**New England University Transportation Center**
77 Massachusetts Avenue, E40-279
Cambridge, MA 02139
utc.mit.edu

# Year 25 Final Report

## Grant Number: DTRT13-G-UTC31

**Project Title:**

# Urban Last-Mile Transportation 4.0

| Project Number: | Project End Date: | Submission Date: |
|---|---|---|
| MITR25-52 | 6/30/19 | 8/7/19 |



| | | | |
|---|---|---|---|
| **Principal Investigator:** | Dr. Matthias Winkenbach | **Co-Principal Investigator:** | Prof. Yossi Sheffi |
| **Title:** | Research Scientist | **Title:** | Professor of Systems Engineering |
| **University:** | MIT | **University:** | MIT |
| **Email:** | mwinkenb@mit.edu | **Email:** | sheffi@mit.edu |
| **Phone:** | (857) 253-1639 | **Phone:** | (617) 253-5316 |

**Final Report**

# Urban Last-Mile Transportation 4.0

Matthias Winkenbach, Yossi Sheffi and Xavier Lavenir[1]

*Center for Transportation & Logistics, Massachusetts Institute of Technology*

## Disclaimer

*For an extensive report on the work conducted under this project, the literature reviewed, the technical details of the methodology employed, and an in-depth discussion of the results obtained, we refer the reader to Lavenir (2019) and Lavenir and Winkenbach (2019). This document is a collection of selected sections from Lavenir (2019) and Lavenir and Winkenbach (2019), summarizing our work and some preliminary findings at a non-technical level.*

## Abstract

With the rise of the on-demand economy and the increasingly high service levels expected by consumers, companies are racing to provide shorter and shorter delivery lead times for e-commerce orders. Two-day deliveries have become the norm in dense urban areas and companies are developing highly responsive urban distribution networks capable of serving consumers in lead times as low as one hour. In this work, we explore the strategic design of highly responsive urban distribution networks, promising lead times of under two hours, and investigate which operational and environmental parameters drive the deployment of different types of network designs. Furthermore, we evaluate the environmental footprint of these networks by measuring their contribution to congestion and $CO_2$ and $NO_x$ emissions, and contrast these with the environmental footprint of next-day delivery distribution networks. From a policy perspective, we investigate how the strategic design and performance of highly responsive networks are affected by an urban congestion charge policy and by different levels of government subsidy for last-mile logistics infrastructure. We conduct a case study inspired by the operations of a global fashion company in Manhattan, New York. Our preliminary findings indicate that highly responsive networks can be sustainable if designed under favorable operational conditions.

## 1. Introduction

The demand for goods and services to be brought into and out of cities is growing at a fast rate due to two major phenomena. First, urbanization is happening at a rapid pace on a global scale. Currently, around 55% of the world's population lives in urban areas and it is expected to

[1] X. Lavenir is a Research Assistant at the MIT Megacity Logistics Lab under the supervision of Dr. Winkenbach

increase to 68% by 2050 (United Nations, 2018). Second, we observe a constant increase in per-capita amount and fragmentation of demand for business-to-consumer (B2C) deliveries. To compete with the instant gratification shopping experience provided by brick-and-mortar stores, e-retailers have continued to promise faster delivery times to customers (Howard, 2014). Two-day shipping has become the norm in dense urban areas and market leaders are pushing for near-instant deliveries with lead time as low as one hour.

Companies wishing to provide a highly responsive service to their customers, with lead times under two hours, have been testing a myriad of innovative solutions to meet these rising service level expectations and to keep delivery costs to a minimum. E-retailers with no physical store presence can take advantage of the nascent concept of mobile depots to get inventory closer to consumer demand. A mobile depot is a mobile warehouse, such as a truck fitted with warehousing facilities and a loading dock, which can relocate throughout the day to get closer to consumer demand and serve as a hyper-local fulfilment hub. This concept was tested by TNT Express in 2013 (Verlinde, Macharis, Milan, & Kin, 2014). The burgeoning sharing economy has also had a profound impact on last-mile logistics and particularly on highly responsive delivery services. Rather than maintaining their own fleet of couriers, companies can partner with on-demand logistics providers to outsource the last-mile portion of the delivery (Savelsbergh & Van Woensel, 2016).

