	411.6	517.4	3394.9	2731.7	7040	2087.2	1367.8
	TC^* per facility (£)	551.6	488.4	2738.8	463.8	6763.9	166.4	247.8
Results with current service offerings	T (day)	0.5	1	1	1	1	1	3.5
	T_e (day)	0.125	0.167	22/24	14/24	16/24	3.5(2)	3.5(2)
	μ (orders /day)	411.6	517.4	18416.2	9208.1	41395.2	525.1	2962.5
	TC per facility (£)	571.5	494.2	3181.9	526.8	10754.3	205.6	256.2
Cost reduction(1) (%)		3.58	1.23	16.18	13.59	58.99	23.56	3.39

(1): Cost reduction = (TC^* with current service offerings - TC^* in theoretical optimum)/TC^* in theoretical optimum

(2): Estimate

Table 12 shows varying BOPS service offerings for retailers across different sectors. Even within the same organisation, J Sainsbury's group offers different order receipt windows and collection speeds for BOPS customers in its groceries, clothing and homeware businesses. For all retailers, the order picking costs are dominant while the inventory related costs contribute typically to less than 20% of the total cost. The optimal solutions for Tesco and Sainsbury's groceries show longer order receipt windows than their current actual offerings (1.679 days and 2.191 days vs. 0.5 days and 1 day), but the percentage cost differences between the actual and optimum are small (3.58% and 1.23%). The results are sensitive to the values of the model parameters. For example, with fixed picking costs K = £5, £50 and £100, the optimal order receipt window for Tesco changes to T^* = 0.804 days, T^* = 2.700 days and T^* = 3.847 days respectively, with cost reduction percentages of 0.26%, 11.75% and 26.11%. Our results indicate that it can be justifiable for retailers to offer multiple ordering slots in a day (i.e., $T < 1$), and a collection deadline in less than 3 hours after the cut-off time (i.e., $T_e < 0.125$), especially when K/c_h is low. In 2017, Tesco extended its same-day BOPS service (order by 1pm for collection from 4pm) by introducing a second ordering

slot (order by 9am for collection from midday) in response to the increasing BOPS demand (The Sun, 2017).

In contrast, retailers for homeware and books & stationery should offer a much longer order receipt window. Habitat currently follows the common practice of setting $T = 1$ day, which is far from optimal ($T^* = 10.966$ days) and incurs an increase of over 23% in the total fulfilment costs. WH Smith sets a $T \approx 3.5$ days receipt windows for BOPS customers each week, which is more in agreement with the optimal model results. The most significant costs savings are highlighted for Boots, who promise store collection availability 16 hours after the cut-off time across the UK. Boots run one large automated warehouse from where they supply over 2500 stores on a daily basis through a fast distribution network of cross-dock facilities. With an estimated transportation window $L = 12$ hours, the current order processing window (4 hours) is much shorter than optimality ($T_e - L = 1.323$ days $- 0.5$ days $= 0.823$ days $\approx T^*$), resulting in a high order processing rate and a high cost. Our optimal model results for Boots point to a continuous order processing policy at a rate equal to the demand rate. The Boots warehouse operates 24/7 and our benchmark assumptions may not accurately reflect the current situation or our estimation of L may be too high. For example, if we assume transportation windows $L = 8$ hours and $L = 4$ hours, the percentage cost difference between the optimal solution and the actual benchmark are 15.6% and 5.37%. Table 12 demonstrates that some retailers could make substantial cost improvements with the application of the scientific decision-making models presented here. It emphasises the necessity of studies such as this research to provide guidance, insights and models for retailers to better design and operate their BOPS service offerings.

Because the model variables are continuous, the optimal solutions in Table 12 have to be adjusted for practical use. Naturally, retailers need to offer T and T_e as a sum of hours or days, such as 6 hours or one day or one week, to facilitate clear communication with customers and to avoid confusion. For example, for Tu clothing, $T^* = 1.108$ days, $T_e^* = 1.752$ days and $L = 0.75$ days, the retailer could offer $T = 1$ day and $T_e = 1.75$ days. Customers can then collect their orders 1.75 days (or 42 hours) after the daily cut-off time (e.g., order by 6pm day X, collect from 12noon on day X+2). The cost difference between the optimal solution and adjusted Tu Clothing solution is only 0.04%. Habitat, with $T^* = 10.966$ days, $T_e^* = 4.380$ days and $L = 3$ days, could offer a weekly BOPS service $T = 7$ days and $T_e^* = 4.5$ days, (e.g., order by midnight on Sunday, collect on Friday from 12 noon) with a cost increase over the optimal solution of 0.68% only. These two examples show that practical solutions can be derived easily and that based on the cost implications some retailers may benefit from reconsidering their current BOPS service offering for store collection.

Our models can also be used to analyse the impact of the level of fulfilment centralisation. The current BOPS services in Tesco appear to be near-optimal, given the strategic decision to operate a highly decentralised fulfilment strategy from 291 stores. We show in Figure 25 the percentage difference in fulfilment cost by decreasing the number of picking facilities (or concentrating BOPS demand in fewer stores) vs. the current situation when 291 stores are used. We show results for K = £20 and K = £50 and for two store picking congestion rates $\alpha = 0.2$ and $\alpha = 0.01$, and $NSD = 0.98$.

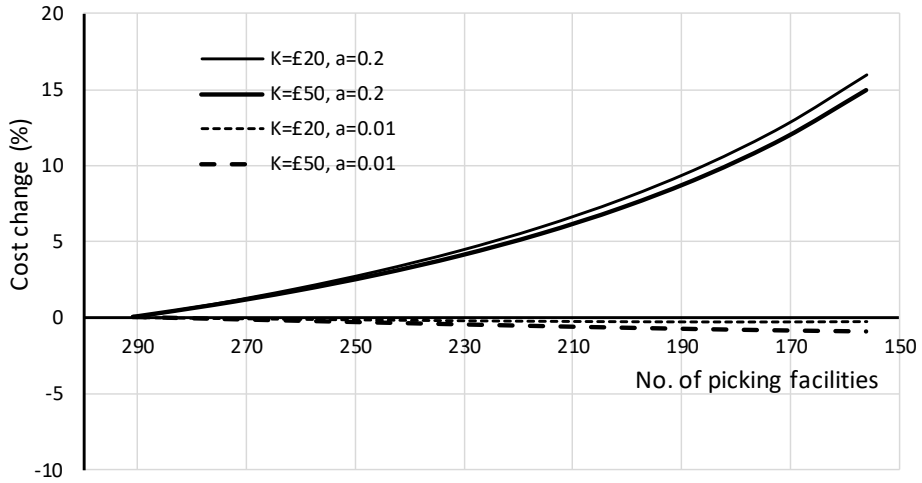


Figure 25. Total fulfilment cost change (%) for Tesco when the total BOPS demand is picked in fewer stores.

Figure 25 demonstrates that the cost impact from picking centralisation may depend on the relative importance of inventory costs vs. store congestion cost. Centralising picking operations saves in inventory cost but increases the picking cost due to more congestion in the picking stores. For very low α, we notice for Tesco a net saving from picking consolidation, whereas, if α is higher, a more centralised picking operation is more costly. Centralising picking stores also reduces the cost of infrastructure and equipment for picking operations, but will require longer travel distances for STS and therefore higher transportation costs. Our models can be extended to cast light on determining the optimal centralisation level of fulfilment by considering the trade-offs between the total fulfilment costs, STS transportation costs and fixed investment costs, which we pursue in chapter 5.

4.3.3 Guidance frameworks for BOPS speed offerings

To highlight the variance of store collection speed offerings across product categories with different order values, we show the optimal fulfilment speed ($T_e^* - L$) from Table 12 in Figure

26. Using the expression for T_e^* from Table 8 (or $T_e^* - L$ from Table 10) and assuming that a high *NSD* service level is desirable (so that $c_p/(c_p+c_h)$ or $(c_p-c_h)/c_p \approx 1$), we obtain $(T_e^* - L)c_h^{\frac{2}{3}} = (4Kc'_\mu(\mu^*))^{\frac{1}{3}}$. We can therefore plot iso-$(T_e^* - L)c_h^{\frac{2}{3}}$ contour lines, which we did for the values 0.1, 0.15 and 0.2 in Figure 26. Each retailer is also represented as a marker, showing its optimal collection speed $T_e^* - L$ and value for $c_h^{2/3}$. Figure 26 shows that optimal collection services are much faster for products with high holding cost or order value. Specifically, the optimal BOPS service is the fastest for grocery retailers, followed by health & beauty retailer and then fashion retailers, whilst the optimal BOPS service is slowest for homeware and books & stationery retailers. We dissect the impact of order value and demand rate as follows.

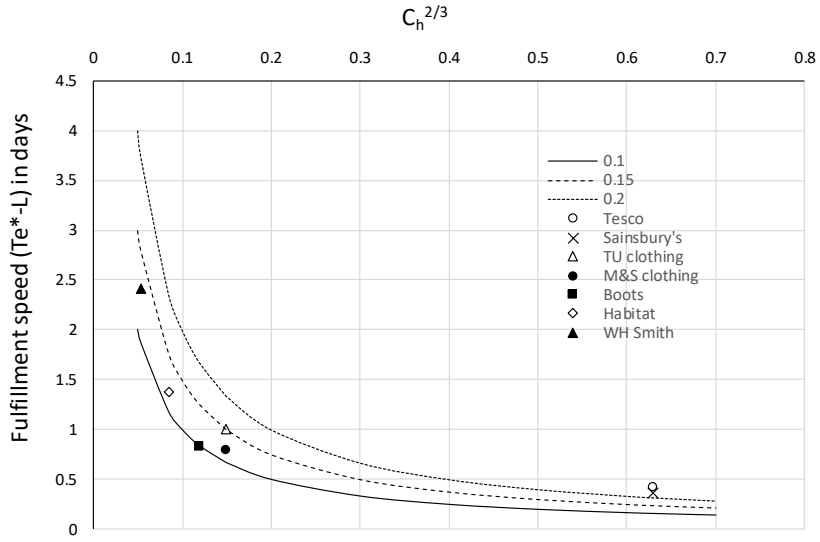


Figure 26. Optimal BOPS service speed for different retailers

The impact of demand rate. Our analytical results in section 4.2 have shown that both the order receipt window and the fulfilment service window decrease when the demand rate increases. The comparison of the results for Tesco and Sainsbury's highlight this pattern. With a higher demand level, Tesco could offer a shorter receipt window (smaller *T*) and a faster fulfilment service (smaller T_e). This is consistent with the current practice where customers in Tesco can collect BOPS orders 3 hours after the cut-off time while customers in Sainsbury's have to wait for 4 hours. BOPS is a win-win strategy for both retailers and customers and its market in the UK projected to rise by 45.8% over the next five years to reach £9.6 billion by 2022 (GlobalData, 2018b). We anticipate that BOPS service will be expedited further and same-day collection services will be more prevalent across different retail sectors owing to the growth of BOPS volumes.

The impact of order value. The higher the order value, the higher the unit holding cost and the faster the collection service. Table 12 and Figure 26 show that Tesco and Sainsbury's provide the fastest collection service, even though they have the lowest daily BOPS volume in each picking facility. This is mainly attributed to their high order values and unit holding costs due to high perishability. The impact of order value can also be inferred from the comparison between Habitat and WH Smith. Compared to WH Smith, Habitat has a higher AOV and can offer faster BOPS services even though it has a lower daily BOPS volume per picking facility. These results are consistent with those in Zhang et al. (2019a), which show empirically that omni-channel retailers offer shorter fulfilment time for products with higher prices.

A high demand rate could be the result of the nature of products (i.e., fast moving) or the characteristics of retailers (i.e., market leader). The order value is associated with product types while the length of STS transportation window (i.e., the centralisation level of picking and/or collection facilities) is linked to the distribution network of the retailer. Using the analytical results and the insights from retail practice above, we provide two normative frameworks to guide retail operating strategies based on product profile and retail profile.

Figure 27 summarises the results of the analysis with regard to demand level and product value and/or perishability in a general way. It delineates the regions where same day, next day, or longer cycles may be most desirable. It indicates that fast moving products with high value and/or perishability are more suitable for fast collection services. This pattern correlates with industry practice where fast moving and highly valuable/perishable products such as groceries and consumer electronics are offered with same day collection services. In contrast, standard or slow collection services can be sufficient for slow moving products of low value, such as collect after 5 working days. This framework resonates with the study of Chopra (2016), which suggested that online retailers should compete on low cost for slow moving products and premium delivery services for fast moving products.

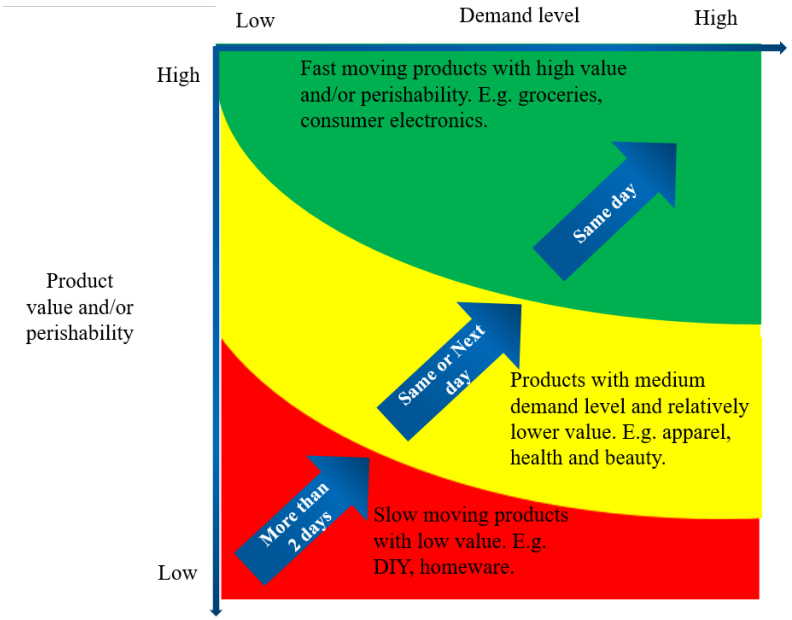


Figure 27. Optimal retail operating strategies based on product profile.

Figure 28 summarises where retailer practices are expected to differ with respect to two important retailer characteristics - market share and fulfilment network density. It shows that the higher a retailer's market share and the more decentralised the retailer's online fulfilment processes, the faster the BOPS service that can be offered. A high market share implies a high demand level. A large number of facilities such as stores and distribution centers (DCs) deployed for BOPS fulfilment means a more decentralised picking strategy and a shorter STS transportation window. Both characteristics enable retailers to offer and provide a faster fulfilment service.

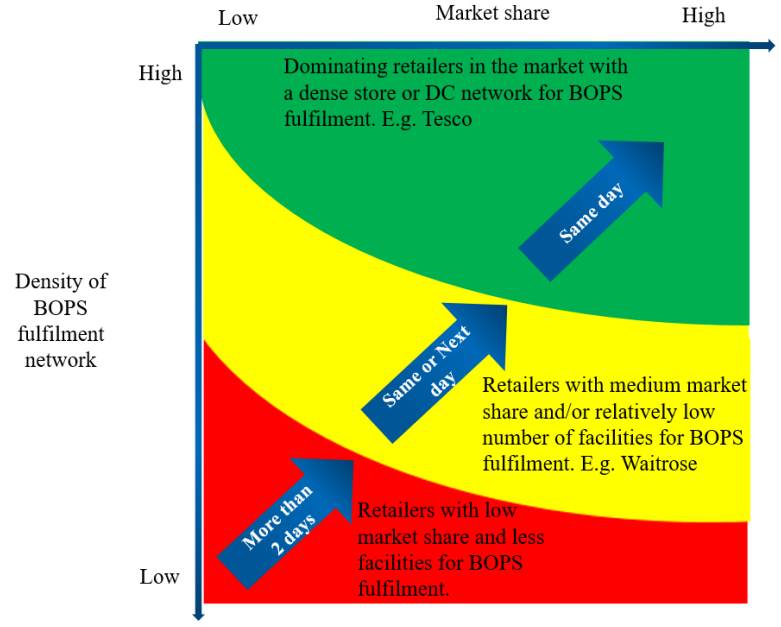


Figure 28. Optimal retail operating strategies based on retailer profile.

This tactic is being used by retail giants such as Walmart in the US and Tesco in the UK to leverage their vast store networks for online fulfilment to compete with Amazon (Forbes, 2018). Target plans to increase the number of stores offering BOPS services in the US, from around 1000 to all stores by the end of 2019 and Walmart is set to roll out its grocery pickup program from around 2100 stores to 3100 stores by the end of the fiscal year (Bloomberg, 2019). The analysis indicates that the increased density of stores offering BOPS services may allow these retailers to expedite BOPS services and compete, if not overtake, pure online retailers such as Amazon. In contrast, for retailers with a lower market share and deploying fewer facilities for processing online orders, the optimal fulfilment service window will be longer and therefore less competitive promises for store collection speed can be made. For example, compared to Tesco, the high end grocery retailer Waitrose represents 4% of online grocery market and fulfils its BOPS orders from approximately 180 stores (Mintel, 2019). It only offers next-day collection for BOPS customers compared to the same-day service in Tesco, which may be viewed as uncompetitive in today's grocery market.

As schematic representations of the insights generated from our model results and analysis, the two normative frameworks illustrated in Figures 27 and 28 suggest how product and retailer related factors come into play in designing BOPS service speed offerings. Such frameworks can provide very general guidance for practitioners, but the exact decision for BOPS service speed offerings should also consider other aspects and dynamics, such as customers sensitivity on collection speed and the corresponding fees (as discussed in section 2.6 in chapter 2 for joint pricing and delivery optimisation) and real-time information on product availability (as discussed in section 2.5.2 for dynamic online order allocation). Further work can extend our research to exploit big data availability and data analytics and modelling techniques to generate classifications with more granular details to guide retailers to make and adjust decisions more frequently regarding BOPS service speed offerings.

4.4 Concluding remarks

Motivated by the increasing prevalence and importance of BOPS services and the diversity in offerings given by retailers, this chapter has examined the optimal BOPS service window that a retailer should offer. Specifically, the model developed helps retailers determine the optimal order receipt window, the time to initiate order processing and the required order processing rate for BOPS orders for two types of fulfilment strategies – locally picked and shipped-to-store. A retailer could adopt a decentralised processing strategy by fulfilling BOPS demand from local stores. In contrast, for a centralised processing strategy, the retailer ships orders from a few facilities to other branch stores for customer collection, thus incurring a non-zero transportation window.

We develop a cost minimisation model to optimise the order receipt window, the time to start order processing, and the processing rate needed for fulfilling BOPS orders with a given collection deadline. The model generates the optimal time gap between the cut-off time and delivery deadline, which is essentially the speed of BOPS services. In the contemporary retailing world, fulfilment speed is a major element of competitive advantage for retailers (Marino et al., 2018). Some retailers aspire to provide ambitious collection services to woo customers, regardless of the adverse cost implications whilst others are very cautious in tweaking their BOPS service offerings and may therefore miss out on competitive opportunities. The model and analysis help retailers to strike a balance between the fulfilment service offering and fulfilment costs.

Our work is the first to investigate analytically the impact of product and retailer characteristics on store collection speed offered by retailers. We have examined the implications of the analysis for a set of examples from contemporary practice in different retail product categories. Building on the insights gained from applying the models to analyse the BOPS service offerings of 5 major UK-based retailers, we present two frameworks of retail operating strategies based on product types and retailers' profiles, showing the scenarios under which retailers could offer different speeds of collection services. The analysis also indicates that more retailers are likely to offer same day collection if the level of BOPS sales continue to grow as predicted. We show that it may not necessarily be optimal for retailers to follow the common practice and set the order receipt window as one day from a cost perspective. The analysis reveals that some retailers (e.g. Tesco and Sainsbury's) could provide an order receipt window less than one day, offering two or more slots for BOPS services in a day. In contrast, retailers should offer a longer order receipt window if the demand rate and product value is low, such as the order receipt window of more than 2 days offered by WH Smith.

We examine the impact of the centralisation level of fulfilment operations by varying the number of stores used for BOPS fulfilment in Tesco. The total fulfilment cost may increase or decrease depending on the store picking congestion when fulfilment is more centralised. Additionally, a more centralised policy requires less investment in obtaining online fulfilment capabilities. However, it will lead to a long STS transportation window, which increases transportation costs and impedes the collection speed that can be offered. The analysis reveals the trade-offs to be considered by retailers when deciding the number of facilities used for BOPS fulfilment, i.e., the level of centralisation of picking operations.

In this chapter, we use lateness penalty cost to guarantee delivery service, which may be difficult to measure precisely in some cases. The models presented can be adjusted to replace the penalty cost with a service level to limit the number of late orders. Our models can be extended in several ways. First, the assumptions of constant demand rate and processing rate can be relaxed to investigate whether and how the model results and insights may change. Second, further exploration on the impact of the centralisation level of BOPS fulfilment operations can be pursued by investigating the relationship between the STS transportation window and the demand rate captured in each picking facility. Third, here we have focused on BOPS service only and therefore there is only one class of orders processed in each facility. The model could be developed to account for fulfilment facilities handling two classes of online orders, e.g., BOPS orders and home delivery orders.

Chapter 5
Configuring the online order fulfilment networks for omni-channel retailers

Different retailers have adopted different combinations of online fulfilment options to strike a balance between delivery services and costs. No guidelines are available for retailers to understand which fulfilment options are likely to be successful in which retail settings. We first provide a typology of commonly used fulfilment network configurations for groceries and non-food products. Our typology shows that the role of conventional retail stores has evolved in the omni-channel world to become either BOPS collection points or online fulfilment facilities or both. We focus our study on the configuration of online order fulfilment network for groceries. We evaluate different online fulfilment network configurations by developing the corresponding total cost functions analytically. We derive total cost functions for the fulfilment network when different centralisation levels are used for order picking, i.e., different numbers of retail facilities are utilised for picking online orders. Computational experiments have been conducted to provide clear managerial insights on the number of online fulfilment facilities needed for different business scenarios.

5.1 Introduction

Convenient cross-channel services offered to customers such as BOPS and SFS have put retailers under huge pressure to deliver flexibly. While flexible delivery and collection services have boosted online sales, the cost to fulfil online orders from picking to shipping has also increased dramatically. It is estimated that 18 cents of every dollar generated online goes to the costs of order fulfilment (IBM, 2017). The global retail and consumer goods CEO survey conducted by PwC on behalf of JDA Software revealed that only 19% of the top retailers report profitability from fulfilling their omni-channel demand (JDA, 2015). From a service perspective, order fulfilment performance (e.g., timeline and charges) has been linked to customers' perception on service quality and therefore customers' behaviours such as repurchases, referral and returns (Rao et al., 2011; Nguyen et al., 2018). The omni-channel mandate makes it imperative for retailers to reconfigure their online fulfilment networks to maximise operational efficiency while ensuring customer satisfaction.

Typically, there are four fulfilment options available, namely fulfil from retail stores, from DCs, from dedicated EFCs, from suppliers' warehouses (i.e., drop-shipping) (Ishfaq and Bajwa, 2019). The Drop-shipping option has become more prevalent as suppliers can reach more customers and retailers can provide a wider range of assortments with low inventory risks and less investment in online fulfilment capabilities (Khouja and Stylianou, 2009;

Randall et al., 2006). In the UK, retailers across different sectors have deployed different combinations of these fulfilment options. In the clothing sector, UK-based fashion retailer Next leverages its stores for BOPS order fulfilment and its warehouses for direct HD orders. In 2018, the company also started using available store inventories for HD orders when stock-outs occur in warehouses, as well as transshipment between stores when a requested collection store cannot fulfil a BOPS order (Next, 2019). A rival retailer, New Look, only fulfils online orders from its DC, which also handles traditional store replenishment (New Look, 2019). In the electrical goods sector, the UK general catalogue merchandiser Argos uses a combination of hub stores, DCs and regional online fulfilment centers for online demand (Sainsbury's, 2017). Currys PC World, a major player in British electrical and white goods market, fulfils online orders from DCs and some of its stores (Dixons Carphone, 2019).

With a host of possible combinations of fulfilment options, it is crucial for retailers to have clear guidance to understand which option is likely to be preferable in which setting. The literature review in Chapter 2 has discussed existing studies with regard to the online order allocation problem. However, little work has been done on evaluating the full range of existing fulfilment options in different retail scenarios with different product types (e.g., perishable or bulky) and retail profiles (e.g., the density of store network). In fact, there is a lack of classification of online fulfilment network configurations based on various business factors such as product characteristics. Additionally, the common assumption shared among existing studies is that all stores are utilised for online fulfilment, which is rarely the case in practice and may not be optimal. Lastly, extant studies mainly deal with the online fulfilment problem for HD demand and often overlook the ever-increasing BOPS demand. This could be the result of the dominating status of HD services, though BOPS services have been gaining increasing popularity worldwide (GlobalData, 2018b).

As highlighted in sections 2.8 and 2.9, research focusing on network configuration for omni-channel retailing is very limited. To bridge the gap, we first propose a typology of commonly used online order fulfilment network configurations currently used by retailers with multiple sales channels. We then focus our study on the configuration of online fulfilment network for groceries. The typical online groceries fulfilment networks adopted by leading British supermarkets consist of a set of retail stores and dedicated EFC(s). Since our objective is to cast light on the strategic decision on from where retailers should fulfil their online orders, two questions are particularly critical. Retailers need to decide whether and how many EFCs they need to invest and put into operation. Additionally, retailers need to contemplate the number and location of retail stores to be utilised as picking facilities for online fulfilment.

The higher the number of stores used for online fulfilment, the more decentralised the fulfilment operations is. We develop analytical models to evaluate the total costs of different online fulfilment configurations under different centralisation levels of order picking. The most distinct merit of our analytical models is the capability to provide clear managerial insights with approximate data (e.g., demand density rather than exact customers' locations and demand volumes). We obtain results on the optimal centralisation level of order picking and therefore the number of picking facilities needed but not the exact locations of these facilities. We then investigate the impacts of different factors, such as the demand density, the ratio of BOPS demand to the total online demand, the delivery vehicle capacity and the maximum travel distance allowable for BOPS collection, on the attractiveness of different centralisation levels.

Our study in this chapter contributes to both theory and practice in the rapidly evolving context of omni-channel retailing. Retailers have been revisiting and revising their online fulfilment networks in their transition to omni-channel to deliver such services profitably. This study is the first to provide a typology of commonly used omni-channel fulfilment network configurations, which contributes to research in the omni-channel domain. The new analytical models presented provide insights for both academics and practitioners on the fulfilment choices that arise in the dynamic retail landscape. The model results and the analyses in this study can help guide retailers to make strategic decisions on how best to fulfil online orders under different business scenarios.

5.2 Relevant literature

Two streams of literature are closely related to this study owing to the type of questions examined here – (1) research on physical network design (PND) and (2) research on continuous approximation (CA) techniques, which are used here to obtain the relevant delivery costs.

5.2.1 Physical Network Design (PND)

Physical distribution network (PDN) design aims to shape the structure of the distribution network in accordance with a company's competitive strategy and long-term goals (Varsei and Polyakovskiy, 2017). It is concerned with a variety of decisions such as the strategic decisions on the number, size, and location of facilities (e.g., manufacturing plants, warehouses, DCs, etc.); tactical decisions on distribution, transportation, and inventory management policies; and operational decisions on fulfilling customer demand (Perl and Daskin, 1985; Daganzo and Newell, 1986; Ballou, 2001; Ambrosino and Scutella, 2005; Farahani et al., 2014; Mangiaracina et al., 2015).

The topic of PND has received significant attention since the 1990s with the emergence of global markets for products, which compelled businesses to revamp and reconfigure their distribution networks in order to compete globally (Ashayeri and Rongen, 1997). One of the most researched decisions is the facility location problem (FLP). The basic setting of the FLP is to determine the site (s) where a new facility (facilities) should be located to meet the requirements of clients at minimum cost, given a set of clients and a set of potential sites for facilities. Many variants of the FLP have been studied extensively, such as capacitated or uncapacitated FLP (Akinc and Khumawala, 1997; Ghosh, 2003; Daneshzand and Shoeleh, 2009; Sun, 2012; de Armas et al., 2017; Chauhan et al., 2019), single- or multi-product FLP (Owen and Daskin, 1998; Revelle et al., 2008; Nezhad et al., 2013; Askin et al., 2014; Montoya et al., 2016), and static or dynamic FLP (Van Roy and Erlenkotter, 1982; Current et al., 1998; Arabani and Farahani, 2012; Jena et al., 2015; Wang et al., 2017).

The problem of facility location is often considered in conjunction with other decisions, such as demand allocation, inventory level, and transport routing, to optimise the integrated network design. The simultaneous decisions on the number and location of facilities and the assignment of customer demand to these facilities is known as the location-allocation problem (LAP) (see Brimberg et al. (2008) for a detailed review). The simultaneous decisions on the number and location of facilities and the inventory management (e.g., inventory service level) at each facility is known as the location-inventory problem (LIP) (see Farahani et al. (2015) for a detailed review). The simultaneous decisions on the location of facilities and the vehicle routing problem is known as the location-routing problem (LRP) (see Drexl and Schneider (2015) for a detailed review).

The FLP, LAP, LIP and LRP are often formulated with mathematical programming models (e.g., mixed-integer programming) and solved with optimal algorithms or (meta-) heuristics. As Chopra and Meindl (2016) have pointed out, both fixed facility costs and inventory costs increase while transportation costs decrease as the number of facilities increases. Therefore, the total cost curve is convex with respect to the number of facilities. This explains the mechanism of deriving the optimal solutions for those mathematical programming models that have a single objective function (cost minimisation or profit maximisation). Depending on specific research settings (e.g., assumptions on the types and nature of costs considered), these models can inform business decisions based on the trade-offs between different types of costs (e.g., fixed facility costs, transportation costs, and inventory costs).

A potential shortcoming of these models with single objective function is the neglect of possible conflicts between different objectives, which often exist in practice (Mangiaracina

et al., 2015). To this end, a growing body of research has proposed multiple criteria optimisation models to solve problems with respect to PDN. For example, Sabri et al. (2000) propose an integrated multi-objective supply chain model taking into account costs, customer service levels (fill rates), and flexibility (volume or delivery). Melachrinoudis et al. (2005) examine the warehouse location problem based on multiple criteria including cost, customer service, and intangible benefits. Liu and Papageorgiou (2013) address the production, distribution, and capacity planning of global supply chains with total cost, total flow time, and total lost sales as key objectives simultaneously. The objectives of profit maximisation, sales maximisation and supply chain risk minimisation are considered in Bilir et al. (2017) for a joint supply chain network optimisation and competitive facility location problem. Apart from economic and societal considerations, the incorporation of environmental considerations into PND has become a popular trend in recent years. For example, Sherafati et al. (2019) propose a sustainable supply chain network design model considering economic, environmental and social impacts. Existing research has examined various environmental aspects, such as the greenhouse gas (GHG) or carbon footprint (Pishgar-Komleh et al, 2017; de Figueiredo et al., 2017; Liu et al., 2018a; 2018b), water footprint (Mohammed et al., 2017; Zahiri et al., 2017), and waste reduction (Arampantzi and Minis, 2017). Readers are referred to Waltho et al. (2018) for a detailed review in this field.

As discussed in Chapter 2 (e.g., Ishfaq and Raja (2019)), one distinctive feature of PDN design for an omni-channel retailer is that demand can be satisfied flexibly from any available facility in the retailer's network. However, optimal network design for omni-channel retailers has had only limited consideration in the literature to date. Studies focusing on the practice of fulfilling online orders from one retail store before shipping them to another for customer collection is even scarce, even though such practice is common in today's retail world (Next, 2019; Supplychaindive, 2019). In other words, the design of online order fulfilment network while considering the changing roles of conventional retail stores (sinks for offline and BOPS orders and/or sources for HD and BOPS orders for the flow network) is pertinent and timely but is understudied. We reduce this deficit by examining the configuration of online order fulfilment network for grocery retailers and allowing for the shipment of BOPS orders from one retail store to another.

5.2.2 Continuous approximation (CA) technique

In this study, we use a CA technique to obtain the vehicle routing cost for delivering HD orders directly to customers and shipping BOPS orders to collection stores. CA techniques have been widely used to provide insights into Operational Research problems such as facility location, vehicle routing, inventory management and integrated supply chain and

logistics decisions (Dasci and Verter, 2001; Ansari et al., 2018). The key idea of CA is to use a continuous representation of input data and decision variables as density functions and approximate the objective into a functional (e.g., integration) of localised functions which are amenable to solution by standard calculus methods (Ansari et al., 2018). As a result, CA needs less accurate data (e.g., average demand density instead of exact demand locations and volumes) and often provides closed or near-closed form solutions to reveal managerial insights on large-scale Operational Research problems.

The CA approach was first proposed by Newell (1971) and applied to various finite-dimensional Operational Research problems such as scheduling, location and transportation in Newell (1973). The extensive application of CA in vehicle routing problems (VRP) is to decide the most cost-efficient routes for a fleet of vehicles to delivery or pickup commodities or people across a continuous space (Daganzo, 1984a; 1984b; 2005). These studies have provided good approximations for routing distance and therefore costs that are well accepted in the research and practice communities.

For VRP, a good approximation for the optimal value of the total expected all-vehicle travel distance is $2\bar{l}(d/Q)N + k_{VRP}\sqrt{NA}$, where A is the size of a bounded service area; N is the number customers randomly distributed in the service area; d is the expected value of demand for each customer; $\bar{l}$ is the average distance from the depot to a customer point in the area; Q is the capacity of the homogenous delivery vehicles (i.e., each vehicle can serve maximum Q unit demand), and k_{VRP} is a constant scalar determined by the distance metric (i.e., Manhattan distance or Euclidean distance) and vehicle capacity (Campbell, 1993). Campbell (1993) suggests the following values for k under different Q: $k = 0.73$ for $Q = 2$; $k = 0.68$ for $Q = 3$; $k = 0.63$ for $Q = 4$; $k = 0.6$ for $Q = 5$; $k = 0.57$ for $Q \geq 6$. As a result, some researchers just use $k = 0.57$ for the VRP travel distance approximation and we follow the same spirit in this study. The evaluation of $\bar{l}$ can be expressed straightforwardly as an integral and its exact value depends on the shape of the area, the distribution of customers, the distance metric and the location of the depot (Stone, 1991; Pang and Muyldermans, 2013). Stone (1991) has provided detailed analyses and results for evaluating $\bar{l}$ under different scenarios, i.e., different shapes of the region (e.g., rectangle, circle and regular hexagon), different metrics (i.e., Manhattan and Euclidean), and different locations for the depot (e.g., centre or interior or exterior of the region). For example, using the Euclidean metric, the average distance of a depot (centred in the region) to all points is $2R/3$ for a circle with a radius of R and $(1/3+\ln 3/4)R \approx 0.6079864R$ for a regular hexagon with a radius of R (Stone, 1991).

When there is only a single salesman (instead of a fleet of vehicles) visiting N customers uniformly and independently scattered in a region with a size of A, the problem reduces to the well-known Travelling Salesman Problem (TSP). In this case, the expected tour length can be approximated with $k_{TSP}\sqrt{NA}$, where k_{TSP} is again a constant scalar determined by the distance metric (Beardwood et al., 1959). For example, k_{TSP} is estimated at 0.765 for the Euclidean metric via Monte Carlo experiments in Stein (1978) and at 0.97 for the Manhattan metric in Jaillet (1988).

As we only use the CA approach to derive the delivery cost for online orders, we mainly elaborate the literature on CA applications on VRP problems. The application of CA in other problem classes such as facility location problem to obtain results on the size of the influence area of each facility (and therefore the number but not locations of facilities) and inventory management (e.g., shipment size) are reviewed in details in Ansari et al. (2018). For the application of CA models on freight distribution, readers are referred to Langevin et al. (1996) for a review. The major novelty and contribution of our application of CA technique in this chapter is the incorporation of BOPS demand. Fulfiling BOPS demand from a relatively centralised facility to other collection stores via STS incurs extra travel distance. This complicates the approximation of total delivery costs for a picking facility to serve both HD demand and BOPS demand in its influence area and changes the equlibrium of the trade-offs between different types of costs.

5.3 A typology of online fulfilment networks

The literature on the online order fulfilment problem provides insights into how to route online orders to different locations (e.g., retail stores and DCs) without discriminating the profiles of products and retailers in different retail categories. In practice, retailers from different sectors use different combinations of fulfilment facilities to serve online customers. Therefore, it is necessary to provide a typology of online fulfilment network configurations, which we pursue as follows.