In this work, we explore different strategic designs of highly responsive urban distribution networks (i.e., networks which promise lead times of under two hours), and investigate which operational and environmental parameters drive the deployment of different types of strategic designs of these networks. In particular, we investigate how adding increasing levels of flexibility affects the performance of highly responsive systems by contrasting two types of network designs. The first leverages an omnichannel retailer's existing retail stores to provide highly responsive lead times in a ship-from-store setup. The second provides the same lead times using a fleet of mobile depots which relocate throughout the day and act as hyper-local fulfilment hubs. Furthermore, we investigate the environmental footprint of each network design; namely, we measure the total $CO_2$ and $NO_x$ emissions and congestion footprint as a direct result of last-mile logistics activities and contrast this with the environmental footprint of a next-day delivery distribution network. In terms of environmental and operational parameters, we vary the following parameters to understand their effect on the network's performance and environmental footprint: courier type (e.g., bicycle or vehicle), promised lead time (e.g., 90 or 120 mins), demand pattern (e.g., uniform or high spatial variation), and allowing or not allowing couriers to consolidate orders in a single trip. Finally, we investigate how a congestion charge policy and how different levels of government subsidy for last-mile logistics infrastructure affect the strategic network design of highly responsive systems and the environmental footprint of the network.

## 2. Research Approach

Existing methods for the strategic design of traditional distribution networks, such as for the day-before planning problem, propose network designs (facility locations, inventory, vehicle capacity) by assuming deterministic and static consumer demand (Crainic, Ricciardi, & Storchi,

2004). However, such methods cannot be applied when designing highly responsive distribution networks as the critical assumption of static and deterministic demand no longer applies. These current methods used to solve traditional network design problems become intractable when capturing the time-dependency and non-linearity of the highly responsive distribution network. Therefore, we propose a simulation-optimization model with problem-specific feedback. There are three key components to the overall model: a two-stage stochastic optimization model, a state-of-the-art discrete-event simulation model, and a feedback model. The two-stage stochastic optimization model first proposes a strategic network design by solving a simplified form of the problem. The optimization model makes key approximations and does not model the complex non-linear dynamics between agents in the highly responsive system – this renders the problem tractable. The second component is a discrete-event simulation model which simulates all agents involved in last-mile distribution (e.g., couriers, facilities, employees, mobile depots). State-of-the-art simulation models can capture the complex non-linear operations involved in last-mile distribution networks, such as couriers, facilities, employees, and mobile depots, which the optimization model is not capable of doing without becoming intractable. The feedback model is used to find discrepancies between the input parameters of the optimization model and the actual parameter values, as measured by the simulation model. These discrepancies arise because the optimization model makes certain assumptions about the input parameters, such as the picking times at facilities.

To get a baseline on the performance and environment footprint of the highly responsive network, we develop a simple next-day delivery model whereby customer orders are served from a distribution center. Namely, we assume that customer demand is served by a fleet of trucks and solve a capacitated vehicle routing problem (VRP) to get the performance and environmental footprint of this base-line next-day delivery network.

## 3. Methodology

*This section summarizes our methodological approach at a non-technical level. For an in-depth technical documentation of the developed models, their mathematical formulation, and algorithmic implementation, we refer the reader to Lavenir (2019).*

### 3.1. Modeling Highly-responsive Delivery Networks

To solve the highly responsive network design problem, we propose a simulation-optimization model with problem-specific feedback. On a high-level, the model can be viewed as a black box where inputs, such as a different consumer demand cases or facility parameters, are fed into the model and the model proposes a highly responsive network design which can serve consumers in the specified lead times. This model overview is shown in Figure 1 which shows different input parameters, the various sub-models within the “black box”, and the strategic network proposed by the model.

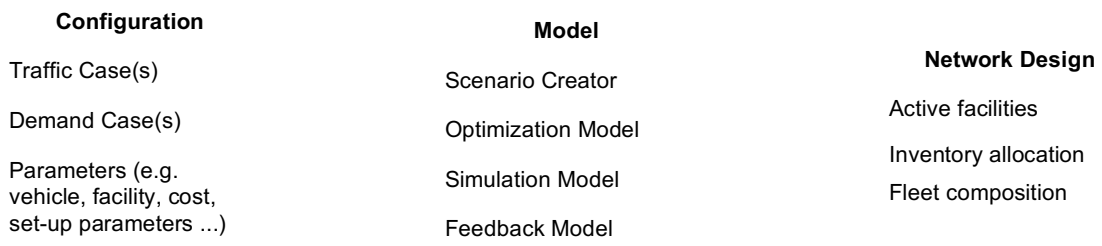


**Figure 1**: High-level model overview

The model is made up of four main components: the optimization model, the simulation model, the feedback model, and the final network picker. The purpose of each of these models is outlined below:

- **Optimization model** - The stochastic optimization model proposes a strategic network design for the highly responsive delivery network.
- **Simulation model** - The simulation model simulates the network performance under particular demand and traffic conditions.
- **Feedback Model** - The feedback model evaluates how the network design performed on the set of simulated scenarios. It analyses the simulation model output data and saves KPIs as well as feedback parameters. Feedback parameters are used to adjust the optimization input parameters on the next iteration.
- **Final Network Picker** - The final network picker analyses all previous proposed strategic network designs and their performances, as captured by the simulation model, and chooses the final network design based on a minimum cost.

*Optimization Model*

Due to the complexity and non-linearity of the highly response urban network design problem, the optimization model proposes a strategic network design based on an approximation of the actual last-mile operations. Namely, we introduce two key model approximations: the temporal- and spatial-aggregation of consumer demand, and the temporal and spatial-aggregation of vehicle capacity. To perform such approximations, we first discretize both space and time into discrete segments. In particular, we discretize the geography into discrete areas, referred to as *pixels,* and time into discrete *time-segments*.

The optimization model proposes a strategic network design for the last-mile network by deploying a two-stage stochastic analytical mixed-integer linear program. The two-stage approach captures the temporal hierarchy of decision-making: first, we make strategic infrastructure decisions under uncertainty (i.e., not knowing the realization of demand), and, second, we make operational decisions, such as the allocation of demand to facilities, once we have observed demand (i.e., a scenario). We capture the inherent uncertainty involved in the

strategic decision by using the Sample Average Approximation, wherein we approximate the original distribution by sampling over multiple scenarios.

The optimization model makes the following strategic network design decisions:

1. *Fixed facility* - Which brownfield facilities should be activated?
2. *Mobile Depots* - How many mobile depots should be activated and how should they re-position themselves throughout the day?
3. *Inventory* - How much inventory should be allocated to mobile depots and fixed facilities?
4. *Stations* - Which dedicated parking bay infrastructure should be invested in for mobile depot use?
5. *Employees* - How many employees should be hired at each fixed facility?

The optimization model approximates the following operational decisions:

1. *On-demand couriers* - How many on-demand couriers of each type should be summoned?
2. *Order allocation* - How are orders allocated to facilities (mobile depots or fixed facilities) and to on-demand couriers?

*Simulation Model*

The simulation model aims to precisely replicate the last-mile operations of the highly responsive delivery service proposed by the optimization model for a given day of operations. More precisely, we simulate the operations within a city over the course of a single day for a set of given demand and traffic cases. There is no intra-day replenishment at facilities / mobile depots and we ignore any operations occurring at the distribution center or further upstream. The simulation model is implemented as a discrete-event model which models the last-mile operations as a discrete sequence of events in time.

A brief description of the main simulation model components is given:

- *Order Generation* - Although the simulation model is provided with a detailed description of each consumer order that arrives over the course of the simulated day (e.g., location, stock keeping units (SKUs), lead time), the system has no advanced knowledge of any of these orders. This Order Generation component is responsible for creating this consumer demand.
- *Order Allocation* - Once a consumer places an order, it needs to be assigned to facilities and couriers which will prepare and deliver the package(s) to the consumer. This is one of the most complicated parts of the simulation model as there are many different ways to

deliver a package to a consumer. On a high-level, the goal of the Order Allocation module is to get the package(s) to the consumer in the fastest and cheapest way.