5.3.1 Grocery network

Grocery retailers such as Tesco were one of the trailblazers to use retail stores for online order picking. Tesco started with a complete store-based fulfilment strategy until 2006 when it launched its first Dotcom fulfilment centre in Croydon, London. This is followed by the operation of a series of Dotcom fulfilment centres, including Aylesford (Kent) in 2008; Greenford (Greater London) in 2010; Enfield (Greater London) in 2012; Crawley (Sussex) in 2013 and Erith (Greater London) in 2013 (Wulfraat, 2014). The online order fulfilment network in Tesco today consists of approximately 291 stores and five purposely built Dotcom

fulfilment centres. In contrast to Tesco, Sainsbury's initially invested in a dedicated EFC for their online groceries orders. The original EFC was closed in 2004 and the company has resorted to stores to fulfil all online orders (Ellis-Chadwick et al., 2007). As the store fulfilment capacity became more constraining with the growth of online volume, Sainsbury's has operated an EFC in Bromley-by-Bow (London) in 2016. Waitrose represents 4% of the total online grocery market in the UK, a distant 3rd place after Tesco (30.1%) and Sainsbury's (13.7%) (Mintel, 2019). It has collaborated with Ocado, a pure online supermarket in the UK, who delivers approximately 4,500 Waitrose goods. However, the deal with Ocado is due to end in September 2020 and Waitrose has been investing heavily in its online business, including the plan for three state-of-the-art customer fulfilment centres (Waitrose, 2019). Like its competitors, Waitrose has been processing online orders from local stores with the addition of one dedicated EFC in Coulsdon (London) since 2015. Its second EFC in Enfield is set to open in 2020 (Waitrose, 2019).

The online fulfilment networks used by the three leading retailers appear to be converging to using a combination of retail stores and dedicated EFCs. The EFCs in operation for these three retailers are located in or close to regions with high online demand density, i.e., the Greater London. Groceries are highly perishable and customers' tolerance to long delivery waiting time for groceries is low (Digitalcommerce360, 2017). This makes store-based picking more attractive as the delivery routes starting from stores are much shorter. However, the number of items per grocery order is often very high and stores are infamous for having low picking efficiencies (Hübner et al., 2016b). Additionally, the growth of online groceries sales render store picking capacities insufficient. These factors lead to the utilisation of EFC(s). The hybrid solution – a combination of retail stores and EFC(s) – allows a grocery retailer to exploit benefits and eliminate drawbacks for both options.

Figure 29 depicts the typical fulfilment network of traditional grocery retailers to meet demand for online groceries. We stress that our focus is to understand how online orders are fulfilled. Hence, we do not consider the traditional distribution network for walk-in customers. Stores in the network can be categorised as collection stores that just serve as BOPS collection points without online fulfilment capabilities and picking stores that serve as both collection points and picking facilities for online fulfilment. Customers can select a collection store or a picking store to collect their BOPS order. Figure 29 illustrates the use of both picking stores (Element ①) and EFCs (Element ②) to process online orders and deliver directly to HD customers or ship BOPS orders to collection stores. Such fulfilment networks display the pooling of inventories for online and offline channels at the store level. Traditional grocery retailers use EFCs mainly in regions with high online demand volumes

so that the formidable investment in such facilities can be justified. Some small-scale retailers or retailers at the beginning of their online business expansion rely on store-based fulfilment only without investing in an EFC (only element ① in the network). For example, Co-op food initially launched its same day delivery service only in March 2019 for shoppers within a four kilometres radius of several stores in London (Co-op, 2019) and expects to roll out its online same day delivery services across almost 100 towns and cities in the UK with orders fulfilled from 650 Co-op stores (ITV News, 2020).

Figure 29 generalises the typical online fulfilment configurations among retailers with relatively mature online business that include both elements ① and ② and retailers using only element ① that is more common in the initial stage when they are developing their online businesses.

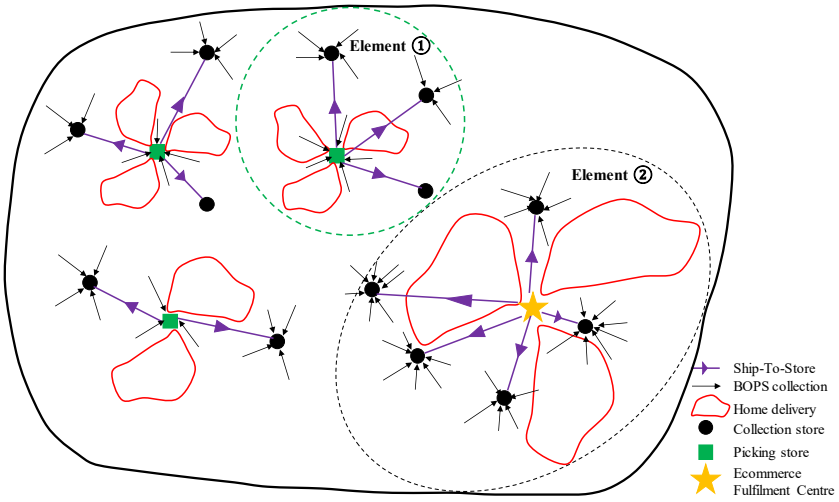


Figure 29. The online order fulfilment network for groceries.

5.3.2 Non-food fulfilment network 1

For non-food products, retailers exhibit different configurations of online fulfilment networks. The first type is the utilisation of integrated DCs for both traditional store replenishment and online order fulfilment (see Figure 30). Stores are only used as BOPS order collection points without online fulfilment capabilities. DCs pick and pack online orders and then either deliver directly to customers (HD) or ship to store for BOPS customer collection. This is essentially the same fulfilment structure as the EFCs for groceries (element ② in Figure 29), when traditional store replenishment is not considered. Both DCs and EFCs have higher fulfilment capacity than stores and do not provide on-site BOPS collection services. Many well-known British retailers, such as John Lewis, New Look, Boots and WH Smith use such fulfilment networks for most, if not all, of their products. In John Lewis, inventories for different product types held at different DCs are viewed as a single stock pool

for all fulfilment channels (store replenishment and online fulfilment) (eDelivery, 2015). For example, its 2nd DC in Magna Park mainly handles garments and 'binnable' items such as shoes and accessories. The DC delivers approximately 60% online orders either directly to customers or to stores for click-and-collect, as well as replenishing about 3 million units of stock to stores each week (Warehouse & Logistics News, 2017). For New Look, its directly operated and highly automated DC in Stoke-on-Trent (Staffordshire) handles deliveries for the UK, Ireland and mainland Europe plus global fulfilment of all ecommerce and 3rd party ecommerce orders, completed by an outsourced distribution hub in Singapore (New Look, 2019). We have explained the fulfilment network configuration of Boots and WH Smith in Chapter 4. Boots relies on its large scale and highly automated Customer Service Centre (CSC) in Nottingham and WH Smith operates three DCs (Birmingham, Swindon and Dunstable) for online fulfilment, whilst all the facilities handle store replenishment as well (NottinghamshireLive, 2017; Menzies Distribution, 2019). We illustrate this type of online fulfilment network in Figure 30, which manifests inventory pooling across channels at the DC level.

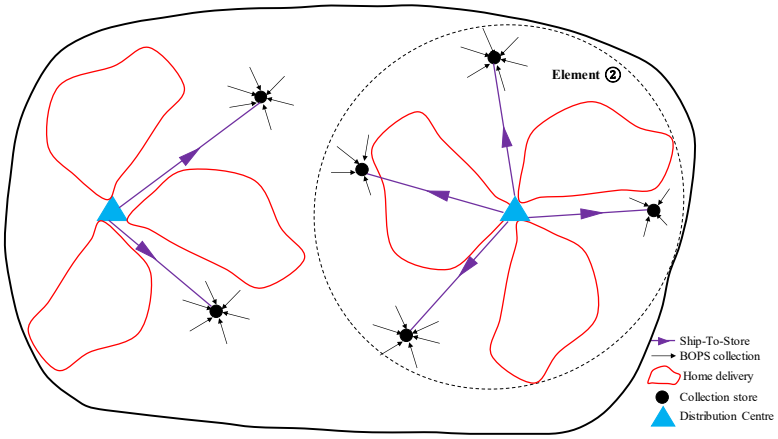


Figure 30. The online fulfilment network configuration 1 for non-food.

5.3.3 Non-food fulfilment network 2

So far, we have delineated two fulfilment networks in which inventories for different sales channels are pooled at the store or DC level. As the channel integration level increases, inventories in more locations can be viewed as a single stock to fulfil online orders even more flexibly. This leads to a more complex fulfilment network as indicated in Figure 31. The key characteristics of such online fulfilment networks is the co-existence of picking stores and DCs for online order processing, direct delivery for HD demand and STS for BOPS demand. This indicates inventory integration for online and offline channels at both the store and the DC level. When a picking store has no product requested in some BOPS orders, such product may be fulfilled and shipped from a DC to the picking store for customer collection (Element ③ in Figure 31). Depending on the business strategy, retailers may

boost their fulfilment capabilities by investing in EFCs. This can help retailers provide faster and more flexible delivery/collection services, particularly same day delivery/collection options. Retailers such as Argos, Next and Currys PC World have been using this type of network for online orders for relatively small items (i.e., non-bulky items such as furniture). 95% of the UK population can enjoy same day delivery, seven days a week offered by Argos. Its Fast Track same day delivery service offers a cut-off time as late as 6pm for delivery by 10pm and Fast Track collection ready in as little as 60 seconds after order placement (argos.co.uk). The outstanding delivery/collection services offered by Argos is attributed to its fine-tuned distribution network consisting of three regional fulfilment centres (Birmingham, Reading and Croydon) (element ②), seven DCs (element ②) and approximate 173 hub stores (element ①) (Sainsburys, 2017; SupplyChainBrain, 2017). Next operates its business with a network of 507 stores and 8 warehouses (Next, 2019). The retailer fulfils online orders from both stores and warehouses without investing in EFCs (Next, 2019). Store stocks are used mainly for local BOPS demand (with the 'Shop My Local Store with 1 hour pickup' option) and for HD demand if warehouses stock out (Next, 2019). Its warehouse in Pontefract, for example, stores and processes new in-season items for both retail stores and online orders (Next.co.uk). Currys PC World also deploys a series of retail stores together with its national DC in Newark, Nottinghamshire, for online fulfilment. 242 out of its 313 stores across the UK offer same day delivery to customers who live within a 10-mile radius of these stores for small items (Dixons Carphone, 2017). Its Newark DC with a size of 1.5 million square feet handles multiple business operations, including central logistics and warehousing for traditional forward distribution, HD for online business, white goods repair, contact centre, repair centre and returns (dixonscarphone.com). As portrayed in Figure 31, this type of fulfilment network exhibits inventory pooling for various channels at both the store and DC level.

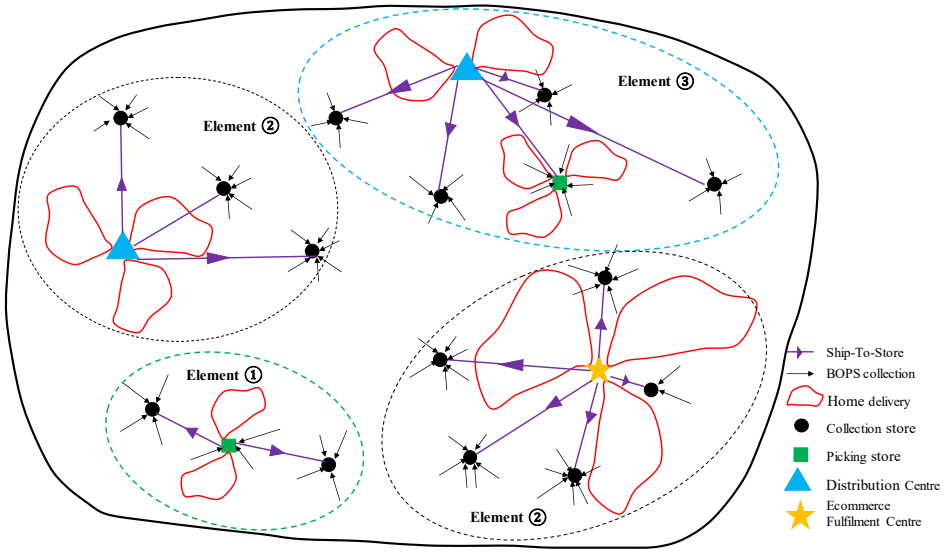


Figure 31. The online fulfilment network configuration 2 for non-food.

5.3.3 Non-food fulfilment network 3

By decoupling customer locations and inventory locations, online platforms can be viewed as an 'endless aisle' with a much wider range of product assortments (e.g., sizes, colours, and designs) that are available only online (Ishfaq et al., 2016). Retailers can list these ecommerce-only items without holding any stocks. Online demand for these items are therefore routed to suppliers who ship directly to customers, i.e., drop-shipping. The drop-shipping option is often applicable to bulky items such as furniture, home furnishing and large home appliances, for which inventories tend to be centralised at the upper-echelon facilities (e.g., central warehouses) due to factors such as lower demand level and high inventory holding costs compared to the transportation cost (Lovell et al., 2005). Theoretically, orders could be drop-shipped directly to customers' homes (HD) or to a retail store for BOPS collection (Ishfaq et al., 2016). In practice, retailers tend to label the items 'Home delivery only' if items have to be drop-shipped by suppliers. This is the case in Tu Clothing and in John Lewis for furniture (johnlewis.com; tuclothing.sainsburys.co.uk). Figure 32 describes this type of online fulfilment network, in which HD orders are fulfilled from retailers' own DC facilities (element ⑤) or from suppliers via drop-shipping (element ④). Note that DCs do not perform STS in element ⑤, unlike element ② in earlier configurations. The key feature of such a fulfilment network is the additional supplement provided by suppliers' distribution facilities to retailers' self-owned stocking facilities. In John Lewis, online orders for furniture, large electrical products and home furnishings are processed in its 3rd DC in Magna Park (with a size of 638,000 square feet) if in stock and are routed to suppliers' warehouses otherwise (eDelivery, 2015). In Tu Clothing, online orders are processed in its Bradford DC if in stock and drop-shipped by brand partners otherwise.

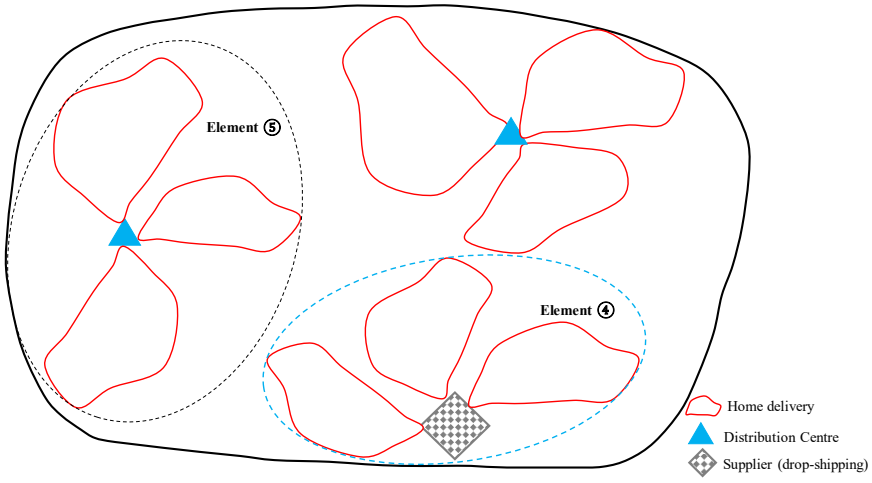


Figure 32. The online fulfilment network configuration 3 for non-food.

To the best of our knowledge, this study is the first to provide a typology of commonly used online fulfilment network configurations for retailers across different product sectors. We

also highlight the commonalities and differences between these configurations by labelling distinctive fulfilment structure elements (i.e., element ①, ②, ③, ④ and ⑤). Fulfilment structures classified as the same element may involve different types of facilities (e.g., EFCs and DCs) but the arcs for online orders flowing into and out of these facilities are the same, i.e., they play the same role in the online fulfilment network. Compared to element ②, element ③ captures the case when a requested BOPS order cannot be fulfilled from a picking store and therefore STS is needed from a DC to that picking store. Element ④ includes the drop-shipping option. This approach reflects a principal-agent relationship between the retailer and its suppliers. As the main objective of this chapter is to examine the online fulfilment decisions, we do not consider more detailed issues related to element ③ (involving inventory management decisions) and element ④ (involving supplier management with potential gaming). Element ② reduces to element ⑤ when DCs do not handle BOPS orders via STS. In the following sections, we examine the main online fulfilment operations (element ① and ②) to provide insights into the decisions on the number of retail stores and DCs/EFCs to use for different types of fulfilment in different retail settings.

5.4 Models for the online groceries fulfilment network

We focus on providing insights on the strategic decision a retailer faces regarding the number of picking facilities needed to fulfil all online demand in a given service region. We first derive expressions for the total cost under different centralisation levels of order picking when online groceries orders are fulfilled from retail stores (i.e., element ①). The total cost expression allows the derivation of insights into the number of stores needed to fulfil online demand in a given service region. We then discuss how to adjust the original analytical model to investigate the case when EFCs and/or DCs are used for online fulfilment (i.e., element ②). To reiterate, we are concerned with configuring the network for the fulfilment of online orders in this chapter and therefore we do not consider demand for traditional walk-in customers.

5.4.1 Problem description

We consider an online order fulfilment problem for groceries described as follows. A retailer selling groceries via physical stores and online channels wishes to serve a number of homogenous and geographically dispersed online customers in a given service region. The service region, denoted as $\boldsymbol{R}$ with a size of R (square mile), is continuous and we model it using a regular hexagon shape. The travel distance between any two nodes (e.g., between retail facilities or between facilities and customers) in $\boldsymbol{R}$ is computed using the Euclidean

metric, which makes a hexagon tiling more attractive since it is well accepted as the best lattice configuration (Newell, 1973; Cui et al., 2010).

Online customers per period (e.g., a month) are assumed to be evenly spread within the retailer's service region $\boldsymbol{R}$ with a homogeneous customer density ρ per square mile ($\rho > 0$). To provide better customer service, the retailer allows customers to choose between BOPS and HD services for online purchases, and customers can collect their BOPS orders from any retail stores. In this study we consider that each online customer induces a single order and the possibility of requesting BOPS services is θ (hence HD service with a possibility of $1 - \theta$). There is a maximum travel distance, denoted as d_{max} (miles), beyond which customers will not travel for a BOPS service. Other research notes the adverse effect of travel distance for store collection on the usage of BOPS services by introducing a hassle cost, i.e., the longer the travel distance for BOPS collection, the higher the hassle cost, the less desired the BOPS services (Gao and Su, 2017). Clearly, the number of collection stores required to guarantee customer service depends on d_{max}. We define the primary influence area of a collection store as $\boldsymbol{S}$ with a size of S, within which customers will visit the store for BOPS collection. We assume $\boldsymbol{S}$ to be hexagonal in order to facilitate the approximation of travel distance. Following the literature on the p-median problem (Berman and Simchi-Levi, 1990), each collection store should be centred at its primary influence area to minimise the total customer travel distance, assuming homogenous online customers in the area. In this way, the radius and the size of a collection store's primary influence area is d_{max} and $S = \frac{3\sqrt{3}}{2} d_{max}^2$ (square miles), respectively (see Figure 33). Assuming that the retailer will serve all online demand in its service region $\boldsymbol{R}$ and ignoring the boundary effect, the entire region $\boldsymbol{R}$ can be partitioned into R/S hexagons (i.e., R/S collection stores). Clearly R/S could be fractional and the issue of rounding may arise. The rounding issue needs particular care (e.g., compare the cost of rounding up and rounding down) when R/S has a small value, i.e., when the retailer only has several stores. If R/S is a large number, the error from rounding can be negligible. In our research setting, groceries retailers often have a dense network of hundreds if not thousands of collection stores (Mintel, 2019). Therefore, we simply round R/S to the nearest integer in our study.

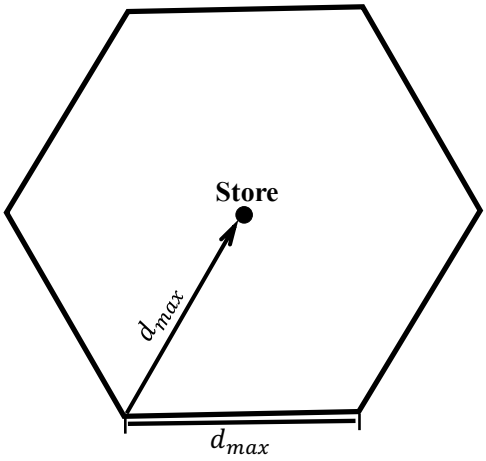


Figure 33. The regular hexagon-shaped primary influence area of a collection store.

Figure 34 displays a hexagonal tessellation of collection stores.

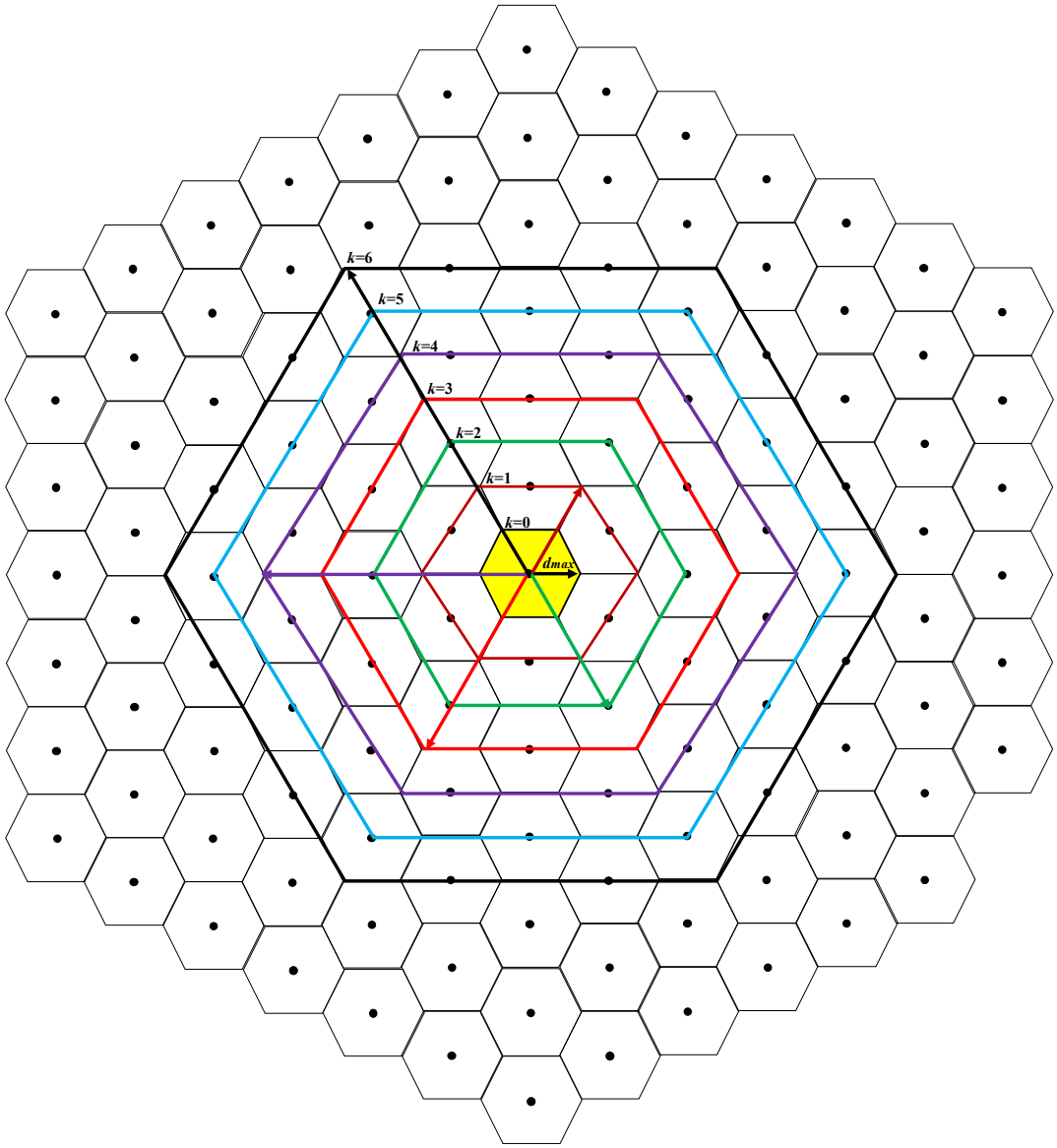


Figure 34. Regular hexagonal tessellation of collection stores in the service region ***R***.

To fulfil online demand, the retailer could deploy each collection store as a picking facility to serve both BOPS and HD customers within its primary influence area. In contrast to such a fully decentralised fulfilment strategy, the retailer could convert some but not all collection stores to picking facilities. In this case, each picking store serves more customers than those located within its primary influence area. We denote the centralisation level of online fulfilment operations as k. The corresponding number of picking stores is denoted as N_k and the influence area of each picking store is denoted as $\boldsymbol{A_k}$ with a size of A_k. Clearly, $N_k = R/A_k$ must hold to serve all online customers in $\boldsymbol{R}$. $k = 0$ indicates that $A_0 = S$, which means a complete decentralisation fulfilment strategy. For $k \geq 1$, each picking store has an influence area with the size of $A_k = (k+1)^2 S$ (including its local primary influence area) since the radius of its influence area increases to $r_k = (k+1)d_{max}$ (see Figure 34).

Figure 35 illustrates the type and the number of online orders handled by each picking store for $k = 0$, 1 and 2, respectively. In the complete decentralised scenario ($k = 0$), each picking store performs HD within its primary influence area and handles its local BOPS customers, which means no STS is required (see Figure 35(a)). When $k = 1$, each picking store receives local BOPS customers, performs STS for 3 collection stores that are $\sqrt{3}d_{max}$ miles away, and delivers to HD customers within its hexagonal influence area with a radius of $2d_{max}$ (see Figure 35(b)). When $k = 2$, each picking store receives local BOPS customers, serves HD customers within its hexagonal influence area with a radius of $3d_{max}$, and performs STS for 8 collection stores, 6 out of which are $\sqrt{3}d_{max}$ miles away and 2 out of which are $3d_{max}$ miles away (see Figure 35(c)).

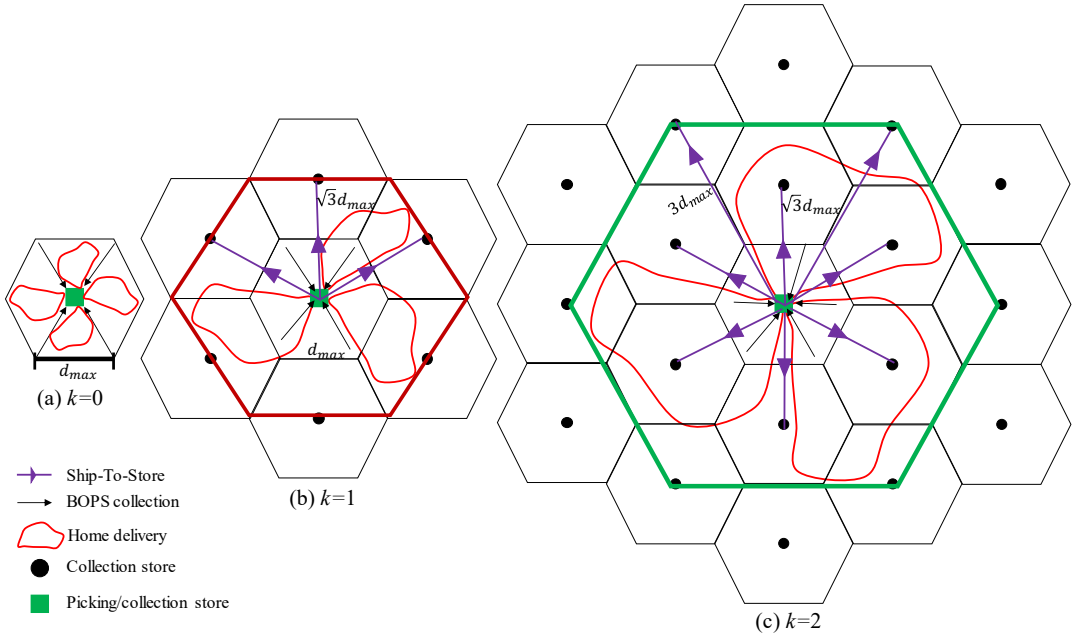


Figure 35. Each picking store handles more online orders as k increases

Table 13 summarises the key features of a picking store for $k \leq 6$ and Table 14 lists the notation we use for this online order fulfilment problem.

Table 13 Key features of each picking facility under different centralisation level k.

Centralisation level k	Radius of a picking store's influence area r_k	Size of a picking store's influence area A_k	#. of collection stores needing STS in $\boldsymbol{A_k}$	Average STS distance from a picking store to a BOPS store $\overline{d_k}$
0	d_{max}	S	0	0
1	$2d_{max}$	$4S$	3	$\frac{1}{3}(3 * \sqrt{3}d_{max}) = 1.732d_{max}$
2	$3d_{max}$	$9S$	8	$\frac{1}{8}(6 * \sqrt{3}d_{max} + 2 * 3d_{max}) = 2.049d_{max}$
3	$4d_{max}$	$16S$	15	$\frac{1}{15}(6 * \sqrt{3}d_{max} + 6 * 3d_{max} + 3 * 2\sqrt{3}d_{max}) = 2.586d_{max}$
4	$5d_{max}$	$25S$	24	$\frac{1}{24}(6 * \sqrt{3}d_{max} + 6 * 3d_{max} + 6 * 2\sqrt{3}d_{max} + 6 * \sqrt{21}d_{max}) = 3.195d_{max}$
5	$6d_{max}$	$36S$	35	$\frac{1}{35}(6 * \sqrt{3}d_{max} + 6 * 3d_{max} + 6 * 2\sqrt{3}d_{max} + 12 * \sqrt{21}d_{max} + 2 * 6d_{max} + 3 * 3\sqrt{3}d_{max}) = 3.764d_{max}$
6	$7d_{max}$	$49S$	48	$\frac{1}{48}(6 * \sqrt{3}d_{max} + 6 * 3d_{max} + 6 * 2\sqrt{3}d_{max} + 12 * \sqrt{21}d_{max} + 6 * 6d_{max} + 6 * 3\sqrt{3}d_{max} + 6 * \sqrt{39}d_{max}) = 4.35d_{max}$
…				
k	$(k+1)d_{max}$	$(k+1)^2S$	$(k+1)^2-1 = k^2+2k$	

Table 14 Notation for the online order fulfilment problem

Notation	Explanation
k	The centralisation level of online picking operations
N_k	The number of picking facilities given the centralisation level k
$\boldsymbol{A_k}$	The influence area of each picking facility given the centralisation level k
A_k	The size of $\boldsymbol{A_k}$
r_k	The radius of $\boldsymbol{A_k}$
$\boldsymbol{R}$	The retailer's entire service region
R	The size of $\boldsymbol{R}$
$\boldsymbol{S}$	The primary influence area of each collection store

S	The size of $\boldsymbol{S}$
d_{max}	The maximum travel distance for customers to collect BOPS orders
ρ	Online customer density per square mile across $\boldsymbol{R}$
θ	The ratio of BOPS demand in total online demand
$\overline{d_k}$	The average STS travel distance from a picking facility to a collection store
α	The picking congestion rate
λ_0	The baseline order volume that can be handled without being affected by congestion in a picking facility
Q	The delivery vehicle capacity, i.e., the maximum number of orders/vehicle
M	The maximum picking capacity for each picking facility
L_{HD}	The total expected all-vehicle travel distance for each picking facility to deliver HD orders
L_{BOPS}	The total expected all-vehicle travel distance for each picking facility to deliver BOPS orders via STS
$\bar{l}$	The average distance from a picking facility to a customer point in its influence area
n	The number of customer points for a picking facility in its influence area
F	The fixed facility cost
a	The reorder cost
h	The unit inventory holding cost
c_p	The unit order picking cost
c_0	The unit order picking cost when the congestion rate is zero
c_f	The fixed cost to conduct a vehicle route
c_v	The unit variable transportation cost for each mile travelled to deliver HD orders
c_l	The unit variable transportation cost for each mile travelled to ship each BOPS order
TF	The total fixed facility cost
TI	The total inventory holding cost
TP	The total order picking cost
$TD1$	The total delivery cost for HD orders
$TD2$	The total delivery cost for BOPS orders via STS
TC	The total cost for fulfilling online demand

For a particular centralisation level k, the number of picking stores needed is $N_k = R/A_k = R/((k+1)^2S)$ and the total number of online orders processed in each picking store is ρA_k, which includes HD orders, local BOPS customer orders, and BOPS orders from other collection stores requesting STS. Since local BOPS customers require no STS, the number of collection stores served by each picking store via STS is $A_k/S–1 = k^2+2k$. We also compute the average STS travel distance, denoted as $\overline{d_k}$, from a picking store to a collection store within its influence area. For example, if $k = 1$, The three collection stores requesting STS are located at the midpoints of three edges of the hexagon and the distance from the picking store to a collection store can be easily computed using the Pythagorean theorem: $\overline{d_1} = \sqrt{(2d_{max})^2 - {d_{max}}^2} = \sqrt{3}d_{max}$(see Figure 35(b)). The calculation of $\overline{d_k}$ becomes more onerous as k increases and there is no specific formula to express $\overline{d_k}$. Table 13 provides the results of $\overline{d_k}$ for $k \leq 6$, which is sufficient to serve our purpose to study the changing pattern of the total cost with varying k.

There is an upper limit on the total number of online orders that can be processed in a picking facility. We denote such maximum capacity as M, which is assumed to be the same for all picking stores. Since we investigate the store-based fulfilment strategy in which all online orders are fulfilled from retail stores, we assume that the total capacity of all picking stores is sufficient to handle all online demand in the service region $\boldsymbol{R}$ (i.e., $M \geq \rho S$). Our primary focus is to provide insights into a retailer's strategic-level decision regarding the number of picking facilities needed to fulfil all online demand in a given service region. The fundamental trade-off regarding this decision is to balance transportation costs and order picking costs (which favour decentralisation) with the fixed facility costs and inventory holding costs (which favour centralisation). If the retailer adopts a more decentralised picking strategy (i.e., smaller k), more retail stores would be used as picking facilities (i.e., higher N_k), which leads to shorter travel distances for delivering HD and BOPS orders. Further, the number of orders allocated to each store will be relatively low and therefore picking facilities will be less congested, incurring lower order picking cost. The adverse effects of this decentralised strategy include the increased inventory holding cost according to the square root law (Maister, 1976) and the higher fixed facility cost required to upgrade and operate each store as a picking facility.

5.4.2 The formulation of the total cost function

Here we present all relevant cost functions for deriving the total online fulfilment cost analytically. In our study, the components of the total cost for the online fulfilment problem are calculated as follows:

1) ***Fixed facility cost***: A fixed facility cost, F (£/picking facility), represents the amortized setup costs for operating a picking facility per period. For example, equipping a collection store with online fulfilment capabilities (i.e., converting a collection store into a picking facility) necessitates initial setup investment, such as technologies (e.g., information system upgrade for enhanced inventory visibilities or touch pads for updating picking lists). The total facility cost TF is given by multiplying the fixed cost of operating each picking facility with the number of the picking facilities; namely,

$$TF(k) = FN_k = FR/(A_k) = FR/((k+1)^2 S)$$

2) ***Inventory holding cost***: We assume that all picking facilities use a continuous review inventory policy and make their ordering decisions for each planning period based on an Economic Ordering Quantity (EOQ) model (Erlebacher and Meller, 2000; Tsao et al., 2012). For each picking facility, the amount of online demand fulfilled is ρA_k and therefore the average inventory holding cost per period is

$\sqrt{2a\rho A_k h}$. The reorder cost, a, is associated with each replenishment and captures costs for communicating with suppliers and transportation partners to organise inventory replenishment. Both the reorder cost a and the unit inventory holding cost h (£ per order per period) are assumed to be the same for all picking stores. The total average inventory holding cost for all picking facilities in the service region is given by:

$$TI(k) = N_k\sqrt{2a\rho A_k h} = R\sqrt{2a\rho h/((k+1)^2 S)}.$$

3) ***Order picking cost***: A unit picking cost, c_p (£/order), is incurred for each order processed in a picking facility. As discussed in Chapter 4, order picking efficiency in stores deteriorates when the amount of orders being processed is high and therefore the facility is crowded with many pickers working concurrently. In the same spirit, we introduce a discount factor α for picking efficiency due to congestion and a baseline order volume λ_0 (the number of orders that can be handled without being adversely affected by congestion in a picking facility). The unit picking cost can therefore be written as $c_p = c_0(1+\alpha)^{\frac{\rho A_k}{\lambda_0}-1}$, while c_0 is the unit picking cost when the congestion rate is zero. Note that the cost function for c_p is reasonable and suitable on the premise that $\rho > 0$. The total order picking cost in each picking facility is $c_p\rho A_k$ and the total order picking cost in the retailer's entire service region is:

$$TP(k) = N_k c_p \rho A_k = \rho R c_0 (1+\alpha)^{\frac{\rho(k+1)^2 S}{\lambda_0}-1}.$$

4) ***HD delivery cost***: For each picking facility, the HD delivery cost is the cost of shipping online orders to the demand points requesting HD services within its influence area ($(1-\theta)\rho A_k$). The transportation cost structure for delivering HD orders consists of a fixed cost and a variable cost (Tsao et al., 2012). The fixed cost, c_f (£/route), includes all the costs needed to conduct a vehicle route, including drivers wages and delivery van depreciations. The variable cost, c_v (£/mile), represents the costs proportional to the travel distance, such as fuel. Therefore, the total HD delivery cost for each picking facility is $c_f\frac{(1-\theta)\rho A_k}{Q} + c_v L_{HD}$. Q is the delivery vehicle capacity (maximum orders/vehicle) and L_{HD} is the total expected all-vehicle travel distance for each picking facility to deliver HD orders. We follow Daganzo (1984a,b) to approximate L_{HD}, which is given by $2\bar{l}\,n/Q + 0.57\sqrt{nA_k}$, with $\bar{l}$ being the average distance from the picking facility to a customer point and n being the

number of customer points (n = (1-θ)ρA_k). The average distance from a picking facility to a customer point in a regular hexagon is given by: 0.608r, with r being the radius of the hexagon (see Stone (1991)). The radius of the regular hexagon enclosing the influence area of each picking store is r_k = (k+1)d_{max} (see Table 13). Hence, $\bar{l} = 0.608(k+1)d_{max}$ and L_{HD} is given by:

$$L_{HD} = \frac{1.216(k+1)d_{max}(1-\theta)\rho A_k}{Q} + 0.57\sqrt{(1-\theta)\rho A_k}.$$

The total HD delivery cost in each picking facility is therefore $c_f \frac{(1-\theta)\rho A_k}{Q} + \frac{1.216c_v(k+1)d_{max}(1-\theta)\rho A_k}{Q} + 0.57c_v\sqrt{(1-\theta)\rho A_k}$ and the total HD delivery cost for all picking facilities in the retailer's entire service region is:

$$TD1(k) = N_k\left(c_f\frac{(1-\theta)\rho A_k}{Q} + \frac{1.216c_v(k+1)d_{max}(1-\theta)\rho A_k}{Q} + 0.57c_v\sqrt{(1-\theta)\rho A_k}\right)$$

$$= R\left(\frac{(1-\theta)\rho\left(c_f+1.216c_v(k+1)d_{max}\right)}{Q} + 0.57c_v\sqrt{(1-\theta)\rho}\right).$$

5) ***STS delivery costs for BOPS***: The STS delivery cost for each picking facility is the cost of shipping BOPS orders to other collection stores (not picking facilities) within its influence area. Retailers such as Next have confirmed that it is generally cheaper to ship online orders to stores for customer collection, compared to direct home deliveries (Next, 2019). This could be explained by the potential cost savings from combining BOPS shipments with traditional store replenishment, as discussed in section 2.6. Therefore, the cost structure for STS should be altered to reflect the difference from HD deliveries. The approximation of STS delivery cost yields different results under different assumptions on BOPS order shipments. Considering the urgency of groceries and the relatively high order volume, the retailer could have direct shipment from a picking facility to a collection store. In this case, the routing of a delivery vehicle to visit more than one customer point is no longer relevant and the transportation cost is mainly dependent on the size of the shipment and the travel distance. Therefore, we introduce a unit variable cost, c_l (£/order-mile), to reflect the associated cost of each mile travelled by a delivery vehicle to ship a BOPS order from a picking store to a collection store. The number of collection stores served by each picking store via STS is k^2+2k and the number of BOPS orders requested in each collection store is $\theta\rho S$. Therefore, the total number of BOPS orders needed for STS from each picking store is $\theta\rho S(k^2+2k)$. The average STS travel distance from a

picking store to a collection store is summarised in Table 13 for $k \leq 6$. In the case of direct shipment for BOPS orders, the total STS delivery costs in the retailer's entire service region can be easily obtained by:

$$TD2(k) = N_k c_l \overline{d_k} \theta \rho S(k^2 + 2k) = R c_l \overline{d_k} \theta \rho (k^2 + 2k)/(k+1)^2 \quad (5.1)$$

When the retailer serves multiple collection stores per delivery trip, it transforms to a VRP problem. The number of delivery vehicles required for STS is $\theta\rho S(k^2+2k)/Q$. Likewise, the total expected all-vehicle travel distance in each picking store for BOPS demand points can be approximated by:

$$L_{BOPS} = \frac{2\bar{l}\theta\rho S(k^2+2k)}{Q} + 0.57\sqrt{A_k(k^2+2k)}$$
$$= \frac{1.216(k+1)d_{max}\theta\rho S(k^2+2k)}{Q} + 0.57\sqrt{A_k(k^2+2k)} \quad (5.2)$$

The approximation for the total STS travel distance L_{BOPS} in Eq. (5.2) can be used to obtain the STS delivery cost following the same approach used in computing the HD delivery costs with relevant parameters. In this study, we assume that groceries retailers make direct shipment to deliver BOPS orders to collection stores. Hence, the STS delivery cost in Eq. (5.1) is used to derive the total cost expression for the online groceries fulfilment problem. Naturally, when all collection stores are used as picking facilities, i.e., $k = 0$, BOPS customers are served from the local store and the retailer incurs zero STS delivery cost.

All relevant cost functions derived above are expressed in terms of the centralisation level of picking (k), which is inversely proportional to the number of picking stores needed to meet all online demand from the service region ($N_k = R/((k+1)^2 S)$). The total cost for the entire service region is the summation of the fixed facility cost, inventory holding cost, order picking cost and delivery cost (for both BOPS and HD demand). The online groceries fulfilment problem to minimise the total cost with a store-based fulfilment strategy can be formulated as follows:

$$\text{Min } TC(k) = R\left(\frac{F}{(k+1)^2 S} + \sqrt{\frac{2a\rho h}{(k+1)^2 S}} + \rho c_0 (1+\alpha)^{\frac{\rho(k+1)^2 S}{\lambda_0}-1} + \right.$$
$$\left. \frac{(1-\theta)\rho(c_f + 1.216 c_v (k+1) d_{max})}{Q} + 0.57 c_v \sqrt{(1-\theta)\rho} + \frac{c_l \overline{d_k} \theta \rho (k^2+2k)}{(k+1)^2}\right) \quad (5.3)$$

Subject to

$$\rho(k+1)^2 S \leq M,$$
$$k \in Z^+.$$

The first constraint describes the limited fulfilment capacity for each picking store and the second constraint specifies that the decision variable k is a non-negative integer. One immediate observation from Eq. (5.3) is that the size of the retailer's entire service region R plays no role in determining the optimal k. That is to say, R exerts no influence upon the strategic decision on the centralisation level of online fulfilment, or the number of collection stores served by each picking store (the ratio between the number of picking stores and the number of collection stores, $N_s/N_k = (k+1)^2$). It only affects the magnitude of total cost savings by choosing the optimal centralisation level k. Based on Eq. (5.3), the sensitivity of the total cost to most model parameters is relatively straightforward. For example, all else being equal, the total cost increases in the size of the region R, the fixed facility cost F, the demand density ρ, the unit order picking cost c_0 and the congestion rate α, whilst it decreases in the delivery vehicle capacity Q. However, the impacts of the maximum travel distance for BOPS collection d_{max} (which defines S) and the BOPS demand ratio θ require further investigation, which will be carried out in the following numerical experiments.

The optimal value of k is the integer within the range $[0, \sqrt{M/\rho S} - 1]$ that minimises $TC(k)$. Note that $TC(k = 0)$ is a constant which depends on the problem parameters in Eq. (5.3). In the following section, we demonstrate how to find the optimal centralisation level k and how it responds to the change in various problem parameters through numerical experiments.

5.4.3 Numerical experiments and results

For ease of comparison, we compute the cost improvement (in percentage) or disimprovement from centralised fulfilment operations over the complete decentralised fulfilment operation, i.e., cost improvement $= 100(TC(k = 0) - TC(k \geq 1))/TC(k = 0)$. Therefore, the minimisation of the total cost $TC(k)$ is equivalent to the maximisation of the corresponding cost improvement, since $TC(k = 0)$ is a constant. A negative cost improvement means that it is best for the retailer to adopt a completely decentralised fulfilment strategy by using all retail stores for online picking. Conversely, a positive cost improvement implies that a relatively centralised fulfilment strategy outperforms the complete decentralised strategy and the retailer should avoid utilising all stores for online picking. The results of numerical experiments in this section illustrate the magnitudes of cost improvement for different centralisation levels, k, when varying different problem parameters. Such sensitivity analyses provide useful insights to inform retailers on the most desirable online fulfilment network configurations for different business scenarios.

We first design a base case scenario for the numerical experiments by specifying values of model parameters as indicated in Table 15.

Table 15 Parameter values in the base case scenario.

Model parameter	Value
F	£100
d_{max}	3 miles
a	£50
h	£0.5/order-period
ρ	10 orders/square mile
c_f	£30
c_v	£0.4/mile
c_l	£0.2/order-mile
c_0	£2/order
α	1%
θ	0.2
Q	20 orders/vehicle
R	2338.27 square mile
M	6,000 orders/period

Given d_{max}, the size of the primary influence area of a collection store is approximately 23.3827 square mile ($S = \frac{3\sqrt{3}}{2} * 3^2 \approx 23.3827$) and there are 100 collection stores (R/S) in the entire service region. The baseline order volume λ_0 is set to be ρS. In the base case scenario, $k \leq 4$ must hold ($\rho(k+1)^2 S \leq M$). When varying ρ or d_{max} (change in S), we specify the maximum k to conform to the capacity constraint. We increase the value of k from 0 to 6 with a step size of 1 and compute the corresponding $TC(k)$ based on Eq. (5.3) and then the resulting cost improvement over $TC(k = 0)$. When investigating the impact of a particular parameter, we keep the values for other parameters the same with that in the base case scenario unless specified otherwise.

Figure 36 illustrates the change in the cost improvement for various values of the maximum customer travel distance d_{max}. When customers are willing to travel further to collect their online orders, a more decentralised fulfilment strategy is favourable. In this example, the optimal $k = 2$ for $d_{max} = 2$ miles ($k \leq 6$ holds, due to M) and $k = 1$ for $d_{max} = 3$ miles ($k \leq 4$ holds due to M). When d_{max} increases to 6 miles ($k \leq 1$ holds, due to M), the retailer ought to remodel all stores as picking facilities to fulfil online demand in order to minimise the total cost. A smaller d_{max} implies a smaller service area covered by each store and therefore less

online demand assigned to each store. With low online demand volume, the savings from the fixed facility cost and inventory holding cost by aggregating online demand into several picking stores (i.e., a more centralised network) overtake the increased transportation cost and order picking cost (congestion is insignificant). A higher centralisation level is therefore more appealing. A bigger d_{max} signifies a larger service area and therefore more online demand for each store. In this case, savings in transportation and order picking costs become more significant and a decentralised network with more picking stores becomes more attractive. We also explore the impact of d_{max} on the total cost, as can be seen in Figure 37. With a given centralisation level k, Figure 37 can inform whether the retailer should encourage or discourage customers to travel further for BOPS collection. In our examples, the retailer can reduce the total cost when d_{max} increases from 2 miles to 3 or 6 miles for $k \leq 1$. However, when the retailer adopts a more centralised fulfilment strategy (e.g., $k \geq 3$), a small d_{max} is actually advantageous to reduce the total cost. Therefore, this model can be used to help retailers decide whether and how far they should use certain incentives such as coupons or price discounts to encourage customers to tolerate a longer travel distance for BOPS collection.

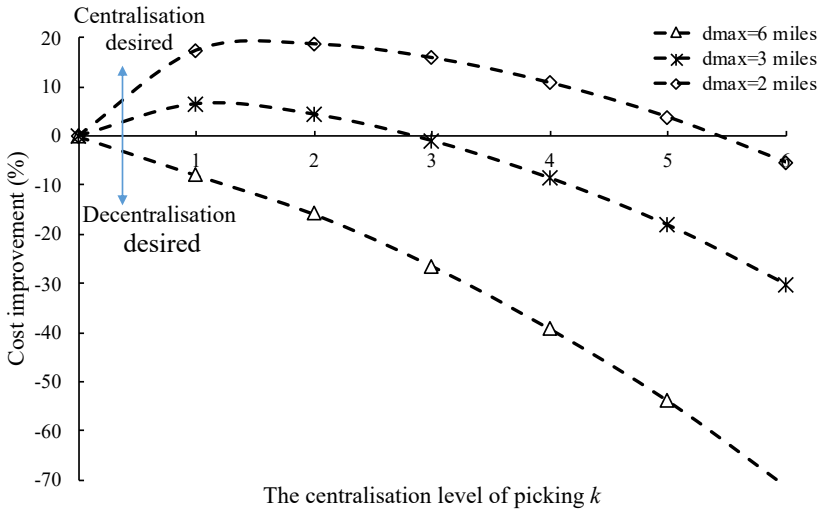


Figure 36. The cost improvement when varying the maximum travel distance d_{max}.

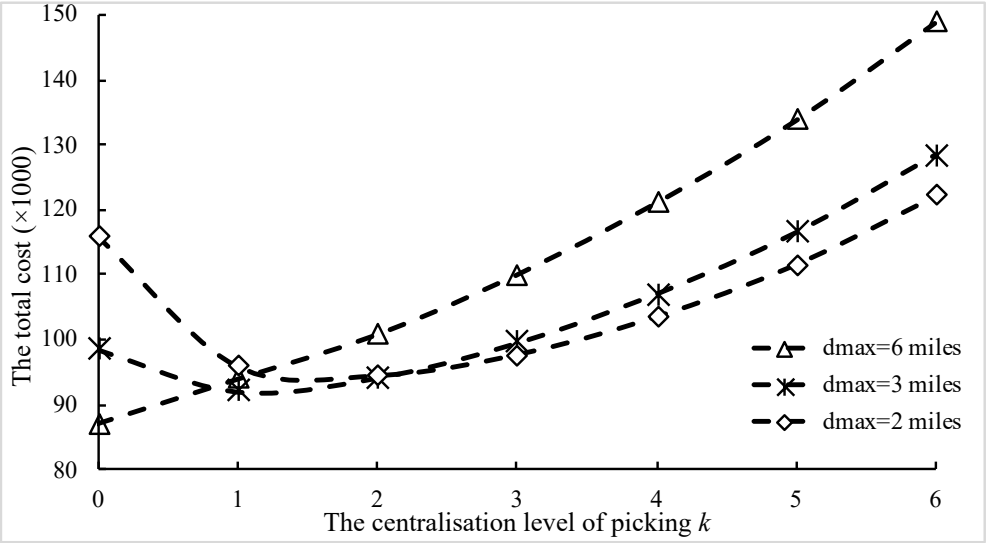


Figure 37. The total cost when varying the maximum travel distance d_{max}.

Figure 38 shows that the lower the demand density ρ, the more centralised the fulfilment network (i.e., higher k), and the fewer the number of picking stores needed for online fulfilment. A corollary is that a complete decentralised strategy can be justified if the demand density is sufficiently high (e.g., $\rho = 40$ orders/square mile). Note that $k \leq 2$ holds for $\rho = 20$ and $k \leq 1$ holds for $\rho = 40$ to meet the capacity constraint. A high demand density means that more online orders are assigned to each store, which lowers the average fixed facility cost and inventory holding cost per order and increases the delivery efficiency from the store to online customers. This makes a decentralised fulfilment strategy more attractive.

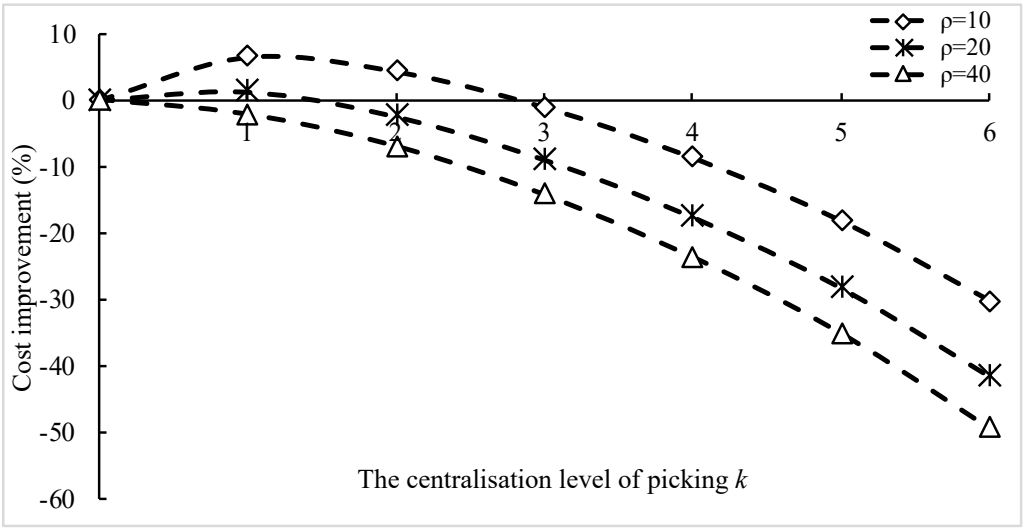


Figure 38. The cost improvement when varying the demand density ρ.

In order to study the impact of BOPS demand ratio to the total online demand on the optimal k, we compute the corresponding cost improvement when $\theta = 0.2$, 0.5 and 0.8, respectively

(see Figure 39). The clear pattern from Figure 39 is that a more decentralised fulfilment network (i.e., lower k) with more picking stores is preferable when more online customers opt for BOPS service (i.e., higher θ). For example, when 20% of customers collect their online orders in stores ($\theta = 0.2$), the optimal $k = 1$ and the corresponding cost improvement is approximately 6.58% over a completely decentralised strategy. For $\theta = 0.8$, the retailer should leverage all stores as picking facilities to minimise the total cost (i.e., optimal $k = 0$). When the centralisation level of online picking is high, more BOPS orders will be picked away from the selected collection store, which results in longer STS transportation distance and increased extra STS delivery costs. Such delivery costs from STS transportation become overwhelming when the BOPS demand ratio is sufficiently high. Hence, the retailer ought to convert more physical stores into picking stores to tame the high STS delivery costs in order to minimise the total cost. Currently, HD service is still dominating in practice. However, BOPS sales are projected to increase steadily (GlobalData, 2018b) due to the win-win benefits of BOPS services for both retailers and customers. Our results indicate that if the trend of the increasing usage for BOPS services continues, retailers will resort to more physical stores within their distribution networks for online fulfilment.

Since the impact of θ on the total cost is not straightforward based on Eq. (5.3), we also show the change in the total cost as θ increases in Figure 40. For a more decentralised fulfilment strategy (i.e., small k), a higher θ generates a lower total cost, which resonates with the common belief that it is cheaper to serve BOPS customers than HD customers (Next, 2019). However, Figure 40 shows that the cost saving benefits from a high usage for BOPS services are diminished or even reversed for a high centralisation level k. This can be explained by the increase in the extra transportation cost for STS as k increases. When more customers choose BOPS over HD, fewer direct deliveries will be requested. Meanwhile, the retailer can curb the extra STS delivery costs for the increasing BOPS orders by deploying more picking stores. As a result, the saving in HD delivery cost becomes more significant than the increase in the STS delivery cost and the overall total cost is lower with a high BOPS demand ratio θ and a low centralisation level k. In our example, for $k = 0$, the retailer saves £22,822 ((98.51 – 75.688)×1000) when the fraction of online customers choosing BOPS increases from 20% to 80%. In this case, the retailer could design certain incentives to steer customers to choose BOPS services over HD services to minimise the total cost.

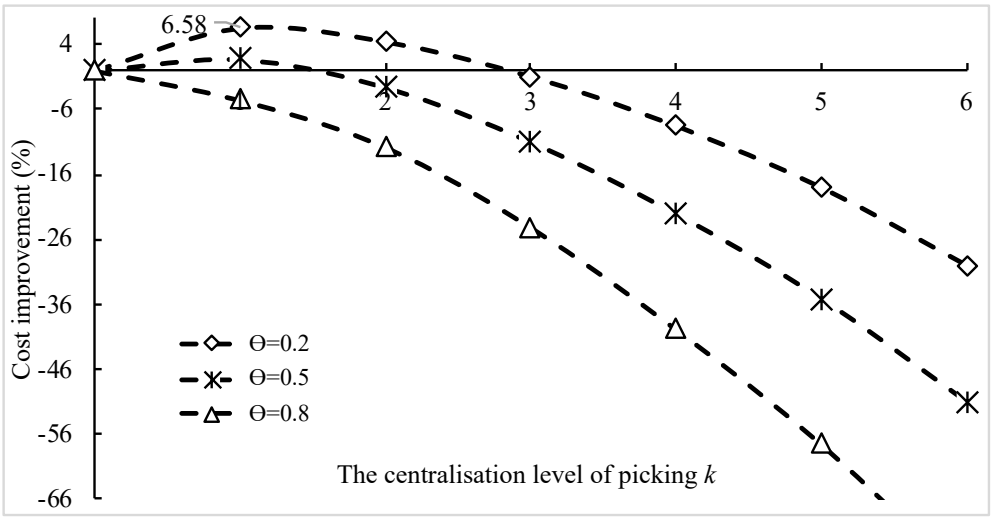


Figure 39. The cost improvement when varying the BOPS demand ratio θ.

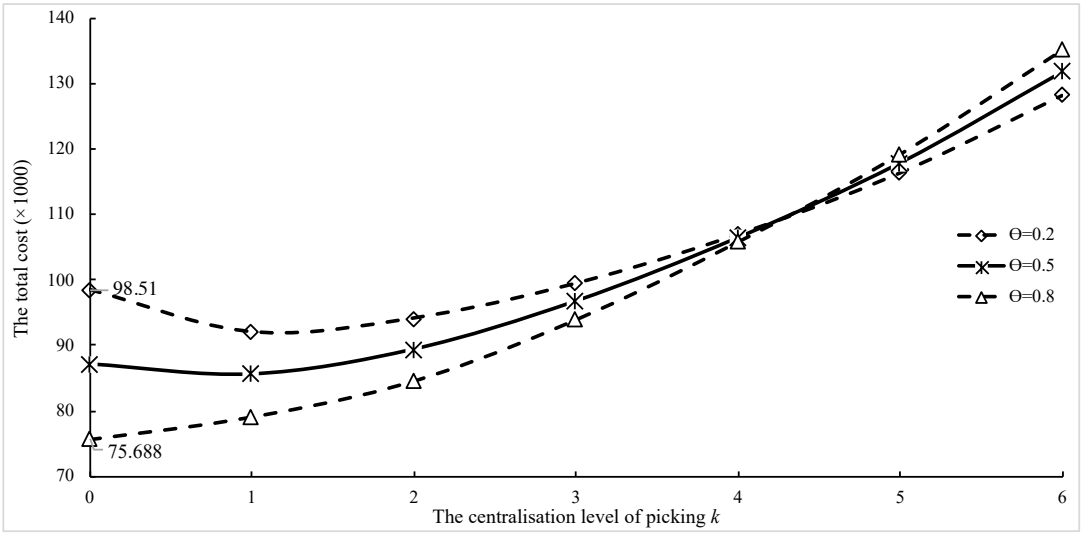


Figure 40. The total cost when varying the BOPS demand ratio θ.

Figure 41 demonstrates that a higher centralisation level of online fulfilment with fewer picking stores is more attractive when the retailer mobilises a fleet of vehicles with high capacity. It is widely acknowledged that delivery efficiency (delivery cost) is positively (negatively) correlated to the delivery density (i.e., the number of stops per vehicle route) (Boyer et al., 2009). The delivery vehicle capacity Q in our problem setting affects the number of customer stops per route and therefore exerts impact on the magnitude of transportation cost savings. Specifically, when the vehicle capacity is high, each delivery vehicle can serve more customers per route and therefore achieve higher delivery efficiency. The increased delivery efficiency makes delivery cost less dominating even if the retailer uses a higher centralisation level for online picking. The reduced dominance of delivery cost implies a more decisive role of the fixed facility and inventory holding cost, which favours a higher centralisation level. For example, when the delivery vehicle capacity Q = 20 orders

per trip, the retailer can achieve the maximum cost improvement (6.58%) by setting $k = 1$. However, for $Q = 40$, the retailer can attain a cost improvement of 6.86% by setting $k = 2$, which is suboptimal (optimal $k = 1$ for a maximum cost improvement of 8.55%) but still outperforms the optimality for $Q = 20$. In other words, the retailer can achieve a higher cost improvement even from a suboptimal k when the delivery vehicle capacity Q increases. Our results suggest that, all things being equal, it is generally beneficial for the retailer to mobilise a fleet of vehicles with higher capacity. The sensitivity of the cost improvement with regard to the increase in Q as illustrated in Figure 41 can help the retailer to balance the cost reduction benefits of mustering larger delivery vehicles against the corresponding up-front capital investment needed.

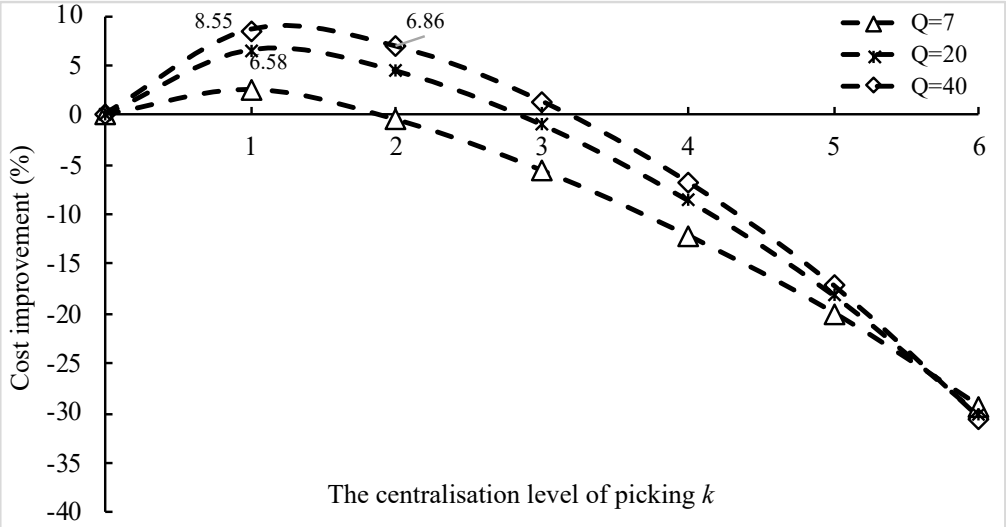


Figure 41. The cost improvement when varying the delivery vehicle capacity Q.

Figure 42 outlines the impact of the unit variable transportation cost c_v on the optimal centralisation level for online fulfilment. The values for c_v in our numerical experiments are guided by the estimation provided by the Department for Transportation in the UK, which indicated that the vehicle cost per mile ranges from approximate £0.35/mile for small vans to £0.7/mile for large vans (DfT, 2013). Figure 42 shows that for a low c_v (e.g., £0.3/mile), the range of the centralisation level leading to a lower total cost than that for the complete decentralisation is relatively wide (e.g., $k \leq 3$). However, this range is narrower (e.g., $k \leq 2$) for a higher c_v (e.g., £0.8/mile). For a higher c_v, the transportation cost becomes more dominant in the total cost expression. In this case, the transportation cost savings from using a more decentralised fulfilment strategy become more considerable than the corresponding increase in the fixed facility and inventory holding cost. Therefore, the retailer will leverage more stores for online fulfilment (i.e., lower k). These results indicate that the retailer could

deploy more stores as online fulfilment facilities when the prices for energy (e.g., fuel and gas) increases.

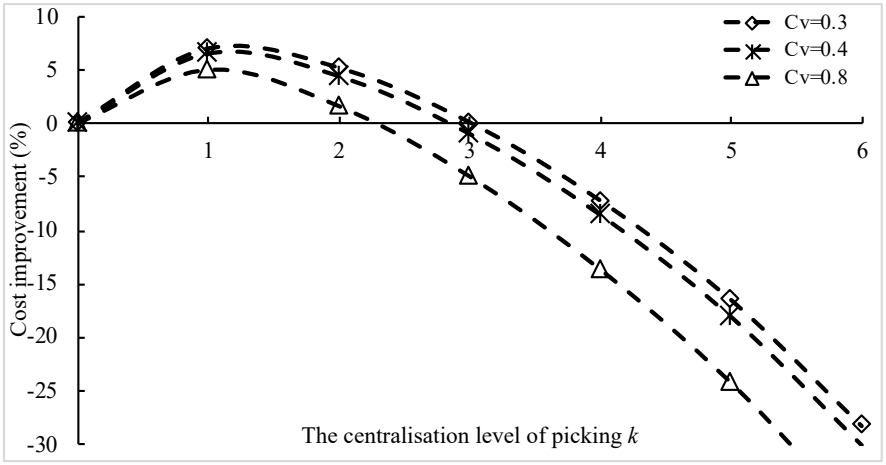


Figure 42. The cost improvement when varying the unit variable transportation cost c_v.

Figure 43 shows the change in the cost improvement when the congestion rate in a picking store increases. The higher the congestion rate, the higher the unit order picking cost, the more decentralised the fulfilment operations should be (e.g., small k). In our example, a 10% congestion rate ($\alpha = 0.1$) will render any centralisation level with $k \geq 1$ cost ineffective. In fact, when setting $k = 3$, the retailer incurs more than twice the total cost of the complete decentralisation (cost improvement $\approx -144\%$ and therefore $TC(k = 3)/TC(k = 0) \approx 2.44$), and the extra cost from a higher centralised level increases more dramatically as k increases.

The congestion rate introduced in this study reflects the diseconomies of scale regarding online order picking. The opposite argument is the economies of scale from high automation and more efficient picking that can be obtained as order volume increases. However, this is not applicable to the case of store picking since automation is very limited. Also, incoming walk-in customers make the problem of congestion more relevant in stores. When picking efficiency deterioration is more considerable due to congestion in the picking store (i.e., higher α), the retailer should deploy more stores for online fulfilment to avoid processing too many orders in a picking store. The consideration of the congestion rate acts as a capacity constraint to limit the number of online orders that can be processed in a picking facility, particularly when the unit picking cost is relatively high.

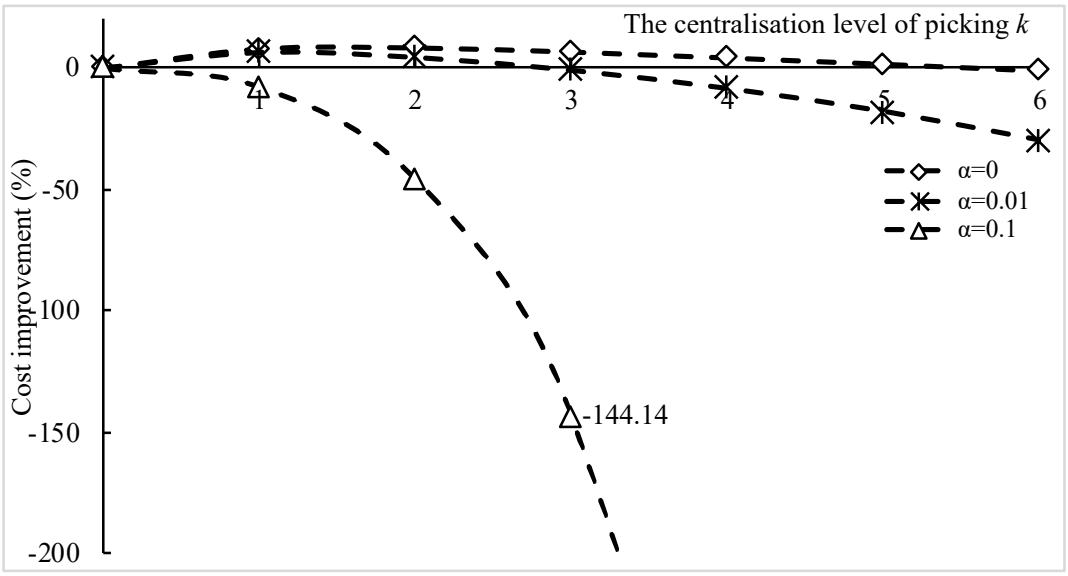


Figure 43. The cost improvement when varying the congestion rate α.

Figure 44 displays how the optimal centralisation level of order picking changes in response to the increase in the fixed facility cost F. As expected, a higher fixed facility cost prefers more centralised online fulfilment operations. Naturally, if it is costly to convert and operate a conventional store as a picking facility for online fulfilment capabilities, retailers will be discouraged from utilising more physical stores to fulfil online demand. This emphasises the key trade-off between the sum of transportation cost and order picking cost and the sum of the fixed facility cost and inventory holding cost when deciding the number of stores to be deployed as picking facilities (i.e., the centralisation level of picking).

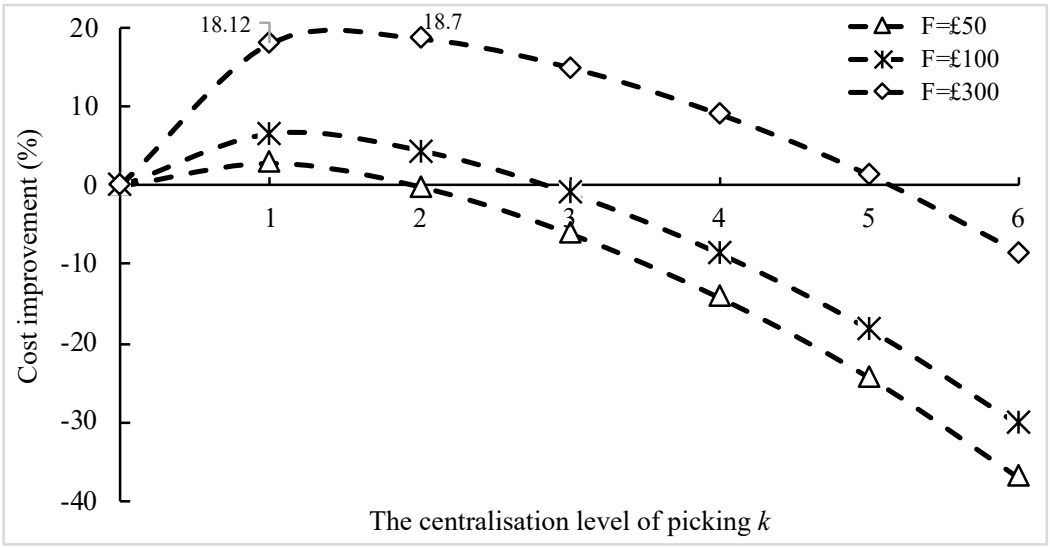


Figure 44. The cost improvement when varying the fixed facility cost F.

So far, the optimal centralisation level k in all experiments complies with the capacity constraint. The impact of the maximum fulfilment capacity M is still obscure. We use Figure 44 to resolve this obscurity. For F = £300, Figure 44 shows that the cost improvement is maximised (18.7%) at k = 2, which abides by the capacity constraint for M = 6,000 orders/period. All things being equal, if M decreases to 2,000 orders/period, $k \leq 1$ must hold and the maximum cost improvement that can be attained is 18.12% at k = 1. In this example, the decrease in M from 6,000 to 2,000 orders/period only results in £680.6/period of cost reduction loss (TC = £96,349.9 for k = 2; TC = £97,030.5 for k = 1; TC = £118,510 for k = 0). If M = 800 orders/period, the retailer has to use all stores for online fulfilment (k = 0). The limiting fulfilment capacity generates TC = £118,510 and the retailer will miss out on considerable total cost reduction from more centralised picking operations (higher k). The change in the minimum total cost TC at the optimal k for a given M can help a retailer explore whether and how much it should expand the fulfilment capacity of its picking stores.

Our model results and analysis provide clear guidance for retailers on the number of stores to be converted into online picking facilities (i.e., optimal N_k) under various business scenarios but does not provide direct information about the exact locations of picking facilities. Recall that each collection store (candidate for picking facilities) is assumed to be centred at its influence area. Therefore, when informed of the optimal N_k, retailers could divide their entire service regions into N_k equal sub-regions and locate the candidate picking store at or close to the geometric centre of each sub-region. In our research settings, online demand is evenly distributed across the region and each picking store has the same influence area with the same number of online orders to be fulfilled. Therefore, fixed facility, inventory holding, and order picking costs will not alter so long as the number of picking stores N_k remains the same, regardless of the exact locations of these picking stores. As a result, the ideal location for a picking store should be the geometric centre of its influence area that minimises the average distance from the picking store to all demand points in its influence area, which minimises relevant delivery costs and therefore the total costs. However, precise location decisions in practice should consider a wider range of practical factors, such as land availability, traffic access, local incentives, costs associated with land acquisition, appraisal, and operations (e.g., labour and utility) (see Onstein et al. (2019) for a comprehensive framework on factors affecting the locations of facilities).