- *Courier* - Two types of couriers are modelled in the simulation model (bicycle and passenger vehicles) which are both assumed to be on-demand couriers. On-demand couriers represent the “gig economy" where a courier can be summoned for a single trip and doesn’t have to be employed for a fixed period of time. Their responsibility is to deliver packages from a facility to a consumer.

- *Facilities* - We model two types of facilities: fixed retail facilities, called fixed facilities, and mobile depots. Mobile depots extend the fixed store facility as they are mobile trucks and can relocate to different positions throughout the day. At each facility, we model the picking process that we assume would occur in the highly responsive delivery system.

*Feedback Model*

The feedback model evaluates how the network design performed on the set of simulated scenarios. It analyses the simulation model output data and saves KPIs as well as feedback parameters. Feedback parameters are used to adjust the optimization model input parameters on the next iteration.

*Final Network Picker*

The final network picker is a model that runs after all iterations are complete. It iterates over all the strategic highly responsive network designs proposed by the optimization model and selects the network design that minimizes the total network cost. The total network cost of a proposed network design is made up of three components: the strategic cost (e.g., investment in facilities, mobile depots, employees), the operational component (e.g., on-demand courier cost), and a penalty cost that is given for each late order.

### 3.2. Modeling Next-day Delivery Operations

To benchmark the performance and environmental footprint (congestion and emissions footprint) of the highly responsive network designs against the current state of e-commerce distribution networks, we build a baseline model which measures the performance and environment impact of a next-day distribution network. We measure the performance and environmental footprint of the next-day distribution network by solving a vehicle routing problem given some realization of customer demand. The direct output of the model is:

1. The number of freight vehicles required to serve customer demand

2. The routes taken by freight vehicles to serve customer demand

Given these outputs, we derive the congestion footprint and total $CO_2$ and $NO_x$ emissions as a direct result of the last-mile operations of these trucks within the city.

The routing problem is solved according to a variant of the savings algorithm proposed by Clarke and Wright (1964). Particularly, we leverage an extension of the savings algorithm called the *improved savings algorithm* implemented according to Crepy et al. (2019). The *improved savings algorithm* is a semi-greedy algorithm which constructs feasible truck routes which serve all consumer demand; the objective function is to minimize total routing costs.

### 3.3. Modeling Pollutant Emissions

An emissions model is needed to model the emissions footprint of the proposed last-mile logistics networks. To ensure fair comparison when contrasting the emissions footprint of different highly responsive networks with one-another and with the next-day network, only the contributions of the last-mile logistics operations are measured. Since there are large operational differences between the next-day network, where large trucks deliver orders directly from the distribution center, and the highly responsive network, where couriers deliver orders from hyper-local fulfillment centers, the emissions footprint of a network is measured as the emissions footprint resulting solely from vehicles involved in last-mile delivery. The emissions footprint of infrastructure which supports last-mile logistics operations, such as facilities, distribution centers, or the movement of vehicles not directly related to last-mile delivery, is not measured.

In this work, the Handbook Emission Factors for Road Transport (HBEFA) Version 3.3 (Keller, Hausberger, Matzer, Wüthrich, & Notter, 2017) is chosen to model the emissions of the last-mile logistics vehicles. HBEFA is a Microsoft Access database which provides emissions factors for select vehicle types under different traffic and environmental conditions. HBEFA was originally developed for the Environmental Protection Agencies of Germany, Switzerland, and Austria, and is now further supported by Sweden, Norway, and France. The emissions factors provided by the model were collected either from data measured in labs or from simulations (Keller et al., 2017). Emission factors are provided in g/km, i.e., the quantity of pollutant emitted when a vehicle travels 1 km under the specified traffic conditions.

### 3.4. Modeling Congestion

A congestion model is needed to model the congestion footprint of the highly responsive network and next-day network designs. As a proxy for congestion, two performance indices are measured: vehicle kilometers travelled (VKT) and road space usage, which we simply call congestion. VKT is the total distance travelled by vehicles involved in last-mile delivery, regardless of vehicle type; i.e., distance travelled by couriers and mobile depots. Road space usage is the total road space required by vehicles involved in last-mile delivery; road space usage captures the actual space used by vehicles and is a function of a vehicles' speed, size, and total distance travelled.