5.4.4 Online fulfilment from DCs or EFCs

As discussed in section 5.3, some retailers fulfil online orders from DCs or EFCs for some or all of their service regions, so that stores are only used as collection points without online fulfilment capabilities. The common practice for groceries retailers is to serve online

customers from dedicated EFCs in regions with sufficiently high demand level (e.g., Greater London). Additionally, we have explained that the roles of DCs and EFCs are essentially the same for online fulfilment and therefore we have categorised them into the same type of fulfilment structure (element ②). In this section, we describe how the derived total cost function in the case of store-based fulfilment can be adjusted to the case of DC-based or EFC-based online fulfilment. The total cost expression can again guide the retailer to decide the number of DCs or EFCs needed to fulfil online demand from the given service region.

In the same spirit as earlier, we denote the service region of the retailer as $\boldsymbol{R_e}$ with a size of R_e. The centralisation level of picking is denoted as k_e, which also defines the size of the influence area for each DC/EFC (denoted as A_e) and the number of DCs/EFCs needed (denoted as N_e). We still assume that each picking facility (e.g., DC or EFC) is centred at its influence area, which has a regular hexagon shape with its radius denoted as r_e. The relationship between k_e, A_e, N_e and r_e remain the same as demonstrated in Table 13, i.e., $A_e = R_e/N_e = (k_e+1)^2 S_e$ and $r_e = (k_e+1)d_{max}$. The only difference is the number of collection stores requesting STS in the influence area of each picking facility. Since DCs and EFCs do not act as BOPS collection points, the number of collection stores within the influence area is $A_e/S = (k_e+1)^2$. All problem parameters, such as the demand density, the fixed facility cost and unit inventory holding cost, are subscripted with e to represent the corresponding features of the DCs/EFCs. The differentiation of parameters enables the account for different features of stores and DCs/EFCs, such as the order picking congestion rate (lower in DCs/EFCs), the fixed facility cost (higher in DCs/EFCs) and the unit inventory holding cost (lower in DCs/EFCs).

The resulting cost expression for the fixed facility cost, the inventory holding cost, the variable picking cost, and the HD delivery cost have the identical structure to the store-based fulfilment strategy described in section 5.4.2. For brevity, we do not elaborate the same calculating logics and directly provide their formulations as follows:

$$TF(k_e) = F_e R_e/((k_e+1)^2 S_e),$$

$$TI(k_e) = R_e\sqrt{2a_e\rho_e h_e/((k_e+1)^2 S_e)},$$

$$TP(k_e) = \rho_e R_e c_{0,e}(1+\alpha_e)^{\frac{\rho_e(k_e+1)^2 S_e}{\lambda_{0,e}}-1},$$

$$TD1(k_e) = R_e\left(\frac{(1-\theta_e)\rho_e\left(c_{f,e}+1.216c_{v,e}(k_e+1)d_{max,e}\right)}{Q_e} + 0.57c_{v,e}\sqrt{(1-\theta_e)\rho_e}\right).$$

The calculation of the STS delivery cost for BOPS orders are different in two aspects. The first one is the number of collection stores served by each DC/EFC via STS, which is $(k_e+1)^2$. The second one is the average STS travel distance from a DC/EFC to a collection store. Compared to a picking store, a DC/EFC typically has much higher fulfilment capacity and delivery vehicle capacity. Therefore, it is very likely that each delivery trip from a DC/EFC can visit more than one collection store. We follow this logic to approximate the STS travel distance in this case. The computation of STS travel distance from one DC/EFC to visit several customer points (i.e., collection stores) follows the same logic as for the derivation of Eq. (5.2). The expected all-vehicle travel distance for each DC/EFC to ship BOPS orders via STS can be approximated by:

$$L_{BOPS,e} = \frac{1.216(k_e+1)^3 d_{max,e}\theta_e\rho_e S_e}{Q_e} + 0.57\sqrt{S_e}(k_e+1)^2.$$

The total STS delivery cost in each DC/EFC is therefore $c_{f,e}\frac{\theta_e\rho_e S_e(k_e+1)^2}{Q_e} + \frac{1.216c_{v,e}(k_e+1)^3 d_{max,e}\theta_e\rho_e S_e}{Q_e} + 0.57c_{v,e}\sqrt{S_e}(k_e+1)^2$ and the total STS delivery cost for all DCs/EFCs in the service region $\boldsymbol{R_e}$ is given by:

$$TD2(k_e) = N_e\left(c_{f,e}\frac{\theta_e\rho_e S_e(k_e+1)^2}{Q_e} + \frac{1.216c_{v,e}(k_e+1)^3 d_{max,e}\theta_e\rho_e S_e}{Q_e} + 0.57c_{v,e}\sqrt{S_e}(k_e+1)^2\right) = R_e\left(\frac{\theta_e\rho_e\left(c_{f,e}+1.216c_{v,e}(k_e+1)d_{max,e}\right)}{Q_e} + 0.57c_{v,e}\sqrt{\frac{1}{S_e}}\right).$$

Note that we use the same fixed cost $c_{f,e}$ and unit transportation cost $c_{v,e}$ for HD and BOPS orders for simple presentation. It is straightforward to specify different values for these parameters. The online fulfilment problem to minimise the total cost with DC-based or EFC-based fulfilment strategy can be formulated as follows:

$$\text{Min } TC(k_e) = R_e\left(\frac{F_e}{(k_e+1)^2 S_e} + \frac{1}{(k_e+1)}\sqrt{\frac{2a_e\rho_e h_e}{S_e}} + \rho_e c_{0,e}(1+\alpha_e)^{\frac{\rho_e(k_e+1)^2 S_e}{\lambda_{0,e}}-1} + \frac{\rho_e\left(c_{f,e}+1.216c_{v,e}(k_e+1)d_{max,e}\right)}{Q_e} + 0.57c_{v,e}\sqrt{(1-\theta_e)\rho_e} + 0.57c_{v,e}\sqrt{\frac{1}{S_e}}\right) \quad (5.4)$$

Subject to

$$\rho(k_e+1)^2 S \le M_e,$$
$$k_e \in Z^+.$$

The implications of the two constraints are the same with that in the case of store-based fulfilment. Eq. (5.4) also indicates that the decision on the optimal k_e to minimise the total cost is independent of R_e, i.e., the size of the service region has no impact on deciding the number of collection stores served by each DC/EFC. We illustrate how to use Eq. (5.4) to decide the optimal centralisation level k_e and therefore the number of fulfilment facilities needed ($N_e = R_e/((k_e+1)^2 S_e)$) with a numerical example.

For the illustrative example, we specify the model parameter values in Table 16. Given $d_{max,e}$, $S_e \approx 23.3827$ square miles and there are 100 collection stores in the region. The baseline order volume $\lambda_{0,e}$ is set to be $\rho_e S_e$ and $k \leq 6$ holds due to M_e.

Table 16 Parameter values for the illustrative example.

Model parameter	Value
F_e	£400
$d_{max,e}$	3 miles
a_e	£30
h_e	£0.3/order-period
ρ_e	50 orders/square mile
$c_{f,e}$	£30
$c_{v,e}$	£0.5/mile
$c_{0,e}$	£2/order
α_e	0.1%
θ_e	0.2
Q_e	50 orders/vehicle
R_e	2338.27 square mile
M_e	60,000 orders per period

We summarise the results of various cost components and the total cost for different k_e in Table 17, and demonstrate the change in the cost improvement for different k_e in Figure 45. The total cost is minimised when $k_e = 2$ and the retailer uses 12 DCs/EFCs ($N_e \approx 12$) to fulfil all HD and BOPS orders in the region with 100 collection stores. In this example, the cost improvement is positive for all $k \leq 6$, i.e., the retailer has a wide range of the choices for the centralisation level k to achieve a lower total cost than that for the complete decentralised order picking ($k = 0$).

Table 17 Results of the numerical example.

The centralisation level k_e	Fixed facility cost $TF(k_e)$	Inventory holding cost $TI(k_e)$	Order picking cost $TP(k_e)$	HD delivery cost $TD1(k_e)$	STS delivery cost $TD2(k_e)$	The total cost $TC(k_e)$	Cost improvement (%)
0	40,000	14,506.69	234,060.69	63,745.17	15,020.43	367,332.98	0
1	10,000	7,253.35	234,763.57	67,157.17	15,873.43	335,047.52	8.79
2	4,444.44	4,835.56	235,939.74	70,569.18	16,726.43	332,515.35	9.48
3	2,500	3,626.67	237,596.28	73,981.18	17,579.43	335,283.56	8.72
4	1,600	2,901.34	239,743.22	77,393.18	18,432.43	340,070.16	7.42
5	1,111.11	2,417.78	242,393.62	80,805.18	19,285.43	346,013.12	5.80
6	816.33	2,072.38	245,563.71	84,217.18	20,138.43	352,808.03	3.95

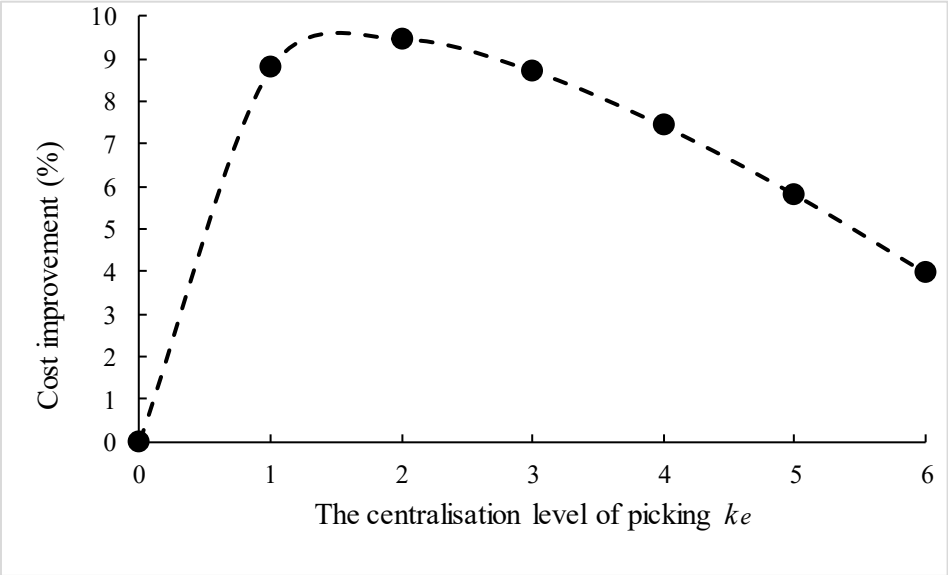


Figure 45. Illustration of the numerical example results.

5.5 Concluding remarks

The observation on the varying combinations of fulfilment options adopted by retailers across different sectors and the lack of guidance on choosing such options have motivated us to study the strategic decision on from where retailers should fulfil their online orders. We proceed with a typology of commonly used online fulfilment network configurations for groceries and non-food products. We then direct our attention to the online groceries fulfilment network, in which online orders are fulfilled from a number of stores and from EFC(s) in regions with sufficiently high demand level (e.g., large metropolitan areas such as Greater London). We first consider the case when the retailer relies on a store-based fulfilment strategy, i.e., leveraging stores to fulfil all online demand in a service region. In this case, we derive the total cost analytically as a function of the centralisation level of order picking. The centralisation level of picking is inversely proportional to the number of stores to be deployed for online fulfilment facilities. Computational experiments have been conducted to investigate how the optimal centralisation level responds to varying cost parameters. We then consider the case when the retailer fulfils online orders from DCs/EFCs and demonstrate how the analytical expression of the total cost of store-based fulfilment can be easily adapted to this case.

Our analyses and results show that the retailer should use more stores as online fulfilment facilities when the demand density is high and/or when customers are willing to travel further to collect their BOPS orders. Within a same customer region, the online demand density in Tesco on average is higher than that in Sainsbury's, since Tesco dominates the online grocery

market in the UK (Mintel, 2019). The approximate number of stores used to pick online orders is 291 in Tesco and 124 in Sainsbury's (Mintel, 2019), which resonates well with our model results and analyses. Another way to interpret the impact of demand density is that stores located in areas with abundant online demand will be more attractive to become a fulfilment facility for online demand. For example, there are 6 Sainsbury's superstores in Nottinghamshire but online orders are only picked from two superstores (Mansfield superstore and Castle Boulevard Superstore) where online order volumes are higher due to their central locations[2]. We assume homogenous online customer demand across a service region but our analyses can be used to gain insights into the choice of picking stores when online demand is unevenly distributed. We highlight that the strategic decision on the number of picking stores needed hinges on the trade-off between the sum of order picking cost and transportation costs for HD and BOPS order shipments and the sum of fixed facility and inventory holding costs. Therefore, a high demand density and/or a further travel distance for BOPS collection implies more online demand is assigned to each store, which makes cost savings in transportation and order picking from decentralised fulfilment more dominant and therefore makes decentralised fulfilment operations more preferable.

A decentralised strategy with more stores utilised as online picking facilities is also desirable in the case of low delivery vehicle capacity, high unit variable transportation cost, low fixed facility cost, high congestion rate in the picking stores, and high BOPS demand ratio to the total online demand. A typical online groceries order comprises many items and therefore requires more picking effort, which makes the order picking efficiency more susceptible to congestion in the store picking facility. This could explain why groceries retailers such as Tesco and Sainsbury's generally adopt a decentralised strategy with many stores utilised for online fulfilment. Additionally, our analyses on the impact of BOPS demand ratio also indicate that if more online customers opt for BOPS services, retailers should resort to deploying more physical stores for online fulfilment. This is consistent with the general tendency in the retail industry worldwide. As BOPS services become more popular, many retailers strive to fulfil more online orders from stores by leveraging more stores as picking facilities. One example is Target that is reconfiguring 300 stores as online picking facilities annually and is on track to reconfigure 1000 stores by the end of 2020 (Supplychaindive, 2019). Other managerial implications include gauging the impact of incentives to encourage customers to travel further for BOPS collection and weighing the investment in mobilizing delivery vehicles with higher capacity against the corresponding savings in the total cost. The expression of the total cost provides further insights. For example, the size of the retailer'

[2] Source: Sainsbury's. *Where and when can I collect my click & collect grocery orders?* Available at: https://help.sainsburys.co.uk/help/delivery-collection/cc-where-when

service region exerts no influence upon the optimal decision on the centralisation level of online fulfilment, as it only affects the magnitude of the cost savings from such decisions.

In this chapter, our analytical model for the total cost requires only limited customer data (e.g., only demand density in a service region), whilst providing many valuable managerial insights. Our model results and analysis can guide retailers to decide the number of picking facilities needed, which also allows retailers to obtain some candidate locations for picking facilities from geometric analysis. However, we acknowledge that this analysis is limited and call for research using more comprehensive optimisation models and/or analytical models to derive further insights into specific optimal locations for online fulfilment facilities, taking account of a range of practical factors discussed in section 5.4.3. Additionally, we identify five distinctive fulfilment structures (element ①, ②, ③, ④, and ⑤ in section 5.3) for online demand. In this chapter we have focused our attention on the first two structures, i.e., fulfil from retail stores (element ①) and fulfil from DCs/EFCs (element ②). It will be interesting to investigate the rest of the structures for the online fulfilment problem. Lastly, future research can extend our model to examine the case when online demand is not homogenous across the given service region. More advanced models can be developed to examine the type and number of fulfilment facilities (e.g., stores or DCs) needed in a region with unevenly distributed online demand.

Chapter 6

Conclusions and future research

6.1 Summary of contributions

With the proliferation of digital technologies, customers desire a more flexible and seamless shopping experience so that they could shop whenever, wherever and however they wish. The changing retail landscape has accelerated the provision of omni-channel retailing services. An increasing number of retailers are staking their future on omni-channel retailing (Verheof et al., 2015, KPMG, 2018). As omni-channel retailing is a nascent phenomenon, many businesses are grappling with the problem of achieving a balance between the costs incurred and the retail services provided. The ongoing high street retail crisis with an increasing number of store closures including many big names (e.g., IKEA, M&S) shows that retailers may struggle to adapt to the omni-channel era (BBC News, 2019a; BBC News, 2020). The dynamic and challenging retail landscape and the lack of guidance from academic research motivated our study. This thesis has aimed at identifying critical supply chain and operations management decisions for retailers to deliver omni-channel performance and developing quantitative models to provide insights into some of these decisions.

We discussed the background and motivation for this study in detail in chapter 1. We introduced the transformation of retailing to omni-channel, examined the state of omni-channel execution worldwide, and discussed the challenges and necessary adaptations regarding supply chain and operations management for retailers to deliver an omni-channel customer experience. Motivated by this exciting and testing retail climate, we clarified our main research objective - develop quantitative decision models to provide guidelines for retailers to make supply chain, operations and logistics decisions when evolving towards omni-channel.

To pursue our main research objective, we developed a conceptual map highlighting key supply chain and operations decisions for fulfilling demand in omni-channel retailing in chapter 2, based on an extensive and critical review of both academic studies and the business literature. Due to our aspiration to aid decision-making using quantitative research approaches, our review concentrated on research that focuses on retailers operating multiple retailing channels and that uses quantitative methods. This critical review is of significant value in informing research in the omni-channel domain. The major contributions of this review chapter are:

- We differentiate the three terminologies, namely multi-, cross- and omni-channel retailing, and show how the degree of flexibility for both front-end customer experiences

and back-end operations progresses from multi-channel to omni-channel retailing. By distinguishing these terminologies, we illustrate the trajectory and necessary efforts for retailers to transform from multi-channel to cross-channel and eventually to omni-channel.

- We identify four specific supply chain and operations planning areas for omni-channel fulfilment – 1. Inventory management, 2. Warehousing and fulfilment, 3. Last mile delivery, and 4. Returns management. We discuss the key issues and the contributions from existing quantitative research in each of these areas and underline those questions that call for more research attention in each area. Such discussions lead to the development of a conceptual map.
- Our newly developed conceptual map (Figure 10) is a powerful visualization tool to facilitate understanding on what omni-channel services mean to customers and how much effort is entailed for retailers to provide such services. Specifically, the conceptual map features four areas of channel integration that can be triggered by customers and therefore affect customers' omni-channel experience. These four areas shaping front-end customer experiences are – 1. Order information exchange, 2. Channel assortment, 3. Delivery options, and 4. Returns policies. It also highlights the four distinct areas of supply chain and operations planning noted above and shows the level of channel integration that can be controlled by retailers to achieve back-end operational efficiencies. Additionally, it pinpoints the contextual factors, including the characteristics of the retailer, market, product, and supplier, that can impact omni-channel decisions.

The key decisions emphasised in the conceptual map justify the three research problems investigated in the thesis from chapter 3 to chapter 5 and provide potential avenues for future research.

In chapter 3, we examined the operational decisions on BOPS fulfilment when the retailer uses local store inventories to satisfy same day BOPS customers. Having its origins as a 'substitute' service to reduce the pressure on overloaded HD services, BOPS has become increasingly popular and has been offered by many retailers to boost customer service and reduce their operating costs. We studied the typical BOPS service format: order before X (i.e., the order cut-off time) and collect after Y (the earliest collection time and therefore the deadline by which orders should be ready for collection). Given the cut-off time and the deadline for fulfilment, the retailer needs to decide the time to start picking, the number of picking waves initiated per cycle and the minimum picking rate needed to achieve a target service level. We used proportion of orders fulfilled by the deadline (NSD) to measure the

service level. We investigated both scenarios when the retailer picks one wave and when the retailer picks multiple waves per order cycle. Expressions for NSD have been obtained based on the time to start picking, the picking rate needed, and the number of picking waves per cycle (in the generalised multi-wave scenario). Best Performance Frontiers (BPFs) have been derived using computational methods and the impacts of the timing and the magnitudes of demand rate fluctuations have been examined through numerical experiments. The major contributions of our model results and the analyses in chapter 3 are:

- We present BPFs, which indicate the best combination of the time to start picking and the minimum picking rate needed to achieve a specific service level. Retailers can use the BPFs to improve their online order picking operations in conventional retail stores with the lowest staffing level (i.e., the minimum picking rate) and the appropriate schedule of picking activities (i.e., the wave release time) to guarantee their desired service levels. The BPFs can also help to identify the highest NSD attainable for a given picking rate. Retailers could use the BPFs as a diagnostic tool to benchmark their current operations against the best achievable performance.
- We show that a retailer could lower the picking rate needed by launching more picking waves per day. However, the time to start picking will be earlier and therefore the total picking duration for each cycle is longer. The analysis highlights the trade-off between costs of a higher picking rate and costs for a longer picking duration. The effect of reducing the picking rate by initiating more picking waves is shown to diminish as the number of picking waves per cycle increases. An equation (Eq. (3.8)) has been proposed to capture the interaction between the time to start picking, the picking rate needed, and the number of picking waves launched per cycle, which helps retailers to determine the number of picking waves per cycle and the wave release times flexibly based on the budget or resources they have for BOPS order fulfilment operations.
- We study the scenarios when a retail experiences two different arrival rates per cycle, e.g., an early (late) demand surge when most customers place their orders well before (close to) the order cut-off time. The analysis indicates which scenario is more beneficial for retailers. Specifically, we show that a shorter duration of early demand surge requires a lower picking rate compared to a late demand surge scenario. When the two arrival rates are significantly different, the BOPS demand pattern (i.e., whether it is an early or late surge) leads to considerably different minimum picking rates. The analysis suggests that retailers should take into account the timing and duration of the demand change when determining the level of picking rate for BOPS order fulfilment, particularly when demand fluctuates considerably.

In chapter 4, we investigated the strategic decision on the order receipt window for BOPS services motivated by the differences in the BOPS order receipt windows offered by retailers across different retail sectors. An inappropriate BOPS order receipt window could adversely affect both the operating costs and customer services. The length of order receipt window is a problem arising at the marketing-operations interface and deserves particular research attention. To this end, we built on the BOPS order fulfilment problem examined in chapter 3 to develop a new model that jointly optimises the order receipt window and the picking rate to minimise total costs. The modelling approach has multiple facets. First, we allow for STS initiatives to serve BOPS customers. In other words, BOPS orders can be fulfilled from local stores as well as from more central facilities such as warehouses before being shipped to the designated stores for collection. The latter option (STS) necessitates a transportation window from the fulfilment facility to the collection store. Second, the order receipt window for BOPS demand is a critical decision variable to be optimised, rather than a given parameter (i.e., one day) as in the analytical models in chapter 3. Third, the problem is formulated as a cost minimisation model, which considers the WIP cost, late penalty cost, fixed order picking cost and variable picking cost. The service level is a system variable to be determined by the optimal picking rate, the time to start picking and the order receipt window length. The major contributions of our model results and the analyses in chapter 4 are:

- The work is the first to optimise the order receipt window for BOPS services in a deadline-driven system. Given a promised deadline, the cost minimisation models can be used to determine the order receipt window for BOPS orders, the speed offered to customers for BOPS collection, and the picking rate required to process BOPS orders in both decentralised scenario (i.e., fulfil BOPS orders from local stores) and more centralised scenario (i.e., fulfil BOPS orders from more central facilities such as warehouses).
- By introducing the STS transportation window, we provide insights on the impact of the centralisation level of order fulfilment on decision variables such as the BOPS order receipt window. The analysis reveals that the cost impact from picking centralisation depends on the relative importance of inventory costs compared to variable picking costs due to store congestion. The highlight of the trade-offs between these two costs can be helpful when extending our models to investigate the optimal centralisation level of online fulfilment.
- We apply our model results to investigate and understand the offerings of 5 major UK-based retailers in terms of their BOPS services. The insights form the basis for two new generic frameworks of retail operating strategies based on (1) product types and (2) the retailers' business profile. The model results and the frameworks provide clear insights on strategic decisions on BOPS service offerings (e.g., the length of order receipt

window and the collection speed) and operational decision on the picking rate needed, in order to balance cost and service. Specifically, the frameworks show the impact of both product characteristics and retailer characteristics on the BOPS service speed, and when same day, next day or more than two days can be desirable.

In chapter 5, we studied the strategic decision on from where to fulfil online orders. For retailers with multiple sales channels, online fulfilment sources include retail stores, DCs, EFCs and fulfilment directly by suppliers via drop-shipping. In practice, retailers from different product sectors are using different combinations of fulfilment options. The mixed successes and failures demonstrate that the fulfilment strategy is critical for profitability (e.g., Webvan versus Tesco discussed in section 1.4). We developed a typology of online fulfilment network configurations exemplified by leading UK-based retailers for both groceries and non-food products. The typology highlights the commonalities and differences between four distinct fulfilment configurations. We focused our attention on the critically important online groceries fulfilment network, which typically comprises a set of retail stores and EFC(s). We derived analytical expressions for the total cost as a function of the centralisation level of online picking, which in turn defines the number of stores to be deployed as picking facilities to fulfil online orders. The major contributions of our model results and the analyses in chapter 5 are:

- The typology is the first to categorise the commonly used online order fulfilment networks by retailers with both online and offline channels across various product sectors. We identify the key characteristics of four distinct fulfilment configurations – 1. Inventory pooling at the store level for online and offline channels by utilising retail stores and potentially EFCs for online order fulfilment, 2. Inventory pooling at the DC level for online and offline channels by using DCs to handle traditional store replenishment as well as online fulfilment, whilst stores are simply used as BOPS collection points, 3. Inventory pooling at both the store and DC level for online and offline channels so that online orders can be fulfilled more flexibly from retail stores and DCs, and 4. The supplement of supplies' distribution facilities (i.e., drop-shipping) to retailers' self-owned stocking facilities to fulfil a wider range of online orders. This typology helps both researchers and practitioners to understand contemporary retail practice with regard to online order fulfilment.
- We develop analytical models to help retailers determine the number of retail facilities (e.g., stores and DCs) to be deployed as online fulfilment facilities in order to serve online customers from a given service region. A distinct benefit of these models is that they require only limited customer data (e.g., only demand density across the region rather than the exact demand locations and volumes) to provide clear and relevant

insights into the critical decision on the number of picking facilities needed to fulfil online demand. One immediate observation from the univariate functional expression of the total cost is that the size of a retailer's service region has no influence on the strategic decision on the centralisation level of online fulfilment.

- We conduct computational experiments to investigate how different business scenarios with varying model parameters affect the optimal number of picking facilities needed for online fulfilment. Our analysis shows that retailers should deploy more stores for online fulfilment (i.e., decentralised online fulfilment) in the case of (1) high demand density, (2) high BOPS demand ratio to the total online demand, (3) high unit transportation costs, (4) low delivery vehicle capacity, (5) low fixed facility cost and (6) long travel distance tolerable for customers to collect BOPS orders. We also discuss when the retailer should use certain incentives to encourage customers to travel further for BOPS collection and to steer customers to choose BOPS over HD services.

The main research objective of this thesis has been successfully achieved. The development of the conceptual map has identified key supply chain and operations management decisions. Three research problems have been investigated on the basis of the conceptual map and have been addressed by developing new models that have enabled informative and valuable analyses to be conducted. This thesis has not only facilitated research development in the omni-channel retailing domain but also provided managerial insights that should be relevant for practitioners in the contemporary retail industry. The following section discusses the future work of our thesis.

6.2 Future research

The conceptual map developed in chapter 2 highlights a wide range of issues that are relevant to decision making in the omni-channel retail landscape (see Table 2) and justifies the choice of the three research problems examined in this thesis. Future research should investigate the whole range of important issues to provide meaningful insights and implications for both scholars and practitioners to embrace the omni-channel retail strategy. Topics of particular importance and relevance include online fulfilment operations, last mile delivery planning, returns management, and demand management using the concepts from revenue management (i.e., service differentiation for deliveries and/or returns with discriminating prices to manage customer demand for maximum revenue) (DeHoratius et al., 2020).

For the first two research problems, we cast light on decision-making for BOPS service offerings (i.e., order receipt window and collection speed) and fulfilment operations (i.e., the time to start picking and the picking rate needed). For both problems, we assume that the

demand rate and the order picking rate are constant and continuous. It would be valuable to relax these assumptions and extend our models to the case of time-varying demand rate and/or picking rate. Demand seasonality can also be considered to enrich studies on this subject. We have focused our studies on BOPS demand in the first two research problems given their prominence and importance. However, stores are also used as fulfilment facilities to handle direct HD orders (i.e., SFS) as well as BOPS orders. In this case, online fulfilment facilities (e.g., retail stores) could process more than one stream of demand with different deadlines. As our models consider only one demand stream (i.e., BOPS demand) with one deadline, it would be valuable to extend them to examine the case when retail stores handle both BOPS demand and SFS demand with different deadlines.

In the second research problem, we examine one of the most commonly used BOPS service form: order before X and collect after Y. Another increasingly prevalent service form is 'pickup in store after X hours', which offers continuous customer ordering opportunities with $T = 0$. Though this continuous ordering service form is only offered in some retailers for limited product ranges at selected stores, it could be valuable to draw comparison between these two service forms in order to understand when and how retailers should set a positive order receipt window ($T > 0$) and when retailers are better off with a continuous ordering window ($T = 0$). In addition, we assume infinite capacity at the fulfilment locations (e.g., retail stores and central warehouses). This assumption is reasonable because the unit variable picking cost in our model is non-decreasing in the picking rate, which acts as a constraint to the number of orders that can be processed in fulfilment facilities. However, it may be worth investigating our second research problem by considering specific fulfilment capacity constraints.

For the third research problem, we present a typology of commonly used online fulfilment network configurations and focus our attention on the important groceries online fulfilment network, which is very challenging due to the distinctive features related to groceries. As a result, we primarily analyse two main fulfilment structures for groceries, i.e., fulfil from retail stores and fulfil from DCs/EFCs, though our typology distinguishes four existing fulfilment configurations with other elements such as drop-shipping. Future work can examine other types of fulfilment network configurations, such as the fulfilment network for bulky products (e.g., furniture) via the drop-shipping option. In our analytical model, we explore the decision on the number of stores or DCs to be deployed as online fulfilment facilities, with the assumption that the retailer's distribution network is static. In practice, as more customer spending is made online, many retailers feel the need to adjust their store network by closing some stores and opening small-format stores (Forbes, 2019d). The

opening and closure of stores affects the level of convenience for customers with regard to BOPS services, i.e., the number of collection stores available and the average travel distance to each collection store. Some retailers, such as Next, have explored the upside and downside of keeping some unprofitable stores as collection points for BOPS customers to ensure service (Next, 2019). Therefore, it would be very pertinent and interesting to investigate the strategic decisions on the opening and closure of retail stores and how to fulfil online orders with the evolving store network.

In this thesis, we have developed decision models to aid retailers in making critical supply chain and operations management decisions to serve omni-channel customers. Our focus throughout has been on quantitative model development and we have used publicly available data to show insights. Our model results and the analyses conducted provide valuable managerial insights for retailers to deliver omni-channel performance with a balance between cost and service. We advocate relevant empirical research to further inform omni-channel decision making and to perform empirical testing of our models in specific retail contexts.

A Appendix for Chapter 3

A.1 The derivation of *NSD* and the critical point when the retailer encounters two arrival rates per cycle.

We only focus on the scenario when $t_d - T < t_w \leq t_c$, since the case when picking starts too late (i.e., $t_c < t_w \leq t_d$) or too early (i.e., $t_c - T < t_w \leq t_d - T$) is not an interesting scenario for retailers. When there are two arrival rates in one cycle, the retailer may start picking before or after the demand shift point. We discuss these two cases separately.

Case 1: Picking starts no later than the demand shift point, $(t_d - T < t_w \leq t^*)$.

The number of orders arriving in the current cycle is $\lambda_1\big(t^* - (t_c - T)\big) + \lambda_2(t_c - t^*) = (\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T$. If picking activities finish before the deadline (see Figure 46(a)), then $t_f = t_w + ((\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T)/\mu \leq t_d$, thus $\mu \geq ((\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T)/(t_d - t_w)$. Orders arriving from $t_c - T$ to t_w have been picked by the deadline. Thus:

$$NSD = \frac{(t_w - (t_c - T))\lambda_1}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} = \frac{(t_w - t_c + T)\lambda_1}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}.$$

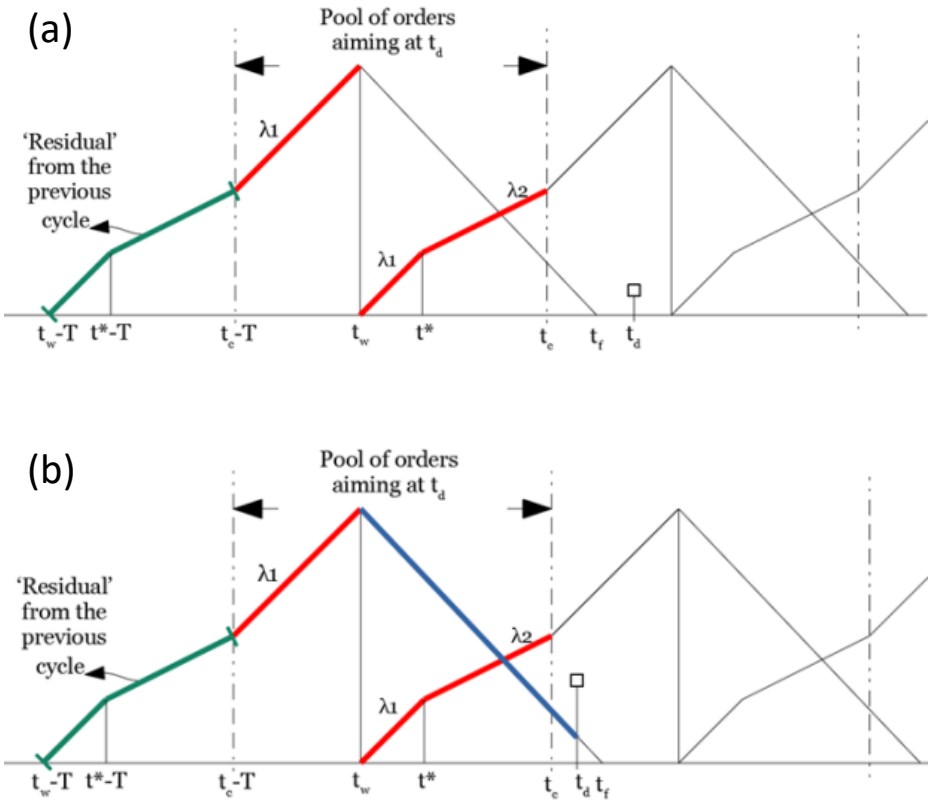


Figure 46. Picking wave starts before the demand shift point and (a) finishes before the deadline, (b) finishes after the deadline.

If picking activities finish after the deadline (see Figure 46(b)), then $\mu < ((\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T)/(t_d - t_w)$. By the deadline, $(t_d - t_w)\mu$ orders have been picked. However, residual orders from $t_w - T$ to $t_c - T$ are among those picked orders. Thus:

$$NSD = \frac{(t_d - t_w)\mu - \lambda_1\big((t^* - T) - (t_w - T)\big) - \lambda_2((t_c - T) - (t^* - T))}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} = \frac{(t_d - t_w)\mu}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} - \frac{\lambda_1(t^* - t_w) + \lambda_2(t_c - t^*)}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}.$$

In summary, when picking starts no later than the demand shift point ($t_d - T < t_w \le t^*$):

$$NSD = \begin{cases} \frac{(t_d - t_w)\mu}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} - \frac{\lambda_1(t^* - t_w) + \lambda_2(t_c - t^*)}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} & for \ \frac{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}{T} \le \mu < \frac{((\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T)}{(t_d - t_w)} \\ \frac{(t_w - t_c + T)\lambda_1}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} & for \ \ \mu \ge \frac{((\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T)}{(t_d - t_w)} \end{cases}.$$

With a given service level, β, the wave release time ${t_w}^* = \beta\left(1 - \frac{\lambda_2}{\lambda_1}\right)(t^* - t_c) + \beta T + t_c - T$ and the minimum picking rate $\mu^* = \frac{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}{t_d - t_c + (1-\beta)T - \beta\left(1 - \frac{\lambda_2}{\lambda_1}\right)(t^* - t_c)}$ on the BPF.

Case 2: Picking starts after the demand breaking point, $(t^* < t_w \le t_c)$.

If picking activities finish before deadline (see Figure 47(a)), then $\mu \ge ((\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T)/(t_d - t_w)$.

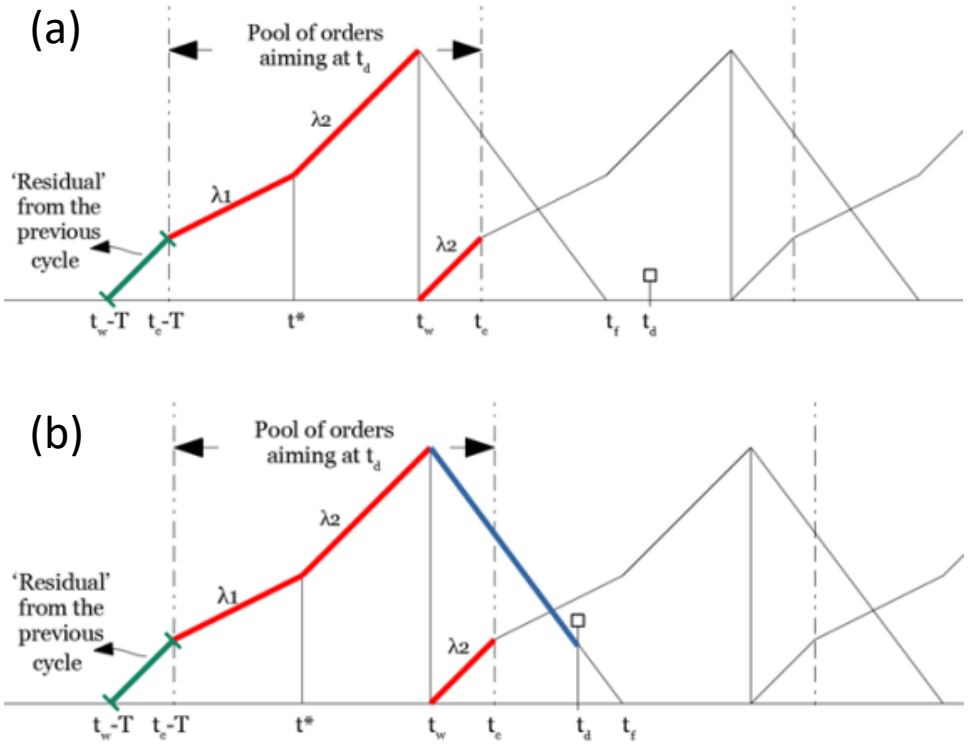


Figure 47. Picking wave starts after the demand shift point and (a) finishes before the deadline, (b) finishes after the deadline.

The number of orders arriving in the current cycle is $(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T$. By the deadline, orders arriving from $t_c - T$ to t_w have been picked. Thus:

$$NSD = \frac{\lambda_1(t^* - (t_c - T)) + \lambda_2(t_w - t^*)}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} = \frac{(\lambda_1 - \lambda_2)t^* - \lambda_1(t_c - T) + \lambda_2 t_w}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}.$$

If picking activities finish after deadline (see Figure 47(b)), then $\mu < ((\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T)/(t_d - t_w)$.

By the deadline, $(t_d - t_w)\mu$ orders have been picked. However, residual orders from $t_w - T$ to $t_c - T$ are among those picked orders. Thus:

$$NSD = \frac{(t_d - t_w)\mu - \lambda_2((t_c - T) - (t_w - T))}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} = \frac{(t_d - t_w)}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}\mu - \frac{\lambda_2(t_c - t_w)}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}.$$

In summary, when picking starts after the demand shift point ($t_w > t^*$):

$$NSD = \begin{cases} \frac{(t_d - t_w)\mu}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} - \frac{\lambda_2(t_c - t_w)}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} & for \frac{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}{T} < \mu < \frac{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}{(t_d - t_w)} \\ \frac{(\lambda_1 - \lambda_2)t^* - \lambda_1(t_c - T) + \lambda_2 t_w}{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T} & for \quad \mu \geq \frac{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}{(t_d - t_w)} \end{cases}.$$

With a given service level, β, the wave release time ${t_w}^* = t_c - (1 - \beta)\left(\frac{\lambda_1}{\lambda_2}(t^* + T - t_c) + t_c - t^*\right)$ and the minimum picking rate $\mu^* = \frac{(\lambda_1 - \lambda_2)(t^* - t_c) + \lambda_1 T}{t_d - t_c + (1-\beta)\left(\frac{\lambda_1}{\lambda_2}(t^* + T - t_c) + t_c - t^*\right)}$ on the BPF.

A.2 Proof of proposition 4

When a retailer sets a target service level β and picks $N - 1$ waves per cycle, the minimum picking rate μ_{N-1} required satisfies:

$$\frac{\lambda^{N-1}(\mu_{N-1} - \lambda)}{\mu_{N-1}({\mu_{N-1}}^{N-1} - \lambda^{N-1})} = \frac{t_d - t_c}{T} + (1 - \beta). \tag{3.9}$$

Based on Eqs. (3.8) and (3.9), we have the following expression:

$$\frac{\lambda^N(\mu_N - \lambda)}{\mu_N({\mu_N}^N - \lambda^N)} = \frac{t_d - t_c}{T} + (1 - \beta) = \frac{\lambda^{N-1}(\mu_{N-1} - \lambda)}{\mu_{N-1}({\mu_{N-1}}^{N-1} - \lambda^{N-1})},$$

while

$$\frac{\lambda^N(\mu_N - \lambda)}{\mu_N({\mu_N}^N - \lambda^N)} = \frac{\lambda^N}{\mu_N({\mu_N}^{N-1} + {\mu_N}^{N-2}\lambda + {\mu_N}^{N-3}\lambda^2 + \cdots + \mu_N\lambda^{N-2} + \lambda^{N-1})} = \frac{1}{\sum_{i=1}^{N}(\frac{\mu_N}{\lambda})^i},$$

and

$$\frac{\lambda^{N-1}(\mu_{N-1} - \lambda)}{\mu_{N-1}({\mu_{N-1}}^{N-1} - \lambda^{N-1})} = \frac{\lambda^{N-1}}{\mu_{N-1}({\mu_{N-1}}^{N-2} + {\mu_{N-1}}^{N-3}\lambda + {\mu_{N-1}}^{N-4}\lambda^2 + \cdots + \mu_{N-1}\lambda^{N-3} + \lambda^{N-2})} = \frac{1}{\sum_{i=1}^{N-1}(\frac{\mu_{N-1}}{\lambda})^i}. \tag{3.10}$$

Denote $q = \mu_N/\lambda$ and $q_0 = \mu_{N-1}/\lambda$. Eq. (3.10) leads to the following expressions:

$$\sum_{i=1}^{N} q^i = \frac{q(1-q^N)}{(1-q)} = \sum_{i=1}^{N-1} {q_0}^i = \frac{q_0(1-{q_0}^{N-1})}{(1-q_0)},$$

$$\frac{q_0(1-{q_0}^{N-1})}{(1-q_0)} = \frac{q_0({q_0}^N-1)}{(q_0-1)} - {q_0}^N = \frac{q(q^N-1)}{(q-1)}.$$

Since ${q_0}^N > 0$, $\frac{q_0({q_0}^N-1)}{(q_0-1)} > \frac{q(q^N-1)}{(q-1)}$ holds. Denote $h(x) = \frac{x(x^N-1)}{(x-1)} = x^N + x^{N-1} + x^{N-2} + .. + x$ and then $\frac{dh(x)}{dx} = Nx^{N-1} + (N-1)x^{N-2} + \cdots + 1 > 0$. As a result, with $h(q_0) = \frac{q_0({q_0}^N-1)}{(q_0-1)} > h(q) = \frac{q(q^N-1)}{(q-1)}$, we have $q_0 > q$ and therefore $\mu_{N-1} > \mu_N$, meaning that the picking rate required to achieve the same service level for $N-1$ waves is higher than that for N waves picking model.

For a given service level β, $\beta\lambda T$ orders arriving within the current cycle will have been picked by the deadline, whether a retailer picks $N-1$ or N waves. In a BPF policy, the retailer picks $(t_d - t_{w,1})\mu_N$ orders if N picking waves are launched per cycle and $((t_c - T) - (t_{w,N} - T))\lambda$ orders are 'residual' from previous cycle. Likewise, $(t_d - t_{w,1}{}')\mu_{N-1}$ orders will be picked if the retailer picks $N-1$ waves per cycle while $((t_c - T) - (t_{w,N-1} - T))\lambda$ orders are 'residual'. $t_{w,1}$ and $t_{w,1}{}'$ are the release times for the first wave in N and $N-1$ wave picking respectively. Therefore,

$$\begin{aligned}\beta\lambda T &= \left(t_d - t_{w,1}\right)\mu_N - ((t_c - T) - (t_{w,N} - T))\lambda \\ &= \left(t_d - t_{w,1}{}'\right)\mu_{N-1} - ((t_c - T) - (t_{w,N-1} - T))\lambda.\end{aligned}$$

Since $t_{w,N-1} = t_{w,N} = t_c - (1-\beta)T$ and $\mu_{N-1} > \mu_N$, $\left(t_d - t_{w,1}\right) > \left(t_d - t_{w,1}{}'\right)$ and therefore $t_{w,1} < t_{w,1}{}'$ hold. That is to say, a retailer will release the first wave earlier when N waves are picked than is the case when $N-1$ waves are picked per cycle. Thus, proposition 4 is proved.

B Appendix for chapter 4

B.1 The retailer starts picking before the order cut-off time, i.e., $t_w \leq t_c$.

When $t_w \leq t_c$, orders arriving from t_w to t_c will only be processed in the next cycle because the retailer releases only one picking wave each cycle. In a cycle $(t_c - T, t_c]$, the retailer carries out the following activities chronologically (see Figure 48):

1) From t_w to t_p: pick $(t_c - t_w)\lambda$ orders from the previous cycle, which are late;
2) From t_p to t_d: pick $(t_d - t_w)\mu - (t_c - t_w)\lambda$ orders from the current cycle, which are on time;
3) From t_d to t_f: pick $\lambda T - (t_d - t_w)\mu$ orders from the current cycle, which are late.

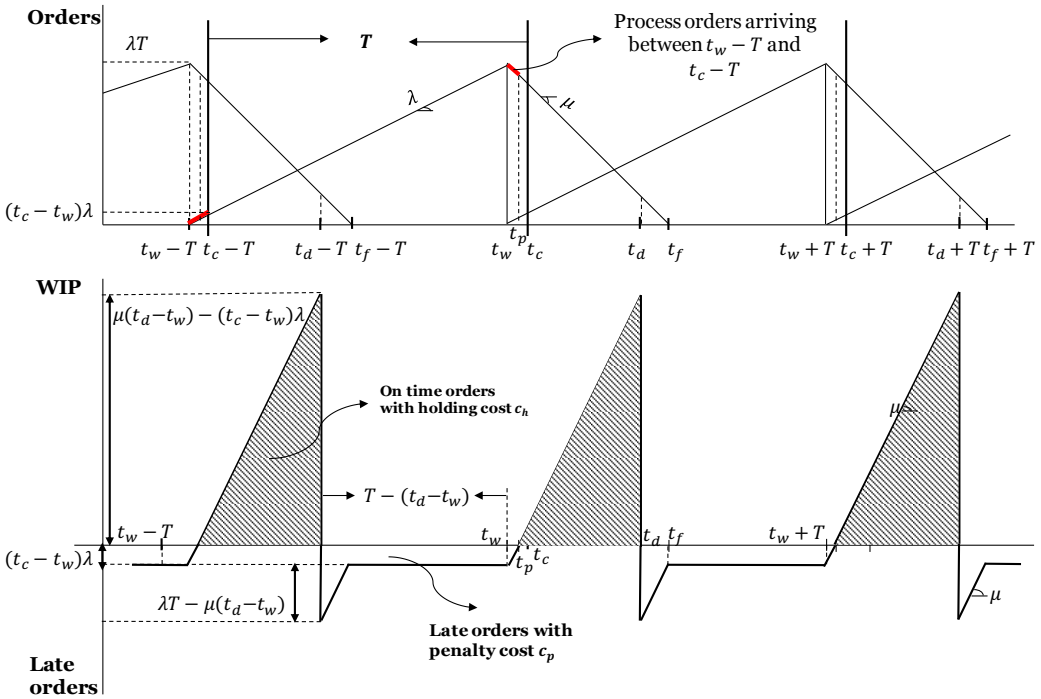


Figure 48. The evolution of order accumulation, WIP inventories and late orders when $t_w \leq t_c$.

As we can see from Figure 48, WIP inventories build up from zero to $(t_d - t_w)\mu - (t_c - t_w)\lambda$ from t_p to t_d, with a duration of $((t_d - t_w) - (t_c - t_w)\lambda)/\mu)$. The average WIP inventory level per time unit is $\frac{1}{T}\left(\frac{(t_d - t_w)\mu - (t_c - t_w)\lambda}{2} \times \frac{(t_d - t_w)\mu - (t_c - t_w)\lambda}{\mu}\right) = \frac{((t_d - t_w)\mu - (t_c - t_w)\lambda)^2}{2T\mu}$. The average level of late orders per time unit from the current cycle is $\frac{1}{T}\left(\frac{\lambda T - (t_d - t_w)\mu}{2} \times \frac{\lambda T - (t_d - t_w)\mu}{\mu}\right) = \frac{(\lambda T - (t_d - t_w)\mu)^2}{2T\mu}$. The average level of late orders per time unit from the previous cycle is $\frac{1}{T}\left((t_c - t_w)\lambda(T - (t_d - t_w)) + \frac{(t_c - t_w)\lambda}{2} \times \frac{(t_c - t_w)\lambda}{\mu}\right) = (t_c - t_w)\lambda - \frac{(t_d - t_w)(t_c - t_w)\lambda}{T} + \frac{(t_c - t_w)^2\lambda^2}{2T\mu}$. The order fulfilment optimisation in the case can be formulated as follows:

(P0) $$\text{Min } TC_0(t_w,\mu,T) = c_h\frac{((t_d-t_w)\mu-(t_c-t_w)\lambda)^2}{2T\mu} + c_p\frac{(\lambda T-(t_d-t_w)\mu)^2}{2T\mu} + c_p(t_c-t_w)\lambda - c_p\frac{(t_d-t_w)(t_c-t_w)\lambda}{T} + c_p\frac{(t_c-t_w)^2\lambda^2}{2T\mu} + \frac{K}{T} + c_\mu(\mu)\lambda, \quad (4.7)$$

Subject to $t_w \le t_c$, $\mu(t_d - t_w) - \lambda T \le 0$, $\mu \ge \lambda$, and $t_w \ge 0$.

Because $\frac{dTC_0}{dt_w} = -\frac{(c_h+c_p)(\mu-\lambda)((t_d-t_w)\mu-(t_c-t_w)\lambda)}{T\mu} < 0$, the total fulfilment costs $TC_0(t_w,\mu,T)$ decreases when t_w increases. To minimise the total fulfilment costs, the retailer will set t_w as late as possible (until its upper bound t_c) and the optimality occurs when $t_w = t_c$. Hence, it is not reasonable for the retailer to start picking before the order cut-off time from a cost perspective.

B.2 KKT necessary conditions for problem (P1)

Define the Lagrangian function

$$L_1(T_e,T,\mu,\beta_1,\beta_2,\beta_3) = TC_1(T_e,T,\mu) + \beta_1(\mu T_e - \lambda T) + \beta_2(T_e - fT) + \beta_3(\lambda-\mu),$$

where β_1, β_2 and β_3 are Lagrange multipliers. The KKT conditions require that

$$\frac{dTC_1}{dT_e} + \beta_1\mu + \beta_2 = 0 \text{ or } \frac{c_hT_e\mu}{T} - \frac{c_p(\lambda T-\mu T_e)}{T} + \beta_1\mu + \beta_2 = 0,$$

$$\frac{dTC_1}{dT} - \beta_1\lambda - \beta_2 f = 0 \text{ or } -\frac{c_h{T_e}^2\mu}{2T^2} + \frac{c_p(\lambda^2T^2-\mu^2{T_e}^2)}{2T^2\mu} - \frac{K}{T^2} - \beta_1\lambda - \beta_2 f = 0,$$

$$\frac{dTC_1}{d\mu} + \beta_1 T_e - \beta_3 = 0 \text{ or } \frac{c_h{T_e}^2}{2T} - \frac{c_p(\lambda^2T^2-\mu^2{T_e}^2)}{2T\mu^2} + \lambda c'_\mu(\mu) + \beta_1T_e - \beta_3 = 0,$$

$$\beta_1(\mu T_e - \lambda T) = 0,$$

$$\beta_2(T_e - fT) = 0,$$

$$\beta_3(\lambda-\mu) = 0,$$

$$\beta_i \ge 0,\ i = 1,2,3. \quad (4.8)$$

The KKT point satisfying Eqs. (4.8) is:

$$T_e = \frac{c_p\lambda T-(\beta_1\mu+\beta_2)T}{(c_h+c_p)\mu},\ T = \sqrt{\frac{2K\mu+(c_h+c_p){T_e}^2\mu^2}{c_p\lambda^2-2(\beta_1\lambda+\beta_2 f)\mu}},\ \mu = \sqrt{\frac{c_p\lambda^2T^2}{(c_h+c_p){T_e}^2+2c'_\mu(\mu)\lambda T+2(\beta_1T_e-\beta_3)T}}. \quad (4.9)$$

Clearly, the exact expression for the KKT point in Eqs. (4.9) depends on the multipliers β_i for $i = 1,2,3$, which we discuss below.

Scenario 1: $\beta_1 = \beta_2 = \beta_3 = 0$.

In this scenario, none of the constraints are binding. The KKT point for Eqs. (4.9) is $T_e = \frac{c_p\lambda T}{(c_h+c_p)\mu}$, $T = \sqrt{\frac{2K\mu+(c_h+c_p){T_e}^2\mu^2}{c_p\lambda^2}}$ and $\mu = \sqrt{\frac{c_p\lambda^2T^2}{(c_h+c_p){T_e}^2+2c'_\mu(\mu)\lambda T}}$.

Because $T_e = \frac{c_p \lambda T}{(c_h+c_p)\mu}$, $T = \sqrt{\frac{2K\mu(c_h+c_p)}{c_h c_p \lambda^2}}$ and $\mu = \sqrt{\frac{c_h c_p \lambda T}{2c'_\mu(\mu)(c_h+c_p)}}$, the unique solution for the two unknowns is: $T = \frac{1}{\lambda}(\frac{2K^2(c_h+c_p)}{c_h c_p c'_\mu(\mu)})^{\frac{1}{3}}$, $\mu = \left(\frac{Kc_h c_p}{2c'_\mu(\mu)^2(c_h+c_p)}\right)^{\frac{1}{3}}$. As a result, $T_e = \frac{c_p \lambda T}{(c_h+c_p)\mu} = (\frac{4Kc_p c'_\mu(\mu)}{c_h{}^2(c_h+c_p)})^{1/3}$ and the total fulfilment cost is $TC_1(T_e, T, \mu) = \lambda(\frac{4Kc_h c_p c'_\mu(\mu)}{c_h+c_p})^{\frac{1}{3}} + c_\mu(\mu)\lambda$. Therefore, the KKT point in this scenario is

$$\overline{\boldsymbol{x}}^1 = \left(\left(\frac{4Kc_p c'_\mu(\mu)}{c_h{}^2(c_h+c_p)}\right)^{\frac{1}{3}}, \frac{1}{\lambda}\left(\frac{2K^2(c_h+c_p)}{c_h c_p c'_\mu(\mu)}\right)^{\frac{1}{3}}, \left(\frac{Kc_h c_p}{2c'_\mu(\mu)^2(c_h+c_p)}\right)^{\frac{1}{3}}, \beta_1 = 0, \beta_2 = 0, \beta_3 = 0\right)^t.$$

Scenario 2: $\beta_1 > 0, \beta_2 = \beta_3 = 0$.

In this scenario, $\mu T_e = \lambda T$, which means that the retailer fulfils all orders accumulated in a cycle by the deadline. $\frac{dTC_1}{dT_e} + \beta_1\mu + \beta_2 = 0$ leads to $c_h\lambda + \beta_1\mu = 0$ and therefore $\beta_1 < 0$. Considering the non-negativity of Lagrange multipliers, this scenario cannot occur.

Scenario 3: $\beta_2 > 0, \beta_1 = \beta_3 = 0$.

In this scenario, $T_e = fT$ and the retailer starts picking exactly at the order cut-off time. The KKT point is:

$$\overline{\boldsymbol{x}}^2 = \left(\sqrt{\frac{2K\mu f^2}{c_h f^2\mu^2 + c_p(\lambda - f\mu)^2}}, \sqrt{\frac{2K\mu}{c_h f^2\mu^2 + c_p(\lambda - f\mu)^2}}, \sqrt{\frac{c_p\lambda^2 T^2}{(c_h+c_p)(fT)^2 + 2c'_\mu(\mu)\lambda T}}, \beta_1 = \beta_3 = 0, \beta_2 = c_p\lambda - (c_h + c_p)f\mu\right)^t.$$

Scenario 4: $\beta_3 > 0, \beta_1 = \beta_2 = 0$.

In this scenario, $\mu = \lambda$. The KKT point for Eqs. (4.9) is $T_e = \frac{c_p T}{(c_h+c_p)}$, $T = \sqrt{\frac{2K+(c_h+c_p){T_e}^2\lambda}{c_p\lambda}}$, $\mu = \lambda$ and the multiplier $\beta_3 = c'_\mu(\mu)\lambda - \frac{c_h c_p T}{2(c_h+c_p)}$.

The two equations for T_e and T generate the unique solution: $T_e = \sqrt{\frac{2Kc_p}{c_h(c_h+c_p)\lambda}}$, $T = \sqrt{\frac{2K(c_h+c_p)}{c_h c_p \lambda}}$. The corresponding total fulfilment cost is $TC_1(T_e, T, \mu) = \sqrt{\frac{2Kc_h c_p\lambda}{(c_h+c_p)}} + c_\mu(\mu)\lambda$. Hence, the KKT point is:

$$\overline{\boldsymbol{x}}^3 = \left(\sqrt{\frac{2Kc_p}{c_h(c_h+c_p)\lambda}}, \sqrt{\frac{2K(c_h+c_p)}{c_h c_p \lambda}}, \lambda, \beta_1 = 0, \beta_2 = 0, \beta_3 = c'_\mu(\mu)\lambda - \sqrt{\frac{Kc_h c_p}{2\lambda(c_h+c_p)}}\right)^t.$$

Scenario 5: $\beta_1 > 0, \beta_2 > 0, \beta_3 = 0$.

In this scenario, $\mu T_e = \lambda T$ and $T_e = fT$. $\frac{dTC_1}{dT_e} + \beta_1\mu + \beta_2 = 0$. This leads to $c_h\lambda + \beta_1\mu + \beta_2 = 0$ and therefore $\beta_1 < 0$ and/or $\beta_2 < 0$, which cannot occur.

Scenario 6: $\beta_1 > 0, \beta_3 > 0, \beta_2 = 0$.

In this scenario, $\mu T_e = \lambda T$ and $\mu = \lambda$. As a result, $T_e = T$, which contradicts the constraint $T_e \leq fT$. Therefore, this scenario cannot occur.

Scenario 7: $\beta_2 > 0, \beta_3 > 0, \beta_1 = 0$.

In this scenario, $T_e = fT$ and $\mu = \lambda$. The KKT point for Eqs. (4.9) is $T_e = \sqrt{\frac{2Kf^2}{c_p(1-f)^2\lambda + c_h f^2 \lambda}}$, $T = \sqrt{\frac{2K}{c_p(1-f)^2\lambda + c_h f^2 \lambda}}$, $\mu = \lambda$. The multiplier $\beta_2 = c_p\lambda - (c_h + c_p)f\lambda$ and $\beta_3 = c'_\mu(\mu)\lambda - \frac{c_p T - (c_h + c_p)f^2 T}{2}$ and the total fulfilment cost $TC_1(T_e, T, \mu) = \sqrt{2K\lambda(c_p(1-f)^2 + c_h f^2)} + c_\mu(\mu)\lambda$. Hence, the KKT point is:

$$\overline{\boldsymbol{x}}^4 = \left(\sqrt{\frac{2Kf^2}{c_p(1-f)^2\lambda + c_h f^2 \lambda}}, \sqrt{\frac{2K}{c_p(1-f)^2\lambda + c_h f^2 \lambda}}, \lambda, \beta_1 = 0, \beta_2 = c_p\lambda - (c_h + c_p)f\lambda, \beta_3 = \sqrt{\frac{K(c_p - (c_h + c_p)f^2)^2}{2\lambda(c_p(1-f)^2 + c_h f^2)}} + c'_\mu(\mu)\lambda \right)^t.$$

Scenario 8: $\beta_1 > 0, \beta_2 > 0, \beta_3 > 0$.

In this case, all three constraints are binding and then $T_e = fT$ and $T_e = T$, which are contradictory. Therefore, this scenario cannot occur.

All these scenarios are mutually exclusive and there is only one unique KKT point for problem (P1), which is summarised Table 8.

Linear Independence Constraint Qualification (LICQ) for problem (P1)

Let $\nabla g_1(i)$ denote the gradient vector for decision variables T_e, T, μ in the i[th] constraint of problem (P1). Therefore,

$$\nabla g_1(1) = (\mu, -\lambda, T_e)^t,\ \nabla g_1(2) = (1, -f, 0)^t,\ \nabla g_1(3) = (0, 0, -1)^t,\ \nabla g_1(4) = (-1, 0, 0)^t.$$

If there is only one binding constraint, the LICQ is satisfies and the KKT conditions hold true for any feasible solution.

In the case of two binding constraints, note that $\nabla g_1(1)$ and $\nabla g_1(2)$ are linearly dependent if $T_e = 0$ and $\mu = \lambda/f$. As a result, if both $\mu T_e - \lambda T \leq 0$ (1[st] constraint) and $T_e - fT \leq 0$ (2[nd] constraint) are binding, the point ($T_e = 0, T = 0\ \mu = \lambda/f$) violates the LICQ. However, this cannot be optimal because $\lim_{T \to 0} \frac{K}{T} = \infty$ and then $\lim_{T \to 0} TC_1(T_e, \mu, T) = \infty$. With $\lambda > 0$ and $f > 0$, it is clear that any other combination of two gradient vectors will be linearly independent. Therefore, when there are only two binding constraints, the only point that

violates LICQ is ($T_e = 0, T = 0\ \mu = \lambda/f$), which leads to infinite positive total fulfilment cost and this cannot be optimal.

In the case of three binding constraints, if the 1st and 2nd constraints are not binding at the same time, the any possible set of three binding constraint gradients will be linearly independent. If the 1st, 2nd and 3rd constraints are binding, the constraint vectors $\nabla g_1(1)$, $\nabla g_1(2)$ and $\nabla g_1(3)$ are linearly dependent: the equations,

$$a * \begin{pmatrix} \mu \\ -\lambda \\ T_e \end{pmatrix} + b * \begin{pmatrix} 1 \\ -f \\ 0 \end{pmatrix} + c * \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

have non-trivial solutions for the scalars a,b and $c$when $T_e = 0$ and $\mu = \lambda/f$ Therefore, the point($T_e = 0, T = 0\ \mu = \lambda/f$) violates the LICQ. Similarly, if the 1st, 2nd and 4th constraints are binding, the constraint vectors $\nabla g_1(1)$, $\nabla g_1(2)$ and $\nabla g_1(4)$ are linearly dependent because

$$a * \begin{pmatrix} \mu \\ -\lambda \\ T_e \end{pmatrix} + b * \begin{pmatrix} 1 \\ -f \\ 0 \end{pmatrix} + c * \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

has non-trivial solutions for a, b and c when $T_e = 0$. Therefore, when $T_e = 0$, the LICQ cannot be satisfied. Becasue the 1st and 2nd constraints are binding, $T_e = 0$, which also means $T = 0$. Hence, the KKT condition fails to capture the point ($T_e = 0$, $T = 0$, μ), but this point cannot be optimal as explained before.

Therefore, we have shown that the KKT conditions hold true for optimality, because the points at which KKT conditions are not satisfied can never be optimal.

KKT sufficient conditions for problem (P1)

According to KKT second-order sufficient conditions, a KKT point $\bar{\boldsymbol{x}} = (T_e, T, \mu, \beta_1, \beta_2, \beta_3)^t$ is a strict local minimum when the Hessian of the Lagrangian function $L_1(\bar{\boldsymbol{x}})$ is positive definite. Let F_1, G_{1i} be the Hessians of the cost function $TC_1(T_e, T, \mu)$ and the i[th] constraint respectively. Hence, the Hessian for the Lagrangian function at $\bar{\boldsymbol{x}}$ is $\nabla^2 L_1(\bar{\boldsymbol{x}}) = F_1(\bar{\boldsymbol{x}}) + \sum_{i=1}^{3} \beta_i G_{1i}(\bar{\boldsymbol{x}})$.

Clearly, $G_{11} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} = 0$, $G_{12} = G_{13} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = 0$. Therefore, $\nabla^2 L_1(\bar{\boldsymbol{x}}) = F_1(\bar{\boldsymbol{x}})$. For the total cost function $TC_1(T_e, T, \mu) = \frac{c_h {T_e}^2 \mu}{2T} + \frac{c_p(\lambda T - \mu T_e)^2}{2T\mu} + \frac{K}{T} + c_\mu(\mu)\lambda$, the second order partial derivatives are:

$$\frac{d^2 TC_1}{d{T_e}^2} = \frac{(c_h + c_p)\mu}{T}; \frac{d^2 TC_1}{dT^2} = \frac{2K + (c_h + c_p)\mu {T_e}^2}{T^3}; \frac{d^2 TC_1}{d\mu^2} = \frac{c_p \lambda^2 T}{\mu^3} + c''_\mu(\mu)\lambda;$$

$$\frac{d^2TC_1}{dT_e dT} = -\frac{(c_h+c_p)\mu T_e}{T^2}; \frac{d^2TC_1}{dT_e d\mu} = \frac{(c_h+c_p)T_e}{T}; \frac{d^2TC_1}{dTd\mu} = -\frac{c_p\lambda^2T^2+(c_h+c_p)\mu^2{T_e}^2}{2T^2\mu^2}.$$

As a result, the Hessian matrix of $TC_1(T_e, T, \mu)$ is:

$$F_1 = \begin{bmatrix} \frac{(c_h+c_p)\mu}{T} & -\frac{(c_h+c_p)\mu T_e}{T^2} & \frac{(c_h+c_p)T_e}{T} \\ -\frac{(c_h+c_p)\mu T_e}{T^2} & \frac{2K+(c_h+c_p)\mu{T_e}^2}{T^3} & -\frac{c_p\lambda^2T^2+(c_h+c_p)\mu^2{T_e}^2}{2T^2\mu^2} \\ \frac{(c_h+c_p)T_e}{T} & -\frac{c_p\lambda^2T^2+(c_h+c_p)\mu^2{T_e}^2}{2T^2\mu^2} & \frac{c_p\lambda^2T}{\mu^3}+c_\mu''(\mu)\lambda \end{bmatrix}.$$

The first principal minors $\frac{(c_h+c_p)\mu}{T}$, $\frac{2K+(c_h+c_p)\mu{T_e}^2}{T^3}$ and $\frac{c_p\lambda^2T}{\mu^3}+c_\mu''(\mu)\lambda$ are positive. With $c_\mu''(\mu) \geq 0$, the second principal minors are:

$${F_1}^1 = \begin{bmatrix} \frac{2K+(c_h+c_p)\mu{T_e}^2}{T^3} & -\frac{c_p\lambda^2T^2+(c_h+c_p)\mu^2{T_e}^2}{2T^2\mu^2} \\ -\frac{c_p\lambda^2T^2+(c_h+c_p)\mu^2{T_e}^2}{2T^2\mu^2} & \frac{c_p\lambda^2T}{\mu^3}+c_\mu''(\mu)\lambda \end{bmatrix} \geq \frac{8Kc_p\lambda^2T^2\mu-(c_p\lambda^2T^2-(c_h+c_p){T_e}^2\mu^2)^2}{4T^4\mu^4},$$

$${F_1}^2 = \begin{bmatrix} \frac{(c_h+c_p)\mu}{T} & \frac{(c_h+c_p)T_e}{T} \\ \frac{(c_h+c_p)T_e}{T} & \frac{c_p\lambda^2T}{\mu^3}+c_\mu''(\mu)\lambda \end{bmatrix} \geq \frac{c_p(c_h+c_p)\lambda^2T^2-(c_h+c_p)^2{T_e}^2\mu^2}{T^2\mu^2},$$

$${F_1}^3 = \begin{bmatrix} \frac{(c_h+c_p)\mu}{T} & -\frac{(c_h+c_p)\mu T_e}{T^2} \\ -\frac{(c_h+c_p)\mu T_e}{T^2} & \frac{2K+(c_h+c_p)\mu{T_e}^2}{T^3} \end{bmatrix} = \frac{2K(c_h+c_p)\mu}{T^4} > 0.$$

At the KKT point $\bar{\boldsymbol{x}}^1$ in scenario 1, $T = \sqrt{\frac{2K\mu+(c_h+c_p){T_e}^2\mu^2}{c_p\lambda^2}}$. Therefore, $c_p\lambda^2T^2-(c_h+c_p){T_e}^2\mu^2 = 2K\mu$ and then $8Kc_p\lambda^2T^2\mu-\left(c_p\lambda^2T^2-(c_h+c_p){T_e}^2\mu^2\right)^2 = 4K\mu(2c_p\lambda^2T^2-K\mu)$, which indicates ${F_1}^1 > 0$. Also, $T_e = \frac{c_p\lambda T}{(c_h+c_p)\mu}$ at $\bar{\boldsymbol{x}}^1$. Therefore, $(c_h+c_p)^2{T_e}^2\mu^2 = {c_p}^2\lambda^2T^2 < c_p(c_h+c_p)\lambda^2T^2$ and then ${F_1}^2 > 0$. The third principal minor is the Hessian matrix itself:

$$F_1 = \frac{(c_h+c_p)\mu}{T} * {F_1}^1 - \left(-\frac{(c_h+c_p)\mu T_e}{T^2}\right) * {F_1}^2 + \frac{(c_h+c_p)T_e}{T} * {F_1}^3.$$

At $\bar{\boldsymbol{x}}^1$, ${F_1}^1$, ${F_1}^2$ and ${F_1}^3$ are positive and therefore F_1 is also positive. Hence, $\nabla^2L_1(\bar{\boldsymbol{x}}^1) > 0$ and $\bar{\boldsymbol{x}}^1$ is a strict local minimum.

At the KKT point $\bar{\boldsymbol{x}}^2$ in scenario 3, $T = \sqrt{\frac{2K\mu+(c_h+c_p){T_e}^2\mu^2}{c_p\lambda^2-2\beta_2 f\mu}}$ and $\mu = \sqrt{\frac{c_p\lambda^2T^2}{(c_h+c_p){T_e}^2+2c_\mu'(\mu)\lambda T}}$, which gives $K = c_\mu'(\mu)\lambda T\mu-\beta_2fT^2$ with $\beta_2 = c_p\lambda-(c_h+c_p)f\mu$. Meanwhile, the expression for μ shows that $c_p\lambda^2T^2-(c_h+c_p){T_e}^2\mu^2 = 2c_\mu'(\mu)\lambda T\mu^2$. Hence, proving $8Kc_p\lambda^2T^2\mu-\left(c_p\lambda^2T^2-(c_h+c_p){T_e}^2\mu^2\right)^2 > 0$ is equivalent to proving $\mu < \left(\frac{2Kc_p}{c_\mu'(\mu)^2}\right)^{\frac{1}{3}}$. In this scenario, $\lambda \leq \mu \leq \frac{c_p\lambda}{(c_h+c_p)f}$, $\mu < \left(\frac{2Kc_p}{c_\mu'(\mu)^2}\right)^{\frac{1}{3}}$ holds so long as $\max\mu = \frac{c_p\lambda}{(c_h+c_p)f} <$

$\left(\frac{2Kc_p}{c'_\mu(\mu)^2}\right)^{\frac{1}{3}}$. When $\mu = \frac{c_p\lambda}{(c_h+c_p)f}$, $T = \sqrt{\frac{2K\mu}{(c_h+c_p)f^2\mu^2+c_p\lambda^2-2c_p\lambda f\mu}} = \sqrt{\frac{2K}{c_hf\lambda}}$. Meanwhile, $K = c'_\mu(\mu)\lambda T\mu$ gives $T = \frac{K(c_h+c_p)f}{c'_\mu(\mu)c_p\lambda^2}$. $T = \frac{K(c_h+c_p)f}{c'_\mu(\mu)c_p\lambda^2} = \sqrt{\frac{2K}{c_hf\lambda}}$ leads to $2c'_\mu(\mu)^2{c_p}^2\lambda^3 = Kc_h\left(c_h+c_p\right)^2f^3$, which suggests $\max\mu = \frac{c_p\lambda}{(c_h+c_p)f} < \left(\frac{2Kc_p}{c'_\mu(\mu)^2}\right)^{\frac{1}{3}}$ and then ${F_1}^1 > 0$. Also, $c_p\lambda^2T^2 - \left(c_h+c_p\right)f^2T^2\mu^2 = 2K\mu + 2c_p\lambda f\mu T^2 - 2\left(c_h+c_p\right)f^2T^2\mu^2 > 0$, which proves ${F_1}^2 > 0$ with $T_e = fT$. Likewise, with ${F_1}^1$, ${F_1}^2$ and ${F_1}^3$ being positive, F_1 is also positive. Hence, $\nabla^2L_1(\overline{\boldsymbol{x}}^2) > 0$ and $\overline{\boldsymbol{x}}^2$ is a strict local minimum.

In scenario 4, $\nabla^2L_1(\overline{\boldsymbol{x}}^3)$ is exactly the same with $\nabla^2L_1(\overline{\boldsymbol{x}}^1)$ by replacing μ with λ. Therefore, $\nabla^2L_1(\overline{\boldsymbol{x}}^3) > 0$ and $\overline{\boldsymbol{x}}^3$ is a strict local minimum.

In scenario 7, $\nabla^2L_1(\overline{\boldsymbol{x}}^4)$ is exactly the same with $\nabla^2L_1(\overline{\boldsymbol{x}}^2)$ by replacing μ with λ. Thus, $\nabla^2L_1(\overline{\boldsymbol{x}}^4) > 0$ and $\overline{\boldsymbol{x}}^4$ is a strict local minimum.

So far, we have demonstrated that there is only one unique KKT point and that this point is a strict local minimum. Among the three variables, $T_e \in (0, \frac{\lambda T}{\mu})$, $\mu \in [\lambda, +\infty]$ and $T \in (0, +\infty]$. When T is infinite, $\lim_{T\to\infty} \frac{c_p(\lambda T-\mu T_e)^2}{2T\mu} = \infty$ and therefore $\lim_{T\to\infty} TC_1(T_e, \mu, T) = \infty$. Likewise, when μ is infinite, $\lim_{\mu\to\infty} c_\mu(\mu)\lambda = \infty$ and therefore $\lim_{\mu\to\infty} TC_1(T_e, \mu, T) = \infty$. As a result, the optimum must be either an interior point or on the boundary, which are listed in Table 8. Therefore, the solution in Table 8 is indeed the global minimum.

B.3 KKT necessary conditions for problem (P2)

Define the Lagrangian function

$$L_2(T_e, T, \mu, u_1, u_2, u_3) = TC_2(T_e, T, \mu) + u_1(\mu T_e - \lambda T) + u_2(\lambda - \mu) + u_3(L - T_e),$$

where u_i for $i = 1, 2, 3$ are Lagrange multipliers. The KKT conditions require that

$$\frac{dTC_2}{dT_e} + u_1\mu - u_3 = 0 \text{ or } \frac{c_hT_e\mu}{T} - \frac{c_p(\lambda T-\mu(T_e-L))}{T} + u_1\mu - u_3 = 0,$$

$$\frac{dTC_2}{dT} - u_1\lambda = 0 \text{ or } -\frac{c_h{T_e}^2\mu}{2T^2} + \frac{c_p\left(\lambda^2T^2-\mu^2(T_e-L)^2\right)}{2T^2\mu} - \frac{K}{T^2} - u_1\lambda = 0,$$

$$\frac{dTC_2}{d\mu} + u_1T_e - u_2 = 0 \text{ or } \frac{c_h{T_e}^2}{2T} - \frac{c_p\left(\lambda^2T^2-\mu^2(T_e-L)^2\right)}{2T\mu^2} + \lambda c'_\mu(\mu) + u_1T_e - u_2 = 0,$$

$$u_1(\mu T_e - \lambda T) = 0,$$

$$u_2(\lambda - \mu) = 0,$$

$$u_3(L - T_e) = 0$$

$$u_i \geq 0, i = 1, 2, 3. \qquad (4.10)$$

If $u_3 > 0$, then $L = T_e$, and none of the orders picked will be delivered on time and the retailer has a zero service level, which is not ideal. Therefore, $T_e > L$ and $u_3 = 0$. The KKT point satisfying Eqs. (4.10) can be written as:

$$T_e = \frac{c_p(\lambda T + L\mu) - u_1 T\mu}{(c_h + c_p)\mu},$$

$$T = \sqrt{\frac{2K\mu + \left(c_p(T_e - L)^2 + c_h {T_e}^2\right)\mu^2}{c_p\lambda^2 - 2u_1\lambda\mu}},$$

$$\mu = \sqrt{\frac{c_p\lambda^2 T^2}{c_p(T_e - L)^2 + c_h {T_e}^2 + 2c'_\mu(\mu)\lambda T + 2T(u_1 T_e - u_2)}}. \tag{4.11}$$

The exact expression for the KKT point in Eqs. (4.11) depends on the multipliers.

Scenario 1: $u_1 = u_2 = 0$.

In this scenario, none of the constraints are binding and the results from Eqs. (4.11) are $T_e = \frac{c_p\lambda T + c_p L\mu}{(c_h + c_p)\mu}$, $T = \sqrt{\frac{2K\mu + (c_p(T_e - L)^2 + c_h {T_e}^2)\mu^2}{c_p\lambda^2}}$, $\mu = \sqrt{\frac{c_p\lambda^2 T^2}{c_p(T_e - L)^2 + c_h {T_e}^2 + 2c'_\mu(\mu)\lambda T}}$. Replacing T_e with $\frac{c_p\lambda T + c_p L\mu}{(c_h + c_p)\mu}$, the KKT point in this scenario is:

$$\overline{\boldsymbol{y}}^1 = \left(\frac{c_p\lambda T + c_p L\mu}{(c_h + c_p)\mu}, \sqrt{\frac{2K\mu(c_h + c_p) + c_h c_p L^2\mu^2}{c_h c_p\lambda^2}}, \sqrt{\frac{c_h c_p\lambda^2 T^2}{c_h c_p L^2 + 2(c_h + c_p)c'_\mu(\mu)\lambda T}}, u_1 = u_2 = u_3 = 0\right)^t.$$

Scenario 2: $u_1 > 0, u_2 = 0$.

When $u_1 > 0$, $\mu T_e = \lambda T$. Replacing T_e with $\lambda T/\mu$ in Eqs. (4.11) yields the KKT point:

$$\overline{\boldsymbol{y}}^2 = \left(\sqrt{\frac{2K + c_p L^2\mu}{c_h\mu}}, \sqrt{\frac{2K\mu + c_p L^2\mu^2}{c_h\lambda^2}}, \sqrt{\frac{c_h\lambda^2 T^2}{c_p L^2 + 2c'_\mu(\mu)\lambda T}}, u_1 = \frac{c_p L\mu - c_h\lambda T}{T\mu}, u_2 = u_3 = 0\right)^t.$$

Scenario 3: $u_1 = 0, u_2 > 0$.

In this scenario, $\mu = \lambda$ and the solutions for Eqs. (4.11) are $T_e = \frac{c_p T + c_p L}{(c_h + c_p)}$ and $T = \sqrt{\frac{2K + (c_p(T_e - L)^2 + c_h {T_e}^2)\lambda}{c_p\lambda}}$. Replacing T_e with $\frac{c_p T + c_p L}{(c_h + c_p)}$, the KKT point is:

$$\overline{\boldsymbol{y}}^3 = \left(\frac{c_p T + c_p L}{(c_h + c_p)}, \sqrt{\frac{2K(c_h + c_p) + c_h c_p L^2\lambda}{c_h c_p\lambda}}, \lambda, u_1 = u_3 = 0, u_2 = \lambda c'_\mu(\mu) - \frac{c_h c_p(T^2 - L^2)}{2T(c_h + c_p)}\right)^t.$$

Scenario 4: $u_1 > 0, u_2 > 0$

In this case, $\mu = \lambda$ and $\mu T_e = \lambda T$, which means $T_e = T$. In this case, the KKT point is:

$$\overline{\boldsymbol{y}}^4 = \left(\sqrt{\frac{2K + c_p L^2\lambda}{c_h\lambda}}, \sqrt{\frac{2K + c_p L^2\lambda}{c_h\lambda}}, \lambda, u_1 = \frac{c_p L - c_h T}{T}, u_2 = \lambda c'_\mu(\mu) - \frac{K}{\lambda T}, u_3 = 0\right)^t.$$

The four scenarios above are mutually exclusive and therefore there is only one KKT point for Problem (P2), which is summarised in Table 9.

LICQ for Problem (P2)

Let $\nabla g_2(i)$ denote the gradient vector for decision variables T_e, T, μ in the i[th] constraint of problem (P2). Therefore,

$$\nabla g_2(1) = (\mu, -\lambda, T_e)^t, \quad \nabla g_2(2) = (0, 0, -1)^t, \quad \nabla g_2(3) = (-1, 0, 0)^t.$$

Note that $\lambda > 0$ and in the feasible region, $\mu > 0$ and $T_e \geq L > 0$. Therefore, the constraint gradient vectors $\nabla g_2(1)$, $\nabla g_2(2)$ and $\nabla g_2(3)$ are linearly independent for any feasible solution. As a result, KKT conditions hold true at optimality.

KKT sufficient conditions for problem (P2)

Let F_2, G_{2i} be the Hessians of the cost function $TC_2(T_e, T, \mu)$ and the i[th] constraint of problem (P2) respectively. The Hessian for the Lagrangian function at a feasible solution $\bar{\boldsymbol{y}}$ is $\nabla^2 L_2(\bar{\boldsymbol{y}}) = F_2(\bar{\boldsymbol{y}}) + \sum_{i=1}^{3} u_i G_{2i}(\bar{\boldsymbol{y}})$.

Clearly, $G_{21} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} = 0$, $G_{22} = G_{23} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = 0$. Therefore, $\nabla^2 L_2(\bar{\boldsymbol{y}}) = F_2(\bar{\boldsymbol{y}})$. The cost function $TC_2(T_e, T, \mu) = c_h \frac{\mu {T_e}^2}{2T} + c_p \frac{(\lambda T - \mu(T_e - L))^2}{2T\mu} + \frac{K}{T} + c_\mu(\mu)\lambda$ and the second partial derivatives are:

$$\frac{d^2TC_2}{d{T_e}^2} = \frac{(c_h + c_p)\mu}{T}; \frac{d^2TC_2}{dT^2} = \frac{2K + (c_h {T_e}^2 + c_p(T_e - L)^2)\mu}{T^3}; \frac{d^2TC_2}{d\mu^2} = \frac{c_p\lambda^2 T}{\mu^3} + c_\mu''(\mu)\lambda;$$

$$\frac{d^2TC_2}{dT_e dT} = -\frac{(c_p(T_e - L) + c_h T_e)\mu}{T^2}; \frac{d^2TC_2}{dT_e d\mu} = \frac{c_p(T_e - L) + c_h T_e}{T}; \frac{d^2TC_2}{dT d\mu} = -\frac{c_h {T_e}^2\mu^2 + c_p(\lambda^2 T^2 + (T_e - L)^2\mu^2)}{2T^2\mu^2}.$$

As a result, the Hessian matrix of $TC_2(T_e, T, \mu)$ is:

$$F_2 = \begin{bmatrix} \frac{(c_h + c_p)\mu}{T} & -\frac{(c_p(T_e - L) + c_h T_e)\mu}{T^2} & \frac{c_p(T_e - L) + c_h T_e}{T} \\ -\frac{(c_p(T_e - L) + c_h T_e)\mu}{T^2} & \frac{2K + (c_h {T_e}^2 + c_p(T_e - L)^2)\mu}{T^3} & -\frac{c_h {T_e}^2\mu^2 + c_p(\lambda^2 T^2 + (T_e - L)^2\mu^2)}{2T^2\mu^2} \\ \frac{c_p(T_e - L) + c_h T_e}{T} & -\frac{c_h {T_e}^2\mu^2 + c_p(\lambda^2 T^2 + (T_e - L)^2\mu^2)}{2T^2\mu^2} & \frac{c_p\lambda^2 T}{\mu^3} + c_\mu''(\mu)\lambda \end{bmatrix}.$$

Clearly, the first principal minors $\frac{(c_h + c_p)\mu}{T}$, $\frac{2K + (c_h {T_e}^2 + c_p(T_e - L)^2)\mu}{T^3}$ and $\frac{c_p\lambda^2 T}{\mu^3} + c_\mu''(\mu)\lambda$ are positive. With $c_\mu''(\mu) \geq 0$, the second principal minors are:

$${F_2}^1 = \begin{bmatrix} \frac{2K + (c_h {T_e}^2 + c_p(T_e - L)^2)\mu}{T^3} & -\frac{c_h {T_e}^2\mu^2 + c_p(\lambda^2 T^2 + (T_e - L)^2\mu^2)}{2T^2\mu^2} \\ -\frac{c_h {T_e}^2\mu^2 + c_p(\lambda^2 T^2 + (T_e - L)^2\mu^2)}{2T^2\mu^2} & \frac{c_p\lambda^2 T}{\mu^3} + c_\mu''(\mu)\lambda \end{bmatrix} \geq$$

$$\frac{8Kc_p\lambda^2 T^2\mu - (c_p\lambda^2 T^2 - \mu^2(c_h {T_e}^2 + c_p(T_e - L)^2))^2}{4T^4\mu^4}.$$

$${F_2}^2 = \begin{bmatrix} \frac{(c_h + c_p)\mu}{T} & \frac{c_p(T_e - L) + c_h T_e}{T} \\ \frac{c_p(T_e - L) + c_h T_e}{T} & \frac{c_p\lambda^2 T}{\mu^3} + c_\mu''(\mu)\lambda \end{bmatrix} \geq \frac{c_p(c_h + c_p)\lambda^2 T^2 - \left((c_h + c_p)T_e\mu - c_p L\mu\right)^2}{T^2\mu^2}.$$

$$F_2{}^3 = \begin{bmatrix} \frac{(c_h+c_p)\mu}{T} & -\frac{(c_p(T_e-L)+c_hT_e)\mu}{T^2} \\ -\frac{(c_p(T_e-L)+c_hT_e)\mu}{T^2} & \frac{2K+(c_hT_e{}^2+c_p(T_e-L)^2)\mu}{T^3} \end{bmatrix} = \frac{2K\mu(c_h+c_p)+c_hc_pL^2\mu^2}{T^4} > 0.$$

At the KKT point $\bar{\boldsymbol{y}}^1$ in scenario 1, $T = \sqrt{\frac{2K\mu+(c_p(T_e-L)^2+c_hT_e{}^2)\mu^2}{c_p\lambda^2}}$ and then $c_p\lambda^2T^2 - \left(c_p(T_e-L)^2 + c_hT_e{}^2\right)\mu^2 = 2K\mu$. Therefore, $8Kc_p\lambda^2T^2\mu - \left(c_p\lambda^2T^2 - \mu^2\left(c_hT_e{}^2 + c_p(T_e-L)^2\right)\right)^2 = \frac{4K\mu(2c_p\lambda^2T^2-K\mu)}{4T^4\mu^4} > 0$ and $F_2{}^1 > 0$. Also, $T_e = \frac{c_p\lambda T+c_pL\mu}{(c_h+c_p)\mu}$, which indicates $\left(\left(c_h+c_p\right)T_e\mu - c_pL\mu\right)^2 = c_p{}^2\lambda^2T^2$ and therefore $F_2{}^2 > 0$. The third principal minor

$$F_2 = \frac{(c_h+c_p)\mu}{T} * F_2{}^1 - \left(-\frac{(c_p(T_e-L)+c_hT_e)\mu}{T^2}\right) * F_2{}^2 + \frac{c_p(T_e-L)+c_hT_e}{T} * F_2{}^3 > 0.$$

Hence, $\nabla^2L_2(\bar{\boldsymbol{y}}^1) > 0$ and $\bar{\boldsymbol{y}}^1$ is a strict local minimum.

In scenario 2, $u_1 > 0$. Based on Eqs. (4.11), $T_e = \frac{c_p(\lambda T+L\mu)-u_1T\mu}{(c_h+c_p)\mu}$ and then $\left(c_h+c_p\right)T_e\mu - c_pL\mu = c_p\lambda T - u_1T\mu < c_p\lambda T$. Therefore, $\left(\left(c_h+c_p\right)T_e\mu - c_pL\mu\right)^2 < \left(c_p\lambda T\right)^2 < c_p\left(c_h+c_p\right)\lambda^2T^2$, which proves $F_2{}^2 > 0$. At the KKT point $\bar{\boldsymbol{y}}^2$, $T = \sqrt{\frac{2K\mu+c_pL^2\mu^2}{c_h\lambda^2}}$ and then $c_h\lambda T^2 = 2K\mu + c_pL^2\mu^2$. Since $\mu T_e = \lambda T$, $(c_p\lambda^2T^2 - \mu^2(c_hT_e{}^2 + c_p(T_e-L)^2))^2 = (c_hT^2\lambda^2 - 2c_pL\mu\lambda T + c_pL^2\mu^2)^2 = (2K\mu - 2c_pL\mu\lambda T + 2c_pL^2\mu^2)^2 < (2K\mu)^2$. As a result, $8Kc_p\lambda^2T^2\mu - \left(c_p\lambda^2T^2 - \mu^2\left(c_hT_e{}^2 + c_p(T_e-L)^2\right)\right)^2 > 8Kc_p\lambda^2T^2\mu - (2K\mu)^2 > 0$, which proves $F_2{}^1 > 0$. Likewise, $F_2 = \frac{(c_h+c_p)\mu}{T} * F_2{}^1 - \left(-\frac{(c_p(T_e-L)+c_hT_e)\mu}{T^2}\right) * F_2{}^2 + \frac{c_p(T_e-L)+c_hT_e}{T} * F_2{}^3 > 0$. Therefore, $\nabla^2L_2(\bar{\boldsymbol{y}}^2) > 0$ and $\bar{\boldsymbol{y}}^2$ is a strict local minimum.

In scenario 3, $\nabla^2L_2(\bar{\boldsymbol{y}}^3)$ is exactly the same with $\nabla^2L_2(\bar{\boldsymbol{y}}^1)$ by replacing μ with λ. Therefore, $\nabla^2L_2(\bar{\boldsymbol{y}}^3) > 0$ and $\bar{\boldsymbol{y}}^3$ is a strict local minimum.

In scenario 4, $\nabla^2L_2(\bar{\boldsymbol{y}}^4)$ is exactly the same with $\nabla^2L_2(\bar{\boldsymbol{y}}^2)$ by replacing μ with λ. Therefore, $\nabla^2L_2(\bar{\boldsymbol{y}}^4) > 0$ and $\bar{\boldsymbol{y}}^4$ is a strict local minimum.

We have demonstrated above that there is only one unique KKT point and that this point is a strict local minimum. Among the three variables, $T_e \in (0, \frac{\lambda T}{\mu})$, $\mu \in [\lambda, +\infty]$ and $T \in (0, +\infty]$. When T is infinite, $\lim_{T\to\infty} \frac{c_p(\lambda T-\mu(T_e-L))^2}{2T\mu} = \infty$ and therefore $\lim_{T\to\infty} TC_2(T_e, \mu, T) = \infty$.

Likewise, when μ is infinite, $\lim_{\mu\to\infty} c_\mu(\mu)\lambda = \infty$ and therefore $\lim_{\mu\to\infty} TC_2(T_e, \mu, T) = \infty$. As a result, the optimum must be either an interior point or on the boundary, which are listed in Table 9. Therefore, the solution in Table 9 is indeed the global minimum.

B.4 KKT necessary conditions for problem (P3)

Define the Lagrangian function

$$L_3(T_e, T, \mu, v_1, v_2, v_3, v_4) = TC_3(T_e, T, \mu) + v_1(\lambda T - \mu T_e) + v_2(\mu(T_e - L) - \lambda T) + v_3(\lambda - \mu) + v_4(L - T_e),$$

where v_i for $i = 1, 2, 3, 4$ are Lagrange multipliers. The KKT conditions require that

$$\frac{dTC_3}{dT_e} - v_1\mu + v_2\mu - v_4 = 0 \text{ or } c_h\lambda - \frac{c_p(\lambda T - \mu(T_e - L))}{T} - v_1\mu + v_2\mu - v_4 = 0,$$

$$\frac{dTC_3}{dT} + v_1\lambda - v_2\lambda = 0 \text{ or } -\frac{c_h\lambda^2}{2\mu} + \frac{c_p(\lambda^2T^2 - \mu^2(T_e - L)^2)}{2T^2\mu} - \frac{K}{T^2} + v_1\lambda - v_2\lambda = 0,$$

$$\frac{dTC_3}{d\mu} - v_1T_e + v_2(T_e - L) - v_3 = 0 \text{ or}$$

$$\frac{c_h\lambda^2T}{2\mu^2} - \frac{c_p(\lambda^2T^2 - \mu^2(T_e - L)^2)}{2T\mu^2} + \lambda c'_\mu(\mu) - v_1T_e + v_2(T_e - L) - v_3 = 0,$$

$$v_1(\lambda T - \mu T_e) = 0,$$

$$v_2(\mu(T_e - L) - \lambda T) = 0,$$

$$v_3(\lambda - \mu) = 0,$$

$$v_4(L - T_e) < 0,$$

$$v_i = 0, i = 1, 2, 3, 4. \quad (4.12)$$

As discussed in Problem (P2), $v_4 = 0$ has to hold to ensure a non-zero service level. Note that if $v_1 > 0$, $v_2 = 0$ has to hold for any feasible solution. When $v_1 > 0$, then $\lambda T = \mu T_e$, and problem (P3) in case 2 is exactly equivalent to problem (P2) in case 1 with the same objective function and constraints. To avoid repetition, we only focus on the optimal results when $\lambda T < \mu T_e$ with $v_1 = 0$. Therefore, the KKT point satisfying Eqs. (4.12) can be written as

$$T_e = \frac{(c_p - c_h)\lambda T - v_2T\mu}{c_p\mu} + L,$$

$$T = \sqrt{\frac{2K\mu + c_p(T_e - L)^2\mu^2}{(c_p - c_h)\lambda^2 - 2v_2\lambda\mu}},$$

$$\mu = \sqrt{\frac{(c_p - c_h)\lambda^2T^2}{c_p(T_e - L)^2 + 2c'_\mu(\mu)\lambda T - 2T(v_2(T_e - L) - v_3)}}. \quad (4.13)$$

The exact expression for the KKT point in Eqs. (4.13) depends on the multipliers.

Scenario 1: $v_2 = v_3 = 0$.

In this scenario, none of the constraints are binding and the KKT point for Eqs. (4.13) is

$$T_e = \frac{(c_p - c_h)\lambda T}{c_p \mu} + L, T = \sqrt{\frac{2K\mu + c_p(T_e - L)^2\mu^2}{(c_p - c_h)\lambda^2}}, \mu = \sqrt{\frac{(c_p - c_h)\lambda^2 T^2}{c_p(T_e - L)^2 + 2c'_\mu(\mu)\lambda T}}.$$

Replacing $T_e - L$ with $\frac{(c_p - c_h)\lambda T}{c_p \mu}$ leads to $T = \sqrt{\frac{2K\mu c_p}{c_h(c_p - c_h)\lambda^2}}$ and $\mu = \sqrt{\frac{c_h(c_p - c_h)\lambda T}{2c_p c'_\mu(\mu)}}$, which generates the exact expression for the two decision variables: $T = \frac{1}{\lambda}\left(\frac{2K^2 c_p}{c_h c'_\mu(\mu)(c_p - c_h)}\right)^{\frac{1}{3}}$, $\mu = \left(\frac{Kc_h(c_p - c_h)}{2c_p c'_\mu(\mu)^2}\right)^{\frac{1}{3}}$. Therefore, the KKT point is

$$\bar{\boldsymbol{z}}^1 = \left(\left(\frac{4Kc'_\mu(\mu)(c_p - c_h)}{c_p {c_h}^2}\right)^{\frac{1}{3}} + L, \frac{1}{\lambda}\left(\frac{2K^2 c_p}{c_h c'_\mu(\mu)(c_p - c_h)}\right)^{\frac{1}{3}}, \left(\frac{Kc_h(c_p - c_h)}{2c_p c'_\mu(\mu)^2}\right)^{\frac{1}{3}}, u_1 = u_2 = u_3 = u_4 = 0\right)^t.$$

The corresponding total cost $TC_3(T_e, T, \mu) = \lambda\left(\frac{4Kc_h c'_\mu(\mu)(c_p - c_h)}{c_p}\right)^{\frac{1}{3}} + c_h\lambda L + c_\mu(\mu)\lambda$.

Scenario 2: $v_2 > 0, v_3 = 0$.

In this scenario, $\mu(T_e - L) = \lambda T$, which means the retailer fulfils every order by the deadline and therefore there are no late orders. $\frac{dTC_3}{dT_e} - v_1\mu + v_2\mu - v_4 = c_h\lambda + v_2\mu = 0$ implies that $v_2 < 0$. Therefore, this scenario cannot occur.

Scenario 3: $v_2 = 0, v_3 > 0$.

In this scenario, $\lambda = \mu$ and KKT point is

$$\bar{\boldsymbol{z}}^2 = \left(\sqrt{\frac{2K(c_p - c_h)}{c_h c_p \lambda}} + L, \sqrt{\frac{2Kc_p}{c_h(c_p - c_h)\lambda}}, \lambda, u_1 = u_2 = u_4 = 0, u_3 = \lambda c'_\mu(\mu) - \sqrt{\frac{Kc_h(c_p - c_h)}{2c_p\lambda}}\right)^t.$$

The corresponding total cost $TC_3(T_e, T, \mu) = \sqrt{\frac{2Kc_h(c_p - c_h)\lambda}{c_p}} + c_h\lambda L + c_\mu(\lambda)\lambda$.

Scenario 4: $v_2 = 0, v_3 > 0$.

In this scenario, $\mu(T_e - L) = \lambda T$ and $\lambda = \mu$. Likewise, $\frac{dTC_3}{dT_e} - v_1\mu + v_2\mu - v_4 = c_h\lambda + v_2\lambda = 0$, which gives $v_2 < 0$. Hence, this scenario cannot occur.

Therefore, the KKT point for Problem (P3) can be summarised in Table 10.

The constraint qualification for Problem (P3)

Let $\nabla g_3(i)$ denote the gradient vector for decision variables T_e, T, μ in the i[th] constraint of problem (P3). Therefore,

$$\nabla g_3(1) = (-\mu, \lambda, -T_e)^t, \quad \nabla g_3(2) = (\mu, -\lambda, T_e - L)^t,$$
$$\nabla g_3(3) = (0, 0, -1)^t, \qquad \nabla g_3(4) = (-1, 0, 0)^t.$$

Note that $\lambda > 0$ and in the feasible region, $\mu > 0$ and $T_e \geq L > 0$. Therefore, $T_e \neq T_e - L$ and $\nabla g_3(1)$ and $\nabla g_3(2)$ are linearly independent. In fact, the constraint gradient

vectors $\nabla g_3(1)$, $\nabla g_3(2)$, $\nabla g_3(3)$ and $\nabla g_3(4)$ are linearly independent for any feasible solution. As a result, KKT conditions hold true at optimality.

KKT sufficient conditions for Problem (P3)

Let F_3, G_{3i} be the Hessians of the cost function $TC_3(T_e, T, \mu)$ and the i[th] constraint in Problem (P3) respectively. The Hessian for the Lagrangian function at a feasible solution $\bar{\boldsymbol{z}}$ is $\nabla^2 L_3(\bar{\boldsymbol{z}}) = F_3(\bar{\boldsymbol{z}}) + \sum_{i=1}^{4} v_i G_{3i}(\bar{\boldsymbol{z}})$.

Clearly, $G_{31} = G_{32} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} = 0$, $G_{33} = G_{34} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = 0$. Therefore, $\nabla^2 L_3(\bar{\boldsymbol{z}}) = F_3(\bar{\boldsymbol{z}})$. For the total cost $TC_3(T_e, T, \mu) = c_h(\lambda T_e - \frac{\lambda^2 T}{2\mu}) + c_p \frac{(\lambda T - \mu(T_e - L))^2}{2T\mu} + \frac{K}{T} + c_\mu(\mu)\lambda$, the second partial derivatives are:

$$\frac{d^2 TC_3}{dT_e{}^2} = \frac{c_p \mu}{T}; \frac{d^2 TC_3}{dT^2} = \frac{2K + c_p(T_e - L)^2 \mu}{T^3}; \frac{d^2 TC_3}{d\mu^2} = \frac{(c_p - c_h)\lambda^2 T}{\mu^3} + c''_\mu(\mu)\lambda;$$

$$\frac{d^2 TC_3}{dT_e dT} = -\frac{c_p(T_e - L)\mu}{T^2}; \frac{d^2 TC_3}{dT_e d\mu} = \frac{c_p(T_e - L)}{T}; \frac{d^2 TC_3}{dT d\mu} = -\frac{(c_p - c_h)\lambda^2 T^2 + c_p(T_e - L)^2 \mu^2}{2T^2 \mu^2}.$$

As a result, the Hessian matrix of $TC_3(T_e, T, \mu)$ is:

$$F_3 = \begin{bmatrix} \frac{c_p \mu}{T} & -\frac{c_p(T_e - L)\mu}{T^2} & \frac{c_p(T_e - L)}{T} \\ -\frac{c_p(T_e - L)\mu}{T^2} & \frac{2K + c_p(T_e - L)^2)\mu}{T^3} & -\frac{(c_p - c_h)\lambda^2 T^2 + c_p(T_e - L)^2 \mu^2}{2T^2 \mu^2} \\ \frac{c_p(T_e - L)}{T} & -\frac{(c_p - c_h)\lambda^2 T^2 + c_p(T_e - L)^2 \mu^2}{2T^2 \mu^2} & \frac{(c_p - c_h)\lambda^2 T}{\mu^3} + c''_\mu(\mu)\lambda \end{bmatrix}.$$

Clearly, the first principal minors $\frac{c_p \mu}{T}$, $\frac{2K + c_p(T_e - L)^2)\mu}{T^3}$ and $\frac{(c_p - c_h)\lambda^2 T}{\mu^3} + c''_\mu(\mu)\lambda$ are positive. With$c''_\mu(\mu) \geq 0$, the second principal minors are:

$$F_3{}^1 = \begin{bmatrix} \frac{2K + c_p(T_e - L)^2)\mu}{T^3} & -\frac{(c_p - c_h)\lambda^2 T^2 + c_p(T_e - L)^2 \mu^2}{2T^2 \mu^2} \\ -\frac{(c_p - c_h)\lambda^2 T^2 + c_p(T_e - L)^2 \mu^2}{2T^2 \mu^2} & \frac{(c_p - c_h)\lambda^2 T}{\mu^3} + c''_\mu(\mu)\lambda \end{bmatrix} \geq$$

$\frac{8K(c_p - c_h)\lambda^2 T^2 \mu - ((c_p - c_h)\lambda^2 T^2 - c_p(T_e - L)^2 \mu^2)^2}{4T^4 \mu^4}$.

$$F_3{}^2 = \begin{bmatrix} \frac{c_p \mu}{T} & \frac{c_p(T_e - L)}{T} \\ \frac{c_p(T_e - L)}{T} & \frac{(c_p - c_h)\lambda^2 T}{\mu^3} + c''_\mu(\mu)\lambda \end{bmatrix} \geq \frac{c_p((c_p - c_h)\lambda^2 T^2 - c_p(T_e - L)^2 \mu^2)}{T^2 \mu^2}.$$

$$F_3{}^3 = \begin{bmatrix} \frac{c_p \mu}{T} & -\frac{c_p(T_e - L)\mu}{T^2} \\ -\frac{c_p(T_e - L)\mu}{T^2} & \frac{2K + c_p(T_e - L)^2)\mu}{T^3} \end{bmatrix} = \frac{2K c_p \mu}{T^4} > 0.$$

At the KKT point $\bar{\boldsymbol{z}}^1$ in scenario 1, $T = \sqrt{\frac{2K\mu + c_p(T_e - L)^2 \mu^2}{(c_p - c_h)\lambda^2}}$ and then $(c_p - c_h)\lambda^2 T^2 - c_p(T_e - L)^2 \mu^2 = 2K\mu$. As a result, $8K(c_p - c_h)\lambda^2 T^2 \mu - \left((c_p - c_h)\lambda^2 T^2 - c_p(T_e -\right.$

$L)^2\mu^2\Big)^2 = 4K\mu\big(2\lambda^2T^2\big(c_p - c_h\big) - K\mu\big) > 0$ and therefore ${F_3}^1 > 0$. Meanwhile, $\mu = \sqrt{\frac{(c_p-c_h)\lambda^2T^2}{c_p(T_e-L)^2+2c'_\mu(\mu)\lambda T}}$. Therefore, $\big(c_p - c_h\big)\lambda^2T^2 - c_p(T_e - L)^2\mu^2 = 2c'_\mu(\mu)\lambda T\mu^2$ and ${F_3}^2 > 0$. The third principal minor $F_3 = \frac{c_p\mu}{T} * {F_3}^1 - \Big(-\frac{c_p(T_e-L)\mu}{T^2}\Big) * {F_3}^2 + \frac{c_p(T_e-L)}{T} * {F_3}^3 > 0$. Therefore, $\nabla^2L_3(\bar{\boldsymbol{z}}^1) > 0$ and $\bar{\boldsymbol{z}}^1$ is a strict local minimum.

In scenario 4, $\nabla^2L_3(\bar{\boldsymbol{z}}^2)$ is exactly the same with $\nabla^2L_3(\bar{\boldsymbol{z}}^1)$ by replacing μ with λ. Therefore, $\nabla^2L_3(\bar{\boldsymbol{z}}^2) > 0$ and $\bar{\boldsymbol{z}}^2$ is a strict local minimum.

We have shown that there is only one unique KKT point and that this point is a strict local minimum. $T_e \in (\frac{\lambda T}{\mu}, \frac{\lambda T}{\mu} + L]$, $\mu \in [\lambda, +\infty]$ and $T \in (0, +\infty]$. When T is infinite, $\lim_{T\to\infty} c_h(\lambda T_e - \frac{\lambda^2 T}{2\mu}) + c_p\frac{(\lambda T-\mu(T_e-L))^2}{2T\mu} = \infty$ and therefore $\lim_{T\to\infty} TC_3(T_e, \mu, T) = \infty$. Likewise, when μ is infinite, $\lim_{\mu\to\infty} c_\mu(\mu)\lambda = \infty$ and therefore $\lim_{\mu\to\infty} TC_3(T_e, \mu, T) = \infty$. As a result, the optimum must be either an interior point or on the boundary, which are listed in Table 10. Therefore, the solution in Table 10 is indeed the global minimum.

B.5 Parameter estimation for other retailers

Estimation of λ. Sainsbury's represent 13.7% of the online groceries sales in the UK (Mintel, 2019) and its annual online sales is estimated at £1722 million. The AOV in Sainsbury's is assumed to be the same with that in Tesco (i.e., £100). The BOPS percentage in Sainsbury's total online sales is estimated at 23%, as indicated in Mintel (2019). The daily BOPS order rate in Sainsbury's is estimated at 10851 ($\frac{£12.3bn*14\%*23\%}{£100*365} \approx 10851$).

M&S takes up approximately 7.6% of the fashion market, which was worth £32 billion to the UK economy in 2017 according to the British Fashion Council (BFC) (GlobalData, 2018a; British Fashion Council, 2018). The annual sales in M&S Clothing section are estimated at £2.432 billion ($£32\ bn * 7.6\%$). In M&S, 22% of overall sales come from online shoppers (Marks and Spencer, 2018a) and 71% of online customers opt for in-store collection (BBC News, 2019b). The AOV in M&S clothing is assumed to be £70, as the AOV for its online rival ASOS is £73 (ASOS, 2018). Thus, the daily BOPS orders received in M&S clothing can be estimated using the same approach. The annual online sales from Tu clothing are estimated at £100 million and 80% of these online orders are collected in stores (The Telegraph, 2015). The AOV in Tu Clothing is also assumed to be £70 and then the daily BOPS orders can be estimated.

The annual revenue of Habitat is approximately £32.5 million in 2018 (http://ecommercedb.com) and over 60% of overall sales were made online (www.about.sainsburys.co.uk). Therefore, the annual online sales in Habitat are estimated at £19.5 million (£32.5×60%). BOPS orders from Habitat can be collected from over 2000 Sainsbury's locations (www.about.sainsburys.co.uk), which indicates the popularity of its BOPS services. We assume that 30% of online orders are collected in stores, which is higher than the industrial benchmark 12% (GlobalData, 2018b). The AOV in Habitat is assumed to be £30, which is consistent with the figure in Dunelm, another homeware retailer (Financial Times, 2016c).

The total sales from books and stationery in WH Smith is £590 million in 2018 (WH Smith, 2018). Because the average internet percentage of total retail sales is 16.3% (ONS, 2018), it is reasonable to assume that approximately 10% of total sales for WH Smith books and stationery comes from online shoppers. As a result, the annual online sales in WH Smith books and stationery are estimated at £59 million. We use the industrial benchmark 12% as the BOPS percentage of total online in WH Smith. The AOV is £15 considering the low prices for its products.

Estimation of *L*. Sainsbury's offers BOPS service at 124 stores (https://help.sainsburys.co.uk). It has been picking online orders in stores and also operates one online fulfilment center in Bromley-by-Bow (East London) for the growing online demand in London (Sainsbury's, 2016). It is reasonable to assume that all BOPS orders are processed in these 124 stores selected for BOPS service offerings, which implies that $L = 0$. Sainsbury's fulfils online Tu clothing orders from its distribution center in Bedford (Financial times, 2014) and deploys 2 national distribution centers for slower moving goods (Sainsbury's, 2008). Boots has consolidated its warehouses into one large store service center (SSC) in Nottingham for UK businesses (NottinghamshireLive, 2017). The Clothing & Home distribution network in M&S comprises of 19 DCs and warehouses (Marks and Spencer, 2018b) and WH Smith has 3 warehouses in the UK for its online and offline businesses (WH Smith, 2018).

The number of collection points for Tu clothing, M&S clothing, Boots, Habitat and WH Smith can be found on their official shopping websites. Clearly, the STS transportation window L is related to the collection speed advertised and the ratio between the number of picking facilities and the number of collection points. Therefore, we assume that L for Tu clothing, M&S clothing, Boots, Habitat and WH Smith are 0.75, 0.5, 0.5, 3 and 3 days respectively.

Estimation of c_h. We assume the unit holding cost in Sainsbury's is the same with that in Tesco. For other retailers, c_h is computed by multiplying the AOV with an annual rate of 30%. For example, In M&S clothing, $c_h = \frac{£70*30\%}{365} \approx £0.058$ per order per day. The unit lateness penalty cost is calculated for an *NSD* service level of 98%. For Sainsbury's we use the same approach as explained for Tesco. For the other retailers, from Table 10, we find that $NSD = (c_p\text{-}c_h) / c_p$.

Estimation of $c_\mu(\mu)$, $c_\mu'(\mu)$ ***and*** K. Recall that $c_\mu(\mu) = c_t + c_0(1+\alpha)^{\frac{\mu}{\mu_0}-1}$. We use the same c_0, μ_0 and α for Sainsbury's as we did for Tesco. For other retailers, orders are processed in warehouses or DCs without in-store shoppers. It is reasonable to assume that α is very low (we use $= 1.5\%$) due to picking automation. c_0 is also smaller for other retailers because there are only a few items per order and the picking efficiency is much higher in warehouses and DCs. We assume that $c_0 = £0.25$ for all other retailers and we have set $\mu_0 = 480$ (about 10% lower than the rate Habitat would need to pick its orders in 0.5 days). The transportation cost was set by taking into account the STS transportation window and the level of picking decentralisation. For Habitat and WH Smith, both with $L = 3$ days, we assume $c_t = £0.25$. Also for M&S clothing with $L = 0.5$ days and 19 warehouses, we have set $c_t = £0.25$. For TU clothing with centralised picking and $L = 0.75$ days, we have taken $c_t = £0.5$. Boots is well-known for its fast distribution network and we assume $L = 0.5$ days and $c_t = £0.6$.

The fixed picking cost K in Sainsbury's is assumed to be the same with that in Tesco. For other retailers, BOPS orders are processed with store replenishment and/or home delivery orders in warehouses and DCs and therefore the average fixed picking cost will be low. We assume that K is £50 for Habitat and WH Smith (warehouses with smaller order volumes) and £100 for Tu clothing, M&S clothing and Boots (warehouses with higher order volumes).

Bibliography

Abdulkader, M.M.S., Gajpal, Y., ElMekkawy, T.Y., 2018. Vehicle routing problem in omni-channel retailing distribution systems. *International Journal of Production Economics*. 196, 43–55.

Acimovic, J., Graves, S.C., 2015. Making better fulfilment decisions on the fly in an online retail environment. *Manufacutring and Serving Operations Management*, 17(1), 34-51.

Agatz, N., Campbell, A., Fleischmann, M., Savelsbergh, M., 2011. Time Slot Management in Attended Home Delivery. *Transportation Science*, 45(3), 435–449.

Agatz, N., Fleischmann, M., van Nunen, J.A.E.E., 2008. E-fulfilment and multi-channel distribution - A review. *European Journal of Operational Research*, 187(2), 339–356.

Aksen, D., Altinkemer, K., 2008. A location-routing problem for the conversion to the "click-and-mortar" retailing: The static case. *European Journal of Operational Research*, 186(2), 554–575.

Akinc, U., Khumawala, B., 1997. An efficient branch and bound algorithm for the capacitated warehouse location problem. *Management Science*, 23 (6), 585–594.

Akturk, M.S., Ketzenberg, M., Heim, G.R., 2018. Assessing impacts of introducing ship-to-store service on sales and returns in omnichannel retailing: A data analytics study. *Journal of Operations Management*, 61, 15–45.

Alawneh, F., Zhang, G., 2018. Dual-channel warehouse and inventory management with stochastic demand. *Transportation Research Part E: Logistics and Transportation Review*. 112(2017), 84–106.

Alishah, J.E., Moinzadeh, K., Zhou, Y.-P., 2015. Inventory Fulfilment Strategies for an Omni-Channel Retailer. *SSRN Electronic Journal*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2659671 (Accessed: 2 May 2017).

Alptekinoğlu, A., Tang, C.S., 2005. A model for analyzing multi-channel distribution systems. *European Journal of Operational Research*, 163(3), 802–824.

Ambrosino, D., Scutella, M. G., 2005. Distribution network design: New problems and related models. *European Journal of Operational Research*, 165(2005), 610-624.

Ansari, S., Basdere, M., Li, X., Ouyang, Y., Smilowitz, K., 2018. Advancements in continuous approximation models for logistics and transportation systems: 1996–2016. *Transportation Research Part B: Methodological*, 107(2018), 229–252.

Arabani, A. B., Farahani, R. Z., 2012. Facility location dynamics: An overview of classifications and applications. *Computers & Industrial Engineering*, 62(1), 408–420.

Arampantzi, C., Minis, I., 2017. A new model for designing sustainable supply chain networks and its application to a global manufacturer. *Journal of Cleaner Production*, 156, 276–292.

Ardjmand, E., Bajgiran, O.S., Rahman, S., Weckman, G.R., Young II, W.A., 2018. A multi-objective model for order cartonization and fulfilment center assignment in the e-tail/retail industry. *Transportation Research Part E: Logistics and Transportation Review*, 115(2018), 16–34.

Argos.co.uk. *Delivery & Collection*. Available at: https://www.argos.co.uk/help/delivery-and-collection/ (Accessed: 30 October 2018).

Asdecker, B., Karl, D., Sucky, E., 2017. Examining drivers of consumer returns in e-tailing with real shop data, *in proceedings of the 50th Howaii International Conference on System Science (HICSS 2017)*, Hawaii.

Asdemir, K., Jacob, V.S., Krishnan, R., 2009. Dynamic pricing of multiple home delivery options. *European Journal of Operational Research*. 196(1), 246–257.

Ashayeri, J., Rongen, J. M. J., 1997. Central distribution in Europe: a multi-criteria approach to location selection. *International Journal of Logistics Management*, 18(1), 97–109.

Askin, R. G., Baffo, I., Xia, M., 2014. Multi-commodity warehouse location and distribution planning with inventory consideration. *International Journal of Production Research*, 52(7), 1897–1910.

ASOS. 2018. *ASOS Annual Report 2018*.

Avery, J., Steenburgh, T.J., Deighton, J.A., Caravella, M., 2012. Adding Bricks to Clicks: The Contingencies Driving Cannibalization and Complementarity in Multichannel Retailing. *Journal of Marketing*, 76(3), 96–111.

Ayanso, A., Diaby, M., Nair, S.K., 2006. Inventory rationing via drop-shipping in Internet retailing: A sensitivity analysis. *European Journal of Operational Research*, 171(1), 135–152.

Bakker, M., Riezebos, J., Teunter, R.H., 2012. Review of inventory systems with deterioration since 2001. *European Journal of Operational Research*, 221(2), 275–284.

Bailey, J.P., Rabinovich, E., 2005. Internet book retailing and supply chain management: An analytical study of inventory location speculation and postponement. *Transportation Research Part E: Logistics and Transportation Review*, 41(3), 159–177.

Ballou, R. H., 2001. Unresolved Issues in Supply Chain Network Design. *Information Systems Frontiers*, 3(4), 417–426.

Barilliance, 2019. *[2019 Update] Click and Collect Statistics, Challenges and Strategies*. Available at: https://www.barilliance.com/click-collect/ (Accessed: 30 October 2019).

Bayram, A., Cesaret, B., 2017. Ship-from-Store Operations in Omni-Channel Retailing. *Proceedings of the 2017 Industrial and Systems Enineering Conference*, Norcross, 1181–1186.

Bazaraa, M.S., Sherali, H.D., Shetty, C.M., 2006. *Nonlinear Programming: Theory and*

Algorithms, 3rd edition, Wiley.

BBC News, 2018. *House of Fraser bought by Sports Direct for £90m.* Available at: https://www.bbc.co.uk/news/business-45140874 (Accessed: 30 October 2019).

BBC News, 2019a. *The M&S stores closing their doors for the final time.* Available at: https://www.bbc.co.uk/news/business-48137882 (Accessed: 30 October 2019).

BBC News, 2019b. *High Street crisis: Can "click and collect" save shop*s? Available at: https://www.bbc.co.uk/news/business-46960792 (Accessed: 30 October 2019).

BBC News, 2020. Ikea announces first big UK store closure. Available at: https://www.bbc.co.uk/news/business-51369413 (Accessed: 7 February 2019).

Beardwood, J., Halton, J.H., Hammersley, J.M., 1959. The shortest path through many points, *in proceedings of the Cambridge Philosophical Society*, 55, 299-327.

Becerril-Arreola, R., Leng, M., Parlar, M., 2013. Online retailers' promotional pricing, free-shipping threshold, and inventory decisions: A simulation-based analysis. *European Journal of Operational Research*, 230(2), 272–283.

Beck, N., Rygl, D., 2015. Categorization of multiple channel retailing in Multi-, Cross-, and Omni-Channel Retailing for retailers and retailing. *Journal of Retailing and Consumer Services*, 27, 170–178.

Bendoly, E., Blocher, D., Bretthauer, K.M., Venkataramanan, M.A., 2007. Service and cost benefits through clicks-and-mortar integration: Implications for the centralization/decentralization debate. *European Journal of Operational Research*, 180(1), 426–442.

Berman, O., Simchi-Levi, D., 1990. Conditional Location Problems on Networks. *Transportation Science*, 24(1), 77-78.

Bernon, M., Cullen, J., Gorst, J., 2016. Online retail returns management: integration within an omni-channel distribution context. *International Journal of Physical Distribution & Logistics Management*, 46(6/7), 584–605.

Berry, L. L., Bolton, R. N., Bridges, C. H., Meyer, J., Parasuraman, A., Seiders, K., 2010. Opportunities for innovation in the delivery of interactive retail services. *Journal of Interactive Marketing*, 24(2), 155–167.

Bhatnagar, A., Syam, S.S., 2014. Allocating a hybrid retailer's assortment across retail stores: Bricks-and-mortar vs online. *Journal of Business Research*, 67(6), 1293–1302.

Bilir, C., Ekici, S. O., Ulengin, F., 2017. An integrated multi-objective supply chain network and competitive facility location model. *Computer & Industrial Engineering*, 108, 136–148.

Bloomberg, 2017. *Online Retailers Are Desperate to Stem a Surging Tide of Returns*. Available at: https://www.bloomberg.com/news/articles/2017-11-03/online-retailers-are-desperate-to-stem-a-surging-tide-of-returns (Accessed: 28 December 2018).

Bloomberg, 2018. *What drew Amazon and Alibaba to Bricks-and-Mortar: Q&A*. Available at: https://www.bloomberg.com/news/articles/2018-01-30/what-drew-amazon-and-alibaba-to-brick-and-mortar-quicktake-q-a (Accessed: 28 December 2019).

Bloomberg, 2019. *Walmart is beating Amazon in a business worth $35 billion*. Available at: https://www.bloomberg.com/opinion/articles/2019-04-01/amazon-risks-missing-out-on-35-billion-click-and-collect-market (Accessed: 30 October 2019).

Bower, A.B., Maxham III, J.G., 2012. Return shipping policies of online retailers: normative assumptions and the long-term consequences of fee and free returns. *Journal of Marketing*, 76, 110-124.

Boyaci, T., 2005. Competitive stocking and coordination in a multiple-channel distribution system. *IIE Transactions*, 37(5), 407–427.

Boyaci, T., Ray, S., 2003. Product Differentiation and Capacity Cost Interaction in Time and Price Sensitive Markets. *Manufacturing & Service Operations Management*, 5(1), 18–36.

Boyer, K.K., Hult, G.T.M., Frohlich, M., 2003. An exploratory analysis of extended grocery supply chain operations and home delivery. *Integrated Manufacturing Systems*, 14(8), 652–663.

Boyer, K.K., Hult, G.T.M., 2006. Customer behavioral intentions for online purchases: an examination of fulfilment method and customer experience level. *Journal of Operations Management*, 24(2), 124-147.

Boyer, K.K., Prud'homme, A.M., Chung, W., 2009. The last mile challenge: evaluating the effects of customer density and delivery window patterns. *Journal of Business Logistics*, 30(1), 185-201.

Bretthauer, K.M., Mahar, S., Venakataramanan, M.A., 2010. Inventory and distribution strategies for retail/e-tail organisations. *Computers and Industrial Engineering*, 58(1), 119–132.

Brimberg, J., Hansen, P., Mladenovic, N., Salhi, S., 2008. A Survey of Solution Methods for the Continuous Location-Allocation Problem. *International Journal of Operations Research*, 5(1), 1–12.

British Fashion Council, 2018. *London Fashion Week September 2018 Facts and Figures*. Available at: https://www.britishfashioncouncil.co.uk/pressreleases/London-Fashion-Week-September-2018-Facts-and-Figures (Accessed: 30 October 2019).

Brynjolfsson, E., Hu, Y.J., Simester, D., 2011. Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales. *Management Science*, 57(8), 1373–1386.

Brynjolfsson, E., Hu, Y.J., Rahman, M.S., 2013. Competing in the Age of Omnichannel Retailing. *MIT Sloan Management Review*, 54(4), 23–29.

Burke, E.K., De Causmaecker, P., Berghe, G.V., Van Landeghem, H., 2004. The State of the art of nursing rostering. *Journal of Scheduling*, 7(6), 441-499.

Business Insider, 2017. *Walmart built a giant vending machine that retrieves groceries*. Available at: https://www.businessinsider.com/walmart-built-a-giant-vending-machine-that-retrieves-groceries-2017-6?r=US&IR=T (Accessed: 30 October 2019).

Campbell, A.M., Savelsbergh, M., 2005. Decision Support for Consumer Direct Grocery Initiatives. *Transportation Science*, 39(3), 313–327.

Campbell, A.M., Savelsbergh, M., 2006. Incentive schemes for attended home delivery services. *Transportation Science*, 40(3), 327–341.

Campbell, J. F., 1993. One-to-many distribution with transshipments: an analytic model. *Transportation Science*, 27(4), 330–340.

Cao, J., So, K.C., Yin, S., 2016. Impact of an "online-to-store" channel on demand allocation, pricing and profitability. *European Journal of Operational Research*, 248(1), 234–245.

Cao, L., 2014. Business Model Transformation in Moving to a Cross-Channel Retail Strategy: A Case Study. *International Journal of Electronic Commerce*, 18(4), 69–96.

Cao, L., Li, L., 2015. The Impact of Cross-Channel Integration on Retailers' Sales Growth. *Journal of Retailing*, 91(2), 198–216.

Cattani, K.D., Souza, G.C., 2002. Inventory Rationing and Shipment Flexibility Alternatives for Direct Market Firms. *Production and Operations Management*, 11(4), 441–457.

Cavallo, A., 2017. Are online and offlien prices similar? Evidence from large multi-channel retailers. *American Economic Review*, 107(1), 283-303.

Çeven, E., Gue, K.R., 2017. Optimal Wave Release Times for Order Fulfilment Systems with Deadlines. *Transportation Science*, 51(1), 52–66.

Chapados, N., Joliveau, M., L'Ecuyer, P., Rousseau, L.M., 2014. Retail store scheduling for profit. *European Journal of Operational Research*, 239(3), 609-624.

Chauhan, D., Unnikrishnan, A., Figliozzi, M., 2019. Maximum coverage capacitated facility location problem with range constrained drones. *Transportation Research Part C: Emerging Technologies*, 99, 1–18.

Chen, B., Chen, J., 2017. When to introduce an online channel, and offer money back guarantees and personalized pricing? *European Journal of Operational Research*, 257(2), 614–624.

Chen, J., Chen, Y., Parlar, M., Xiao, Y., 2011. Optimal inventory and admission policies for drop-shipping retailers serving in-store and online customers. *IIE Transactions*, 43(5), 332–347.

Chen, X., Liu, Y., Wan, Z., 2016. Optimal decision making for online and offline retailers

under BOPS mode. *The ANZIAM Journal*, 58(2), 187-208.

Chiang, W.K., Feng, Y., 2010. Retailer or e-tailer? Strategic pricing and economic-lot-size decisions in a competitive supply chain with drop-shipping. *Journal of the Operational Research Society*, 61(11), 1645–1653.

Chiang, W.K., Monahan, G.E., 2005. Managing inventories in a two-echelon dual-channel supply chain. *European Journal of Operational Research*, 162(2), 325–341.

Choi, T. M., Chen, Y., Chung, S. H., 2019. Online-offline fashion franchising supply chains without channel conflicts: Choices on postponement and contracts. *International Journal of Production Economics*, 215, 174-184.

Chopra, S., 2016. How omni-channel can be the future of retailing. *Decision*, 43(2), 135–144.

Chopra, S., Meindl, P., 2016. *Supply chain management : strategy, planning, and operation*. 6th edn. Boston: Pearson.

Clark, A.J., 1972. An informal survey of multi-echelon inventory theory. *Naval Research Logistics Quarterly*, 19(4), 621–650.

Clarks. *Delivery & Collection*. Available at: https://www.clarks.co.uk/delivery-and-collection (Accessed: 30 October 2019).

CNN, 2019. *Retail's Amazon antidote: Buy online, pickup in store*. Available at: https://edition.cnn.com/2018/12/29/business/walmart-target-best-buy-amazon-online-pickup/index.html (Accessed: 30 October 2019).

CollectPlus. *Easy returns to over 300 retailers*. Available at: https://www.collectplus.co.uk/retailers (Accessed: 30 October 2019)

Co-op, 2019. *Co-op brings home delivery closer to shoppers with launch of online service*. Available at: https://www.co-operative.coop/media/news-releases/co-op-brings-home-delivery-closer-to-shoppers-with-launch-of-online-service (Accessed: 6 January 2020).

Cui, T., Ouyang, Y., Shen, Z-J.M., 2010. Reliable facility location design under the risk of disruptioins. *Operations Research*, 58, 998-1011.

Cummins, S., Peltier, J. W., Dixon, A., 2016. Omni-channel research framework in the context of personal selling and sales management. *Journal of Research in Interactive Marketing*, 10(1), 2–16.

Current, J., Ratick, S., ReVelle, C., 1998. Dynamic facility location when the total number of facilities is uncertain: A decision analysis approach. *European Journal of Operational Research*, 110(3), 597–609.

Daganzo, C.F., 1984a. Distance Traveled To Visit N Points With a Maximum of C Stops Per Vehicle: an Analytical Model and an Application. *Transportation Science*, 18(4), 331–350.

Daganzo, C.F.,1984b. The length of tours in zones of different shapes. *Transportation*

Research Part B, 18(2), 135–145.

Daganzo, C.F., 2005. *Logistics Systems Analysis*. Springer, Heidelberg.

Daganzo, C. F., Newell, G. F., 1986. Configuration of Physical Distribution Networks. *Networks*, 16(1986), 113–132.

Daneshzand, F., Shoeleh, R., 2009. Multifacility Location Problem in Farahani, R.Z., Hekmatfar, M. (Eds.), *Facility location: Concepts, models, algorithms and case studies*, Physica-Verlag, Heidelberg.

Dasci, A., Verter, V., 2001. Continuous model for production-distribution system design. *European Journal of Operational Research*, 129(2), 287–298.

de Armas, J., Juan, A. A., Marques, J. M., Pedroso, J. P., 2017. Solving the deterministic and stochastic uncapacitated facility location problem: from a heuristic to a simheuristic. *Journal of the Operational Research Society*, 68, 1161–1176.

de Figueiredo, E.B., Jayasundara, S., de Oliveira Bordonal, R., Berchielli, T.T., Reis, R.A., Wagner-Riddle, C., La Scala Jr., N., 2017. Greenhouse gas balance and carbon footprint of beef cattle in three contrasting pasture-management systems in Brazil. *Journal of Cleaner Production*, 142, 420–431.

de Koster, R.B.M., Le-Duc, T., Zaerpour, N., 2012. Determining the number of zones in a pick-and-sort order picking system. *International Journal of Production Research*, 50(3), 757–771.

de Koster, R., Le-Duc, T., Roodbergen, K.J., 2007. Design and control of warehouse order picking: A literature review. *European Journal of Operational Research*, 182(2), 481–501.

DeHoratius, N., de Leeuw, S., Hübner, A., Rooderkerk, R.P., 2020. Call for papers:Managing the Marketing-Operations Interface in Omnichannel Retail. *Journal of Operations Management*.

Delaney-Klinger, K., Boyer, K.K., Frohlich, M., 2003. The return of online grocery shopping: a comparative analysis of Webvan and Tesco's operational methods. *The TQM Magazine*, 15(3), 187-196.

Deleersnyder, B., Geyskens, I., Gielens, K., Dekimpe, M.G., 2002. How cannibalistic is the Internet channel? A study of the newspaper industry in the United Kingdom and The Netherlands. *International Journal of Research in Marketing*, 19(4), 337–348.

Deloitte, 2014. *The shed of the future E-commerce: its impact on warehouses*. Available at: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/real-estate/deloitte-uk-shed-of-the-future.pdf (Accessed: 14 March 2018).

Deloitte, 2015. *Omni-channel retail: A Deloitte Point of View*. Available at: https://www2.deloitte.com/content/dam/Deloitte/se/Documents/technology/Omni-channel-2015.pdf (Accessed: 14 March 2018).

Deloitte, 2017. *Building your omni-channel journey*. Available at: https://www2.deloitte.com/content/dam/Deloitte/be/Documents/strategy/Omni-channel_Sales_Collateral-ONLINE.pdf (Accessed: 14 March 2018).

Department for Transport (DfT), 2013. *Cut business costs through improved van fleet management*. Available at: https://energysavingtrust.org.uk/sites/default/files/Van-fleet-management-best-practice-guide.pdf (Accessed: 6 July 2019).

Devari, A., Nikolaev, A.G., He, Q., 2017. Crowdsourcing the last mile delivery of online orders by exploitting the social networks of retail store customers. *Transportation Research Part E: Logistics and Transportation Review*, 105, 105-122.

DHL, 2016. *Omni-channel retail: Fulfilling demand profitably*. Available at: http://thehive-network.com/wp-content/uploads/2016/10/DHL-Omni-channel-Retail-Fulfilling-Demand-Profitably.pdf (Accessed: 14 March 2018).

Difrancesco, R.M., Huchzermeier, A., Schroder, D., 2018. Optimizing the return window for online fashion retailers with closed-loop refurbishment. *Omega*, 78, 205-221.

Digitalcommerce360, 2017. *Survey: Consumers turn to e-commerce amid changing retail landscape*. Available at: https://www.digitalcommerce360.com/2017/03/27/consumers-want-free-shipping-and-theyre-not-willing-to-wait-very-long-for-delivery/ (Accessed: 6 July 2019).

Dijkstra, A.S., Van der Heide, G., Roodbergen, K.J., 2019. Transshipments of cross-channel returned products. *International Journal of Productions Economics*, 209, 70-77.

Ding, Q., Kouvelis, P., Milner, J.M., 2006. Dynamic pricing through discounts for optimizing multiple-class demand fulfilment. *Operations Research*, 54(1), 169-183.

Dixons Carphone, 2017. *Currys PC World appoints On the Dot to offer customers ultimate delivery convenience*. Available at: https://www.dixonscarphone.com/en/news-and-media/press-releases/year/2017/currys-pc-world-appoints-on-the-dot-to-offer-customers-ultimate-delivery-convenience (Accessed: 30 October 2019).

Dixons Carphone, 2019. *Dixons Carphone Annual Report and Accounts 2018/2019*.

Dixonscarphone.com. 1*0 things you should know about KNOWHOW*. Available at: https://careers.dixonscarphone.com/social-hub/10-things-you-should-know-about-knowhow (Accessed: 30 October 2019).

Doerr, K.H., Gue, K.R., 2013. A Performance Metric and Goal-Setting Procedure for Deadline-Oriented Processes. *Production and Operations Management*, 22(3), 726–738.

Drexl, M., Schneider, M., 2015. A survey of variants and extensions of the location-routing problem. *European Journal of Operational Research*, 241(2), 283–308.

EcommerceDB, 2018. *E-commerce revenue analytics for habitat.co.uk*. Available at: https://ecommercedb.com/en/store/habitat.co.uk (Accessed: 30 October 2019).

eDelivery, 2015. *John Lewis: Omnichannel leadership through automation*. Available at: https://edelivery.net/2015/10/john-lewis-omnichannel-leadership-through-automation-edm02/ (Accessed: 30 October 2019).

Ehmke, J.F., Campbell, A.M., 2014. Customer acceptance mechanisms for home deliveries in metropolitan areas. *European Journal of Operational Research*, 233(1), 193–207.

Ellis-Chadwick, F., Doherty, N.F., Anastasakis, L., 2007. E-strategy in the UK retail grocery sector: a resource-based analysis. *Managing Service Quality*, 17(6), 702-727.

Eksioglu, B., Vural, A.V., Reisman, A., 2009. The vehicle routing problem: A taxonomic review. *Computers & Industrial Engineering*, 57(4), 1472–1483.

Enders, A., Jelasssi, T., 2009. Leveraging multichannel retaiLing: the experience of tesco.com. *MIS Quarterly Executive*, 8(2), 89-100.

Eppen, L., Schrage, G., 1981. Centralised ordering policies in a multi-warehouse system with lead times and random demand. In: Schwarz (Ed.), *TIMS Studies in the Management Sciences*, 16, 51-67.

Erlebacher, S.J., Meller, R.D., 2000. The interation of location and inventory in designing distribution systems. *IIE Transactions*, 32, 155-166.

Ernst, A.T., Jiang, H., Krishnamoorthy, M., Sier, D., 2004. Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research*, 153(1), 3–27.

EY, 2015. *Re-engineering the supply chain for the omni-channel of tomorrow - Global consumer goods and retail omni-channel supply chain survey*. Available at: https://www.ey.com/Publication/vwLUAssets/EY-re-engineering-the-supply-chain-for-the-omni-channel-of-tomorrow/%24FILE/EY-re-engineering-the-supply-chain-for-the-omni-channel-of-tomorrow.pdf (Accessed: 14 March 2018).

Farahani, R. Z., Rezapour, S., Drezner, T., Fallah, S., 2014. Competitive supply chain network design: An overview of classifications, models, solution techniques and applications. *Omega*, 45, 92–118.

Farahani, R. Z., Bajgan, H. R., Fahimnia, B., Kaviani, M., 2015. Location-inventory problem in supply chains: a modelling review. *International Journal of Production Research*, 53(12), 3769–3788.

Fernie, J., Sparks, L., 2014. *Logistics and retail management : emerging issues and new challenges in the retail supply chain*. 4th edn. London: Kogan Page.

Financial Times, 2013. *Domino's deadline to deliver*. Available at: https://www.ft.com/content/42134ce2-5f24-11e2-8250-00144feab49a (Accessed: 22 August 2018).

Financial Times, 2014. *Sainsbury to test online clothing sales*. Available at https://www.ft.com/content/f9709320-eda0-11e3-8a00-00144feabdc0 (Accessed: 6 July

2019).

Financial Times, 2016a. *UK retailers count the cost of returns*. Available at: https://www.ft.com/content/52d26de8-c0e6-11e5-846f-79b0e3d20eaf (Accessed: 28 December 2018).

Financial Times, 2016b. *'Last mile' puzzle drives demand for warehousing hubs*. Available at: https://www.ft.com/content/51dc8288-15ec-11e6-b197-a4af20d5575e (Accessed: 6 July 2019).

Financial Times, 2016c. *Dunelm posts 5% rise in profits, grows market share*. Available at: https://www.ft.com/content/07fe5e50-eca2-36e8-9d3e-069eb4486979 (Accessed: 6 July 2019).

Financial Times, 2018. *Warehouses reach for the sky as urban demand soars.* Available at: https://www.ft.com/content/9c066586-3d77-11e8-bcc8-cebcb81f1f90 (Accessed: 23 January 2019).

Fleischmann, M., Bloemhof-Ruwaard, J.M., Dekker, R., van der Laan, E., van Nunen, J.A.E.E., Van Wassenhove, L.N., 1997. Quantitative models for reverse logistics: a review. *European Journal of Operational Society*, 103(1), 1-17.

Forbes, 2018. *Walmart ecommerce tactic against Amazon look to be paying off*. Available at: https://www.forbes.com/sites/andriacheng/2018/08/16/walmarts-ecommerce-tactic-against-amazon-is-paying-off/#47ec8b55b74d (Accessed: 6 July 2019).

Forbes, 2019a. *Brdiging the gap: 5 online-to-offline retail innovations reviving the store*. Available at: https://www.forbes.com/sites/victoriapavlova/2019/02/04/bridging-the-gap-5-online-to-offline-retail-innovations-reviving-the-store/#7aa215ae3bd7 (Accessed: 6 July 2019).

Forbes, 2019b. *Target, Walmart results show they do have an edge against Amazon*. Available at: https://www.forbes.com/sites/andriacheng/2019/08/21/target-walmart-results-show-they-do-have-this-edge-against-amazon/#4020e4081d35 (Accessed: 6 July 2019).

Forbes, 2019c. *2000-consumer study: 68% of shoppers* '*click and collect*'*, or Buy Online and Pick Up In Store*. Available at: https://www.forbes.com/sites/johnkoetsier/2019/03/11/hope-for-brick-and-mortar-retail-68-of-shoppers-click-and-collect-or-buy-online-and-pick-up/#677e22237708(Accessed: 6 July 2019).

Forbes, 2019d. *Pop-Up stores and small-format stores are changing the retail landscape*. Available at: https://www.forbes.com/sites/jiawertz/2019/07/31/pop-up-shops-small-format-stores-changing-retail-landscape/#41bd7ad9143a (Accessed: 20 January 2020).

Fornari, E., Fornari, D., Grandi, S., Menegatti, M., Hofacker, C.F., 2016. Adding store to web: migration and synergy effects in multi-channel retailing. *International Journal of*

Retail & Distribution Management, 44(6), 658-674.

Forrester Research, 2014. *Customer Desires Vs. Retailer Capabilities: Minding The Omni-Channel Gap*. Available at: https://www.accenture.com/us-en/~/media/accenture/conversion-assets/dotcom/documents/global/pdf/technology_7/accenture-customer-desires-vs-retailer-capabilities.pdf (Accessed: 14 March 2018).

Forrester Research, 2018. *Omnichannel mastery: optimise fulfilment and engagement*. Available at: https://www.forrester.com/report/Omnichannel+Mastery+Optimise+Fulfilment+And+Engagement/-/E-RES116388 (Accessed: 6 July 2019).

Galipoglu, E., Kotzab, H., Teller, C., Hüseyinoglu, I.O.Y., Pöppelbuß, J., 2018. Omni-channel retailing research - state of the art and intellectual foundation. *International Journal of Physical Distribution & Logistics Management*, 48(4), 365–390.

Gallino, S., Moreno, A., 2014. Integration of Online and Offline Channels in Retail: The Impact of Sharing Reliable Inventory Availability Information. *Management Science*, 60(6), 1434–1451.

Gallino, S., Moreno, A., Stamatopoulos, I., 2017. Channel Integration, Sales Dispersion, and Inventory Management. *Management Science*, 63(9), 2813–2831.

Gao, F., Su, X., 2017. Omnichannel Retail Operations with Buy-Online-and-Pick-up-in-Store. *Management Science*, 63(8), 2397-2771.

Gibson, B.J., Defee, C.C., Ishfaq, R., 2016. *The state of the retail supply chain: Essential findings of the sixth annual study*. Auburn University Center For Supply Chain Innovation.

GlobalData,2018a. *M&S to lose its decades long status as the number one clothing retailer to Primark*. Available at: https://www.globaldata.com/ms-lose-decades-long-status-number-one-clothing-retailer-primark/ (Accessed: 6 July 2019).

GlobalData, 2018b. *UK click & collect sales set to soar 55.6% to 2022*. Available at: https://www.globaldata.com/uk-click-collect-sales-set-soar-55-6-2022/(Accessed: 6 July 2019).

Golden, B.L., Raghavan, S., Wasil, E.A. eds., 2008. *The Vehicle Routing Problem: Latest Advances and New Challenges: latest advances and new challenges*, Springer, New York.

Ghosh, D., 2003. Neighborhood search heuristics for the uncapacitated facility location problem. *European Journal of Operational Research*, 150(1), 150–162.

Govindan, K., Soleimani, H., Kannan, D., 2015. Reverse logistics and closed-loop supply chain: A comprehensive review to explore the future. *European Journal of Operational Research*, 240(3), 603–626.

Govindarajan, A., Sinha, A., Uichanco, J., 2017. *Inventory Optimisation for Fulfilment Integration in Omnichannel Retailing.* Ross School of Business, Working Paper No. 1341.

Grant, D.B., Fernie, J., Schulz, B., 2014. Enablers and Barriers in German Online Food Retailing. *Supply Chain Forum: An International Journal*, 15(3), 4–11.

Griffis, S.E., Rao, S., Goldsby, T.J., and Niranjan, T.T., 2012. The customer consequences of returns in online retailing: an empirical analysis. *Journal of Operations Management*, 30, 282-294.

Gu, J., Goetschalckx, M., McGinnis, L.F., 2010. Research on warehouse design and performance evaluation: A comprehensive review. *European Journal of Operational Research*, 203(3), 539–549.

Gue, K.R., Meller, R.D., Skufca, J.D., 2006. The effects of pick density on order picking areas with narrow aisles. *IIE Transactions*, 38(10), 859–868.

Gulati, R., Garino, J., 2000. Get the Right of Bricks & Clicks. *Harvard business review (4)*, (May-June), 107–114.

Gümüş, M., Li, S., Oh, W., Ray, Saibal., 2013. Shipping Fees or Shipping Free? A Tale of Two Price Partitioning Strategies in Online Retailing. *Production and Operations Management*, 22(4), 758–776.

Ha, A.Y., 1997. Stock-rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Research Logistics*, 44(5), 457–472.

Harris, L.C., 2010. Fradulent consumer returns: exploiting retailers' return policies. *European Journal of Marketing*, 44(6), 730-747.

Harsha, P., Subramanian, S., Ettl, M., 2019. A Practical Price Optimisation Approach for Omnichannel Retailing. *Journal on Optimisation*, 1(3), 241–264.

He, Y., Xu, Q., Wu, P., 2020. Omnichannel retail operations with refurbished consumer returns. *International Journal of Production Research*, 58(1), 1–20.

Herhausen, D., Binder, J., Schoegel, M., Herrmann, A., 2015. Integrating Bricks with Clicks: Retailer-Level and Channel-Level Outcomes of Online-Offline Channel Integration. *Journal of Retailing*, 91(2), 309–325.

Hjort, K., Lantz, B., 2012. (R)e-tail borrowing of party dresses: an experimental study. *International Journal of Retail & Distribution Management*, 40(12), 997-1012.

Hotelling, H., 1929. Stability in competition. *The Economic Journal*, 39(153), 41-57.

Hovelaque, V., Soler, L.G., Hafsa, S., 2007. Supply chain organization and e-commerce: A model to analyze store-picking, warehouse-picking and drop-shipping. *4or*, 5(2), 143–155.

Hübner, A., Holzapfel, A., Kuhn, H., 2015. Operations management in multi-channel retailing: an exploratory study. *Operations Management Research*, 8(3–4), 84–100.

Hübner, A., Holzapfel, A., Kuhn, H., 2016a. Distribution systems in omni-channel retailing. *Business Research*, 9(2), 255–296.

Hübner, A., Kuhn, H., Wollenburg, J., 2016b. Last mile fulfilment and distribution in omni-channel grocery retailing: A strategic planning framework. *International Journal of Retail & Distribution Management*, 44(3), 228–247.

Hübner, A., Wollenburg, J., Holzapfel, A., 2016c. Retail logistics in the transition from multi-channel to omni-channel. *International Journal of Physical Distribution & Logistics Management*, 46 (6/7), 562–583.

Hung, H.-C., Chew, E.P., Lee, L.H., Liu, S., 2012. Dynamic inventory rationing for systems with multiple demand classes and general demand processes. *International Journal of Production Economics*, 139(1), 351–358.

IBM, 2001. *IBM Global Services: online returns management strategies*. Available at: http://www-07.ibm.com/services/pdf/IBM_Consulting_Online_returns_management.pdf (Accessed: 6 July 2019).

IBM, 2017. *The rising costs of fulfilment: a hidden cost of the omnichannel model*. Available at: https://www.ibm.com/blogs/watson-customer-engagement/2017/08/28/rising-costs-fulfilment/ (Accessed: 6 July 2019).

Internet Retailing, 2010. T*esco average basket size hits £100 as customer satisfaction improves*. Available at: https://internetretailing.net/themes/themes/tesco-average-basket-size-hits-100-as-customer-satisfaction-improves-6206 (Accessed: 6 July 2019).

Ishfaq, R., Bajwa, N., 2019. Profitability of online order fulfilment in multi-channel retailing. *European Journal of Operational Research*, 272(3), 1028–1040.

Ishfaq, R., Defee, C.C., Gibson, B.J., Raja, U., 2016. Realignment of the physical distribution process in omni-channel fulfilment. *International Journal of Physical Distribution & Logistics Management*, 46(6/7), 543-561.

Ishfaq, R., Raja, U., 2018. Evaluation of Order Fulfilment Options. *Decision Sciences*, 49(3), 487–521.

ITV.News, 2020. *Co-op announces expansion plans*. Available at: https://www.itv.com/news/2020-01-02/co-op-announces-expansion-plans/ (Accessed: 2 January 2020).

Jasin, S., Sinha, A., 2015. An LP-Based Correlated Rounding Scheme for Multi-Item Ecommerce Order Fulfilment. *Operations Research*, 63(6), 1336–1351.

Jaillet, P., 1988. A priori solution of a traveling salesman problem in which a random subset of the customers are visited. *Operations Research*, 36(6), 929-936.

JDA, 2015. *Global retail & consumer goods CEO survey: the omni-channel fulfilment imperative.* Available at: https://bridgesgi.com/wp-content/uploads/2015/05/JDA-PWC-

Executive-Summary_D.pdf (Accessed: 30 October 2019).

JDA, 2016. *In-store picking creates havoc for stores*. Available at: https://jda.com/knowledge-center/collateral/in-store-picking-white-paper (Accessed: 1 May 2019).

Jena, S. D., Cordeau, J. F., Gendron, B., 2015. Dynamic Facility Location with Generalized Modular Capacities. *Transportation Science*, 49(3), 484–499.

Jin, A., Galiskan-Demirag, O., Chen, F.Y., Huang, M., 2019. Omnichannel retailers' return policy strategies in the presence of competition. In press.

Jin, M., Li, G., Cheng, T.C.E., 2018. Buy online and pick up in-store: Design of the service area. *European Journal of Operational Research*, 268(2), 613-623.

Johnlewis.com. *Delivery by our approved suppliers*. Available at: https://www.johnlewis.com/customer-services/delivery-information/delivery-by-our-approved-suppliers (Accessed: 30 October 2019).

John Lewis. *Delivery & Collection*. Available at: https://www.johnlewis.com/customer-services/delivery-information (Accessed: 6 July 2019).

John Lewis, 2018. *John Lewis Partnership PLC Annual Report and Account 2018*.

Kabak, Ö., Ülengin, F., Aktaş, E., Önsel, S., Topcu, Y.I., 2008. Efficient shift scheduling in the retail sector through two-stage optimisation, *European Journal of Operational Research*, 184(1), 76–90.

Kämäräinen, V., Saranen, J., Holmström, J., 2001. The reception box impact on home delivery efficiency in the e‐grocery business. *International Journal of Physical Distribution & Logistics Management*, 31(6), 414–426.

Kembro, J. H., Norrman, A., Eriksson, E., 2018. Adapting warehouse operations and design to omni-channel logistics. *International Journal of Physical Distribution & Logistics Management*, 48(9), 890–912.

Khouja, M., 2001. The evaluation of drop shipping option for e-commerce retailers. *Computers & Industrial Engineering*, 41(2), 109-126.

Khouja, M., Stylianou, A.C., 2009. A (Q, R) inventory model with a drop-shipping option for e-business. *Omega*, 37(4), 896–908.

Klose, A., Drexl, A., 2005. Facility location models for distribution system design. *European Journal of Operational Research*, 162(1), 4–29.

Ko, C.S., Min, H., Ko, H.J., 2007. Determination of cutoff time for express courier servicers: a genetic algorithm approach. *International Transactions in Operations Research*, 14(2), 159-177.

Kollmann, T., Kuckertz, A., Kayser, I., 2012. Cannibalization or synergy? Consumers' channel selection in online–offline multichannel systems. *Journal of Retailing and Consumer Services*, 19(2), 186–194.

KPMG. 2018. *KPMG Annual Retail Survey 2018*. Available at: https://home.kpmg/content/dam/kpmg/uk/pdf/2018/01/kpmg-annual-retail-survey-2018.pdf (Accessed: 6 July 2019).

Laporte, G., 1992. The vehicle routing problem: An overview of exact and approximate algorithms. *European Journal of Operational Research*, 59(3), 345–358.

Laporte, G.,2009. Fifty years of vehicle routing. *Transportation Science*, 43(4), 408-416.

Langevin, A., Mbaraga, P., Campbell, J.F., 1996. Continuous approximation models in freight distribution: An overview. *Transportation Research Part B: Methodological*, 30(3), 163–188.

Lee, H. L., Whang, S., 2001. Winning the Last Mile of E-Commerce. *MIT Sloan management review*, 42, 54–62.

Leng, M., Becerril-Arreola, R., 2010. Joint pricing and contingent free-shipping decisions in B2C transactions. *Production and Operations Management*, 19(4), 390–405.

Li, T., Zhao, X., Xie, J., 2015a. Inventory management for dual sales channels with inventory-level-dependent demand. *Journal of Operational Research Society*, 66(3), 488–499.

Li, Z. (Erick), Lu, Q., Talebian, M., 2015b. Online versus bricks-and-mortar retailing: a comparison of price, assortment and delivery time. *International Journal of Production Research*, 53(13), 3823–3835.

Liu, K., Zhou, Y., Zhang, Z., 2010. Capacitated location model with online demand pooling in a multi-channel supply chain. *European Journal of Operational Research*, 207(1), 218–231.

Liu, S., Papageorgiou, 2013. Multiobjective optimisation of production, distribution and capacity planning of global supply chains in the process industry. *Omega*, 41(2), 369–382.

Liu, L., Huang, C.Z., Huang, G., Baetz, B., Pittendrigh, S.M., 2018a. How a carbon tax will affect an emission-intensive economy: a case study of the province of saskatchewan, Canada. *Energy*, 159, 817–826.

Liu, L., Huang, G., Baetz, B., Huang, C.Z., Zhang, K., 2018b. A factorial ecologicallyextended input-output model for analyzing urban GHG emissions metabolism system. *Journal of Cleaner Production*, 200, 922–933.

Lovell, A., Saw, R., Stimson, J., 2005. Product value-density: managing diversity through supply chain segmentation. *The International Journal of Logistics Management*, 16(1), 142-158.

Lunce, S.E., Lunce, L.M., Kawai, Y., Maniam, B., 2006. Success and failure of pure-player organizations: Webvan versus Peapod, a comparative analysis. *Industrial Management & Data Systems*, 106(9), 1344-1358.

Ma, S., Jemai, Z., Sahin, E., Dallery, Y., 2017. The news-vendor problem with drop-shipping and resalable returns. *International Journal of Production Research*, 55(22), 6547–6571.

Mahar, S., Bretthauer, K.M., Venkataramanan, M.A., 2009a. An algorithm for solving the multi-period online fulfilment assignment problem. *Mathematical and Computer Modelling*, 50(9–10), 1294–1304.

Mahar, S., Bretthauer, K.M., Venkataramanan, M.A., 2009b. The value of virtual pooling in dual sales channel supply chains. *European Journal of Operational Research*, 192(2), 561–575.

Mahar, S., Wright, P.D., 2009. The value of postponing online fulfilment decisions in multi-channel retail/e-tail organisations. *Computers & Operations Research*, 36(11), 3061-3072.

Mahar, S., Wright, P.D., 2017. In-Store Pickup and Returns for a Dual Channel Retailer. *IEEE Transactions on Engineering Management*, 64(4), 491–504.

Mahar, S., Wright, P.D., Bretthauer, K.M., Hill, R.P., 2014. Optimizing marketer costs and consumer benefits across "clicks" and "bricks". *Journal of the Academy of Marketing Science*, 42(6), 619–641.

Maister, D.H., 1976. Centralisation of inventories and the 'square root law'. *International Journal of Physical Distribution*, 6(3), 124-134.

Mangiaracina, R., Song, G., Perego, A., 2015. Distribution network design: a literature review and a research agenda. *International Journal of Physical Distribution & Logistics Management*, 45(5), 506–531.

Manzini, R., Gebennini, E., 2008. Optimisation models for the dynamic facility location and allocation problem. *International Journal of Production Research*, 46(8), 2061–2086.

Marchet, G., Melacini, M., Perotti, S., Rasini, M., Tappia, E., 2018. Business Logistics models in omni-channel: a classification framework and empirical analysis. *International Journal of Physical Distribution & Logistics Management*, 48(4), 439–464.

Marino, G., Zotteri, G., Montagna, F., 2018. Consumer sensitivity to delivery lead time: a furniture retail case. *International Journal of Physical Distribution & Logistics Management*, 48(6), 610-629.

Marks and Spencer. *UK Delivery & Collection*. Available at: https://www.marksandspencer.com/c/delivery-and-collection (Accessed: 6 July 2019).

Marks and Spencer, 2018a. *Annual report & financial statements 2018*.

Marks and Spencer, 2018b. *Marks & Spencer announces changes to its logistics network*. Available at: https://corporate.marksandspencer.com/media/press-releases/2018/marks-

and-spencer-announces-changes-to-its-logistics-network (Accessed: 6 July 2019).

McKinsey & Company, 2013. *The future of online grocery in Europe*. Available at: https://www.mckinsey.com/~/media/McKinsey/Industries/Retail/Our%20Insights/The%20future%20of%20online%20grocery%20in%20Europe/The_future_of_online_grocery.ashx (Accessed: 6 July 2019)

McKinsey & Company. 2019. *How lessons from the record-breaking 2018 holiday season can inform retailers' 2019 strategy*. Available at: https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/how-lessons-from-the-record-breaking-2018-holiday-season-can-inform-retailers-2019-strategy?cid=other-eml-alt-mip-mck&hlkid=6555f82d303f413ba153e9a70fe7e270&hctky=1440062&hdpid=54522629-79bb-4441-89a9-891d21633fb9 (Accessed: 6 July 2019).

Melachrinoudis, E., Messac, A., Min, H., 2005. Consolidating a warehouse network: A physical programming approach. *International Journal of Production Economics*, 97, 1–17.

Melo, M.T., Nickel, S., Saldanha-da-Gama, F., 2009. Facility location and supply chain management - a review. *European Journal of Operational Research*, 196(2), 401-412.

Menzies Distribution, 2019. *WH Smith case study*. Available at: https://www.menziesdistribution.com/case-studies/whsmith (Accessed: 6 July 2019).

Metro, 2018. *Customers slam Amazon 'next-day delivery' after presents fail to arrive*. Available at: https://metro.co.uk/2018/12/24/customers-slam-amazon-next-day-delivery-presents-fail-arrive-8281142/ (Accessed: 6 July 2019).

Metters, R., Walton, S., 2007. Strategic supply chain choices for multi-channel Internet retailers. *Service Business*, 1(4), 317–331.

Mirrazavi, S.K., Beringer, H., 2007. A web-based workforce management system for Sainsburys supermarket Ltd. *Annals of Operations Research*, 155, 437-457.

Min, H., Ko, H.J., Ko, C.S., 2006. A genetic algorithm approach to developing the multi-echelon reverse logistics network for product returns. *Omega*, 34(1), 56–69.

Minnema, A., Bijmolt, T.H.A., Gensler, S., Wiesel, T., 2016. To keep or not to keep: effects of online customer reviews on product returns. *Journal of Retailing*, 92(3), 253-267.

Minner, S., 2003. Multiple-supplier inventory models in supply chain management: A review. *International Journal of Production Economics*. 81-82(2003), 265-279.

Mintel. 2019. *Online Grocery Retailing – UK*.

Mirsch, T., Lehrer, C., Jung, R., 2016. Channel integration towards omnichannel management: a literature review, *in proceeding of the 20th Pacific Asia Conference on Information Systems (PACIS 2016)*, Chiayi, Taiwan.

Modak, N.M., Kelle, P., 2019. Managing a dual-channel supply chain under price and

delivery-time dependent stochastic demand. *European Journal of Operational Research*, 272(1), 147–161.

Mohammed, F., Selim, S.Z., Hassan, A., Syed, M.N., 2017. Multi-period planning of closed-loop supply chain with carbon policies under uncertainty. *Transportation Research Part D: Transport and Environment*, 51, 146–172.

Möllering, K.T., Thonemann, U.W., 2008. An optimal critical level policy for inventory systems with two demand classes. *Naval Research Logistics*, 55(7), 632–642.

Montoya, A., Vélez–Gallego, M. C. and Villegas, J. G., 2016. Multi-product capacitated facility location problem with general production and building costs. *NETNOMICS: Economic Research and Electronic Networking*, 17(1), 47–70.

Muir, W.A., Griffis, S.E., Whipple, J.M., 2019. A simulation model of multi-echelon retail inventory with cross-channel product returns. *Journal of Business Logistics*, 40(4), 322-338.

Nageswaran, L., Cho, S.H., Scheller-Wolf, A., 2019. Consumer return policies in omnichannel operations. *Management Science*, Forthcoming.

Nekoiemehr, N., Zhang, G., Selvarajah, E., 2018. Due date quotation in a dual-channel supply chain. *International Journal of Production Economics*, 215(2019), 102-111.

Neslin, S. A., Grewal, D., Leghorn, R., Shankar, V., Teerling, M. L., Thomas, J. S., Verhoef, P. C., 2006. Challenges and Opportunities in Multichannel Customer Management. *Journal of Service Research*, 9(2), 95–112.

Netessine, S., Rudi, N., 2006. Supply Chain Choice on the Internet. *Management Science*, 52(6), 844–864.

New Look Ltd, 2019. *Annual report*.

Newell, G.F., 1971. Dispatching Policies for a Transportation Route. *Transportation Science*, 5(1), 91–105.

Newell, G.F., 1973. Scheduling , Location , Transportation , and Continuum Mechanics ; Some Simple Approximations to Optimisation Problems. *SIAM Journal on Applied Mathematics*, 25(3), 346–360.

Next, 2019. *Annual report and accounts*.

Nezhad, A. M., Manzour, H., Salhi, S., 2013. Lagrangian relaxation heuristics for the uncapacitated single-source multi-product facility location problem. *International Journal of Production Economics*, 145(2), 713–723.

Nguyen, D.H., de Leeuw, S., Dullaert, W.E.H., 2018. Consumer behaviour and order fulfilment in online retailing: a systematic review. *International Journal of Management Reviews*, 20, 255-276.

Nguyen, T.H., Wright, M., 2015. Capacity and lead-time management when demand for service is seasonal and lead-time sensitive. *European Journal of Operational Research*,

247(2), 588–595.

NottinghamshireLive, 2017. *A decade of supply-chain genius at Boots*. Available at: https://www.nottinghampost.com/news/nottingham-news/how-boots-sold-10-billion-544098 (Accessed: 6 July 2019).

Nyquist, E., 2016. How to Make the Most of Omnichannel Retailing. *Harvard business review*, July-August.

Ocado, 2018. *OCADO GROUP PLC Annual Report and Accounts*.

Ocado, 2019. *OCADO GROUP PLC Annual Report and Accounts*.

Ofek, E., Katona, Z., Sarvary, M., 2011. "Bricks and clicks": The impact of product returns on the strategies of multichannel retailers. *Marketing Science*, 30(1), 42–60.

Office for National Statistics (ONS), 2018. *Comparising 'bricks and mortar' store sales with online retail sales: August 2018.* Available at: https://www.ons.gov.uk/businessindustryandtrade/retailindustry/articles/comparingbricksandmortarstoresalestoonlineretailsales/august2018 (Accessed: 6 July 2019).

Onstein, A. T. C., Tavasszy, L. A., van Damme, D. A., 2019. Factors determining distribution structure decisions in logistics: a literature review and research agenda. *Transportation Reviews*, 39(2), 243–260.

OrderDynamics. 2018. *Omni-2000 Research: UK*.

Owen, S.H., Daskin, M.S., 1998. Strategic facility location: A review. *European Journal of Operational Research*, 111(3), 423–447.

Palaka, K., Erlebacher, E., Kropp, D.H., 1998. Lead-time setting, capacity utilization, and pricing decisions under lead-time dependent demand. *IIE Transactions*, 30(2), 151–163.

Pan, J. C.-H., Wu, M.-H., 2012. Throughput analysis for order picking system with multiple pickers and aisle congestion considerations. *Computers & Operations Research*, 39(7), 1661–1672.

Pang, G., Muyldermans, L., 2013. Vehicle routing and the value of postponement. *Journal of the Operational Research Society*, 64(9), 1429–1440.

Parikh, P.J., Meller, R.D., 2008. Selecting between batch and zone order picking strategies in a distribution center. *Transportation Research Part E: Logistics and Transportation Review*, 44(5), 696–719.

Pastor, R., Olivella, J., 2008. Selecting and adapting weekly work schedules with working time accounts: a case of a retail clothing chain. *European Journal of Operational Research*, 184(1), 1-12.

Paul, J., Agatz, N., Spliet, R., De Koster, R., 2019. Shared Capacity Routing Problem - An Omni-channel Retail Study. *European Journal of Operational Research*, 273(2), 731-739.

Pauwels, K., Neslin, S.A., 2015. Building With Bricks and Mortar: The Revenue Impact of

Opening Physical Stores in a Multichannel Environment. *Journal of Retailing*, 91(2), 182–197.

Payne, E., Peltier, J. W., Barger, V. A., 2017. Omni-channel marketing, integrated marketing communications and consumer engagement: A research agenda. *Journal of Research in Interactive Marketing*,11(2), 185–197.

Perl, J., Daskin, M. S., 1985. A warehouse location-routing problem. *Transportation Research Part B: Methodological*, 19(5), 381–396.

Pillac, V., Gendreau, M., Gueret, C., Medaglia, A.L., 2013. A review of dynamic vehicle routing problems. *European Journal of Operational Research*, 225(1), 1–11.

Pishgar-Komleh, S., Akram, A., Keyhani, A., Raei, M., Elshout, P., Huijbregts, M., Van Zelm, R., 2017. Variability in the carbon footprint of open-field tomato production in Iran-a case study of Alborz and East-Azerbaijan provinces. *Journal of Cleaner Production*, 142, 1510–1517.

Punakivi, M., Yrjölä, H., Holmström, J., 2001. Solving the last mile issue: Reception box or delivery box? *International Journal of Physical Distribution and Logistics Management*, 31(6), 427–439.

PwC, 2016. *They say they want a revolution - Total retail 2016*. Available at: https://www.pwc.com/gx/en/retail-consumer/publications/assets/total-retail-global-report.pdf. (Accessed: 6 July 2019)

PwC, 2017. *The 2017 global omnichannel retail index: omnichannel on the march*. Available at: https://www.strategyand.pwc.com/gx/en/reports/the-2017-global-omnichannel-retail-index.pdf (Accessed: 6 July 2019)

Rabinovich, E., Sousa, R., Park, S., Golara, S., 2018. Omnichannel retailing as a balancing act between in-store and home fulfilment. *SSRN Electronic Journal*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3113878 (Accessed: 6 July 2019).

Radhi, M., Zhang, G., 2018. Pricing policies for a dual-channel retailer with cross-channel returns. *Computers and Industrial Engineering*, 119, 63–75.

Radhi, M., Zhang, G., 2019. Optimal cross-channel return policy in dual-channel retailing systems. *International Journal of Production Economics*, 210, 184–198.

Randall, T., Netessine, S., Rudi, N., 2006. An Empirical Examination of the Decision to Invest in Fulfilment Capabilities: A Study of Internet Retailers. *Management Science*, 52(4), 567–580.

Rao, S., Griffis, S.E., Goldsby, T.J., 2011. Failure to deliver? Linking online order fulfilment glitches with future purchase behavior. *Journal of Operations Management*, 29(7-8), 692-703.

Rao, U.S., Swaminathan, J.M., Zhang, J., 2005. Demand and Production Management with Uniform Guaranteed Lead Time. *Production and Operations Management*, 14(4), 400–

412.

Ray, S., Jewkes, E.M., 2004. Customer lead time management when both demand and price are lead time sensitive. *European Journal of Operational Research*, 153(3), 769–781.

Retail Gazette, 2018. *PrettyLittleThing suspends next day delivery due to 'overwhelming demand'*. Available at: https://www.retailgazette.co.uk/blog/2018/07/prettylittlething-suspends-next-day-delivery-due-overwhelming-demand/ (Accessed: 6 July 2019).

Revelle, C. S., Eiselt, H. A., Daskin, M. S., 2008. A Bibliography for Some Fundamental Problem Categories in Discrete Location Science. *European Journal of Operational Research*, 184(3), 817–848.

Rogers, D.S., Melamed, B., Lembke, R.S., 2012. Modeling and analysis of reverse logistics. *Journal of Business Logistics*, 33(2), 107–117.

Sabri, E. H., Beamon, B. M., 2000. A multi-objective approach to simultaneous strategic and operational planning in supply chain design. *Omega*. 28(5), 581–598.

Saghiri, S., Wilding, R., Mena, C., Bourlakis, M., 2017. Toward a three-dimensional framework for omni-channel. *Journal of Business Research*, 77(2017), 53–67.

Sainsbury's, 2008. *Sainsbury's information direct - supply chain locations*. Available at: https://web.archive.org/web/20081022002924/http://www.sainsburys.co.uk/SID/locations/locations.htm (Accessed: 6 July 2019).

Sainsbury's, 2016. *Sainsbury's opens purpose-built centre to fulfil demand for online grocery orders in London*. Available at: https://www.about.sainsburys.co.uk/news/latest-news/2016/21-09-2016 (Accessed: 6 July 2019).

Sainsbury's, 2017. *Argos opens two new regional fulfilment centres, increasing fast track capacity*. Available at: https://www.about.sainsburys.co.uk/news/latest-news/2017/new-regional-fulfilment-centres (Accessed: 30 October 2019).

Sainsbury's. *Where and when can I collect my click & collect grocery orders?* Available at: https://help.sainsburys.co.uk/help/delivery-collection/cc-where-when (Accessed: 6 July 2019).

Sainsbury's. *Click & Collect Groceries.* Available at: https://www.sainsburys.co.uk/shop/gb/groceries/get-ideas/click-and-collect/click-and-collect (Accessed: 6 July 2019).

Sainsbury's, 2018. *Annual Report and Financial Statements 2018*.

Sakraida, L., 2017. *Amazon and Walmart Will Offer Store Credit for Late Christmas Deliveries*, *dealnews*. Available at: https://www.dealnews.com/features/Amazon-and-Walmart-Will-Offer-Store-Credit-for-Late-Christmas-Deliveries/944691.html (Accessed: 1 May 2019).

Salmani, Y., Partovi, F.Y., Banerjee, A., 2018. Customer-driven investment decisions in existing multiple sales channels: A downstream supply chain analysis. *International Journal of Production Economics*, 204, 44–58.

Schneider, F., Klabjan, D., 2013. Inventory control in multi-channel retail. *European Journal of Operational Research*, 227(1), 101–111.

Scott, J.E., Scott, C.H., 2008. Online grocery order fulfilment tradeoffs, in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, Waikoloa, USA.

Seifert, R., Thonemann, U., Sieke, M., 2006. Relaxing channel separation: Integrating a virtual store into the supply chain via transshipments. *IIE Transactions*, 38(11), 917–931.

Shang, G., Pekgün, P., Ferguson, M., Galbreth, M., 2017. How much do online consumers really value free product returns? Evidence from eBay. *Journal of Operations Management*, 53-56, 45-62.

Sherafati, M., Bashiri, M., Tavakkoli-Moghaddam, R., Pishvaee, M. S., 2019. Supply chain network design considering sustainable development paradigm: A case study in cable industry. *Journal of Cleaner Production*, 234, 366–380.

Sopadjieva, E., Dholakia, U.M., Benhamin, B., 2017. A Study of 46,000 Shoppers Shows That Omnichannel Retailing Works. *Harvard business review*.

Stein, D.M., 1978. Scheduling dial-a-ride transportation systems. *Transportation Science*, 12(3), 232-249.

Stock, J., Speh, T., Shear, H., 2006. Managing Product Returns for Competitive Advantage. *MIT Sloan Management Review*, 48(1), 57–62.

Stone, R.E., 1991. Some average distance results. *Transportation Science*, 25(1), 83–91.

Sun, M., 2012. A tabu search heuristic procedure for the capacitated facility location problem. *Journal of Heuristics*, 18, 91–118.

Superdrug. *Delivery Information*. Available at: https://www.superdrug.com/delInfo (Accessed: 6 July 2019).

SupplyChainBrain, 2017. *How Argos transformed its store network for the omnichannel*. Available at: https://www.supplychainbrain.com/articles/27350-how-argos-transformed-its-store-network-for-the-omnichannel (Accessed: 30 October 2019).

Supplychaindive, 2019. *For a fast supply chain, Target is betting on in-store fulfilemnt*. Available at: https://www.supplychaindive.com/news/fulfilment-brick-and-mortar-retail-inventory/560262/ (Accessed: 30 October 2019).

Supplychain 24/7, 2015. *Order Fulfilment as a Competitive Advantage*.Available at: https://www.supplychain247.com/article/order_fulfilment_as_a_competitive_advantage (Accessed: 6 July 2019).

Supplychain 24/7, 2018. Amazon's ever-increasing ecommerce shipping costs. Available at: https://www.supplychain247.com/article/amazons_ever_increasing_ecommerce_shipping_costs (Accessed: 6 July 2019).

Takahashi, K., Aoi, T., Hirotani, D., Morikawa, K., 2011. Inventory control in a two-echelon dual-channel supply chain with setup of production and delivery. *International Journal of Production Economics*, 133(1), 403–415.

Talarico, L., Duque, P.A.M., 2015. An optimisation algorithm for the workforce management in a retail chain. *Computers and Industrial Engineering*, 82, 65–77.

Tesco. *Same Day Delivery - Tesco Groceries*. Available at: https://www.tesco.com/groceries/en-GB/zone/Same-day-delivery (Accessed: 6 July 2019).

Tesco, 2014. *Click+Collect Groceries now available in 300 locations*. Available at: https://www.ourtesco.com/2014/09/17/clickcollect-groceries-now-available-300-locations/ (Accessed: 6 July 2019).

Tesco, 2019. *First quarter trading statement*. Available at: https://www.tescoplc.com/news/2019/first-quarter-trading-statement-201920/ (Accessed: 6 July 2019).

The Forrester, 2018. *Omnichannel Mastery : Optimise Fulfilment And Engagement Omnichannel Mastery : Optimise Fulfilment And Engagement*. Available at: https://www.forrester.com/report/Omnichannel+Mastery+Optimise+Fulfilment+And+Engagement/-/E-RES116388 (Accessed: 23 January 2019).

The Guardian, 2016. *Sainsbury's one-hour delivery service takes on Amazon*. Available at: https://www.theguardian.com/business/2016/sep/26/sainsburys-one-hour-delivery-takes-on-amazon-bikes-london (Accessed: 14 March 2018).

The Guardian, 2018a. *Toys R Us and Maplin fall into administration putting 5,500 UK jobs at risk*. Available at: https://www.theguardian.com/business/2018/feb/28/toys-r-us-falls-into-administration-putting-3200-uk-jobs-at-risk (Accessed: 14 March 2019).

The Guardian, 2018b. *Tesco shuts its non-food website, Tesco Direct*. Available at: https://www.theguardian.com/business/2018/feb/28/toys-r-us-falls-into-administration-putting-3200-uk-jobs-at-risk (Accessed: 14 March 2019).

The Independent, 2018. *Christmas online order deadlines 2018: From Amazon to Asos and John Lewis.* Available at: https://www.independent.co.uk/life-style/amazon-christmas-order-deadlines-delivery-online-shopping-2018-a8675871.html (Accessed: 6 July 2019).

The Sun, 2017. *Tesco extends same-day click and collect service so shoppers can pick up orders from lunchtime*. Available at: https://www.thesun.co.uk/living/3031582/tesco-

extends-same-day-click-and-collect-service-so-shoppers-can-pick-up-orders-from-lunchtime/ (Accessed: 6 July 2019).

The Sun, 2019. *Lidl 'to lunch online delivery service' allowing customers finally to shop from home.* Available at: https://www.thesun.co.uk/money/10097535/lidl-online-delivery-service-shop-from-home/ (Accessed: 6 July 2019).

The Telegraph, 2015. *Sainsbury's fashion clothing empire*. Available at: https://www.telegraph.co.uk/finance/newsbysector/retailandconsumer/11804755/SUN-Inside-Sainsburys-fashion-clothing-empire.html (Accessed: 6 July 2019).

The Telegraph, 2018. *Retailers bracing themselves for last minute delivery chaos after customers complain of parcels arriving three days late*. Available at: https://www.telegraph.co.uk/news/2018/12/23/retailers-bracing-last-minute-delivery-chaos-customers-complain/ (Accessed: 6 July 2019).

Thirumalai, S., Sinha, K.K., 2005. Customer satisfaction with order fulfilment in retail supply chains: implications of product type in electronic B2C transactions. *Journal of Operations Management*, 23(3–4), 291–303.

Torabi, S.A., Hassini, E., Jeihoonian, M., 2015. Fulfilment source allocation, inventory transshipment, and customer order transfer in e-tailing. *Transportation Research Part E: Logistics and Transportation Review*, 79, 128–144.

Tsao, Y.C., Mangotra, D., Lu, J.C., Dong, M., 2012. A continuous approximatation approach for the integrted facility-inventory allocation problem. *European Journal of Operational Research*, 222(2012), 216-228.

Tuclothing.sainsburys.co.uk. *Delivery information*. Available at: https://tuclothing.sainsburys.co.uk/help/deliveryinformation#deloptions (Accessed: 30 October 2019).

Van Baal, S., 2014. Should retailers harmonize marketing variables across their distribution channels? An investigation of cross-channel effects in multi-channel retailing. *Journal of Retailing and Consumer Services*, 21(6), 1038–1046.

Van Roy, T. J., Erlenkotter, D., 1982. A Dual-Based Procedure for Dynamic Facility Location. *Management Science*, 28(10), 1090-1213.

Varsei, M., Polyakovskiy, S., 2017. Sustainable supply chain network design: A case of the wine industry in Australia. *Omega*, 66, 236-247.

Verhoef, P.C., Kannan, P.K., Inman, J.J., 2015. From Multi-Channel Retailing to Omni-Channel Retailing. Introduction to the Special Issue on Multi-Channel Retailing. *Journal of Retailing*, 91(2), 174–181.

Verhoef, P. C., Neslin, S. A., Vroomen, B., 2007. Multichannel customer management: Understanding the research-shopper phenomenon. *International Journal of Research in Marketing*, 24(2), 129–148.

Waitrose, 2019. *Waitrose & Partner announces plans to treble the size of its online operations*. Available at: https://waitrose.pressarea.com/news/16052019/waitrose-partners-announces-plans-to-treble-the-size-of-its-online-operations (Accessed: 30 October 2019).

Wang, X., Lim, M. K., Ouyang, Y., 2017. A Continuum Approximation Approach to the Dynamic Facility Location Problem in a Growing Market. *Transportation Science*, 51(1), 343–357.

Waltho, C., Elhedhli, S., Gzara, F., 2018. Green supply chain network design: a review focused on policy adoption and emission quantification. *International Journal of Production Economics*, 208, 305–318.

Warehouse & Logistics News, 2017. *KNAPP helps John Lewis add capacity through automation in Magna Park 2*. Available at: http://warehousenews.co.uk/2017/02/knapp-helps-john-lewis-add-capacity-through-automation-in-magna-park-2-2/ (Accessed: 30 October 2019).

Wei, L., Jasin, S., Kapuscinski, R., 2017. Shipping Consolidation with Delivery Deadline and Expedited Shipment Options. *Ross School of Business Paper No.1375*.

Weltevreden, J.W.J., 2007. Substitution or complementarity? How the Internet changes city centre shopping. *Journal of Retailing and Consumer Services*, 14(3), 192–207.

WHSmith. *Order Delivery and Collection*. Available at: https://www.whsmith.co.uk/help/shopping-with-whsmith/order-delivery-and-collection/hel00009/ (Accessed: 6 July 2019).

WH Smith PLC, 2018. *Annual report and accounts 2018*.

Wulfraat, M., 2014. *Case study: Tesco's dot-com U.K. business model and lessons learned*. Available at: http://www.canadiangrocer.com/blog/e-grocery-test-study-tesco%E2%80%99s-dot-com-u-k-business-model-and-lessons-learned-45135 (Accessed: 6 January 2020).

Xia, Y., Zhang, G.P., 2010. The impact of the online channel on retailers' performances: An empirical evaluation. *Decision Sciences*, 41(3), 517–546.

Xing, Y., Grant, D.B., McKinnon, A.C., Fernie, J., 2010. Physical distribution service quality in online retailing. *International Journal of Physical Distribution & Logistics Management*, 40(5), 415-432.

Xu, H., Gong, Y., Chu, C., Zhang, J., 2017. Dynamic lot-sizing models for retailers with online channels. *International Journal of Production Economics*, 183(2017), 171–184.

Xu, J., Cao, L., 2019. Optimal in-store inventory policy for omnichannel retailers in franchising networks. *International Journal of Retail & Distribution Management,* 47(12), 1251-1265.

Xu, P.J., Allgor, R., Graves, S.C., 2009. Benefits of Reevaluating Real-Time Order

Fulfilment Decisions. *Manufacturing & Service Operations Management*, 11(2), 340–355.

Yan, T., Rabinovich, E., Dooley, K., Evers, P.T., 2010. Managing backlog variation in order fulfilment: The case of Internet retailers. *International Journal of Production Economics*, 128(1), 261–268.

Yang, J.Q., Zhang, X.M., Fu, H.Y., Liu, C., 2017. Inventory competition in a dual-channel supply chain with delivery lead time consideration. *Applied Mathematical Modelling*, 42(2017), 675–692.

Yang, X., Strauss, A.K., Currie, C.S.M., Eglese, R., 2014. Choice-Based Demand Management and Vehicle Routing in E-Fulfilment. *Transportation Science*, 50(2), 473–488.

Yang, X., Strauss, A.K., 2017. An approximate dynamic programming approach to attended home delivery management. *European Journal of Operational Research*, 263(3), 935–945.

Yao, D.Q., Yue, X., Mukhopadhyay, S.K., Wang, Z., 2009. Strategic inventory deployment for retail and e-tail stores. *Omega*, 37(3), 646–658.

Zahiri, B., Zhuang, J., Mohammadi, M., 2017. Toward an integrated sustainableresilient supply chain: a pharmaceutical case study. *Transport Research Part E: Logistics and Transportation Review*, 103, 109–142.

Zhang, J., Farris, P. W., Irvin, J. W., Kushwaha, T., Steenburgh, T. J., Weitz, B. A., 2010. Crafting integrated multichannel retailing strategies. *Journal of Interactive Marketing*. 24(2), 168–180.

Zhang, J., Onal, S., Das, R., 2019a. Fulfilment time performance of online retailers - an empirical analysis. *International Journal of Retail & Distribution Management*, 47(5), 493-510.

Zhang, J., Xu, Q., He, Y., 2018. Omnichannel retail operations with consumer returns and order cancellation. *Transportation Research Part E: Logistics and Transportation Review*, 118, 308–324.

Zhang, Y., Sun, L., Hu, X., Zhao, C., 2019b. Order consolidation for the last-mile split delivery in online retailing. *Transportation Research Part E: Logistics and Transportation Review*, 122, 309–327.

Zhao, F., Wu, D., Liang, L., Dolgui, A., 2016. Lateral inventory transshipment problem in online-to-offline supply chain. *International Journal of Production Research*, 54(7), 1951–1963.

Zhao, H., Deshpande, V., Ryan, J.K., 2005. Inventory Sharing and Rationing in Decentralized Dealer Networks. *Management Science*, 51(4), 531–547.