### 3.5. Modeling the Imposition of a Congestion Charge

The overuse of roads by vehicles places a burden on cities' scarce and difficult-to-expand infrastructure. Traffic congestion is a major financial cost for cities and urban passengers as it increases travel uncertainty and leads to longer travel times. Since it is difficult to expand the capacity of cities' surface transportation infrastructure, policymakers have proposed a demand-side solution to reduce the number of vehicle passengers in the form of a congestion charge. A congestion charge forces road users to internalize the cost of the negative externality that they impose on others which influences the demand for road-space by making passengers shift to different modes of transport or to travel at different times. We evaluate the impact of a congestion charge policy on the strategic design of the highly responsive networks. We assume that the policy takes the form of an additional a per-km surcharge for each road user involved in the last-mile operations; this flat per-km surcharge pertains to whole of Manhattan and is in effect at all times.

### 3.6. Implementation

All our models were fully implemented in Python 3.6. The optimization model leverages the Gurobi Optimization solver (Gurobi Optimization, 2019). The Simulation model was implemented using the SimPy library (SimPy, 2013).

## 4. Case Study

In a real-world case study, we build on data from a major global fashion company to design a highly responsive network to allow the company to provide their consumers in Manhattan, New York, with highly responsive delivery services via their online channel (e.g., 90 minutes) for select SKUs. The partner company has six retail stores in New York City. These stores are located within or near Manhattan and can be used as hyper-local fulfillment hubs to serve consumers within the target demand region.

### 4.1. Demand Cases

A highly responsive network design is highly dependent on the demand distribution–called a demand case–which it is designed for. A demand case describes the distribution of demand for a particular type of day. We explore how the design and performance of highly responsive networks change with different types of demand. In particular, we design strategic networks for a standard consumer demand case and for a variable consumer demand case.

*Standard Demand Case*

The standard demand case represents the typical demand distribution in Manhattan. Figure 2 shows how the spatial distribution of demand changes from 08:00 to 22:00 in two-hour

increments. A few patterns emerge: first, the region north of central park always has low demand densities; and second, the area just below central park, called Midtown Manhattan, and the area at the bottom of Manhattan, called the Financial District, tend to always have relatively high demand densities.

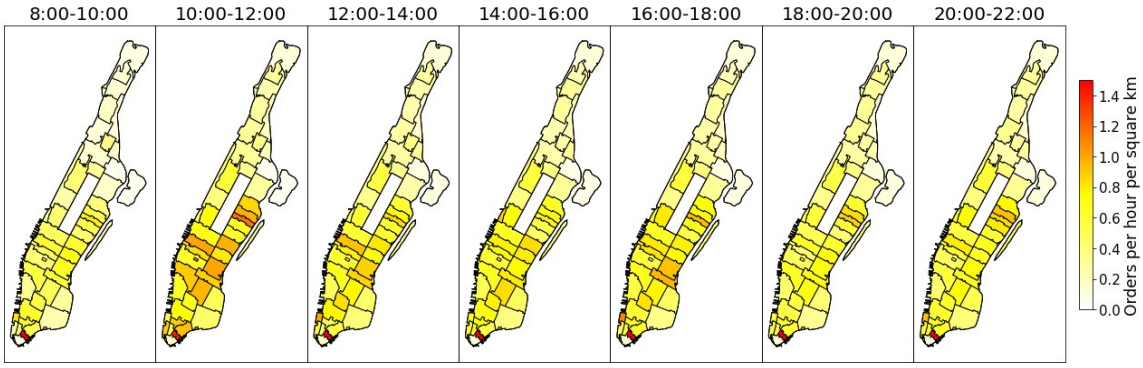


**Figure 2**: Standard demand case

*Variable Demand Case*

The variable demand case is an artificial demand case, where, for each time-segment, we aggregate the total demand occurring across the Manhattan island and we force a particular spatial demand distribution where there is a sub-region with artificially high demand density. For each consecutive time-segment, we shift the sub-region with the highest consumer density, so that the "hot spot" of demand shifts up and down the Manhattan island as they day progresses – this is shown in Figure 3. We create this artificial demand case, so that there is a clear "hot spot" of demand which shifts location throughout the day, and to investigate how a flexible facility network design, where mobile depots are allowed to relocate throughout the day, responds to such demand.

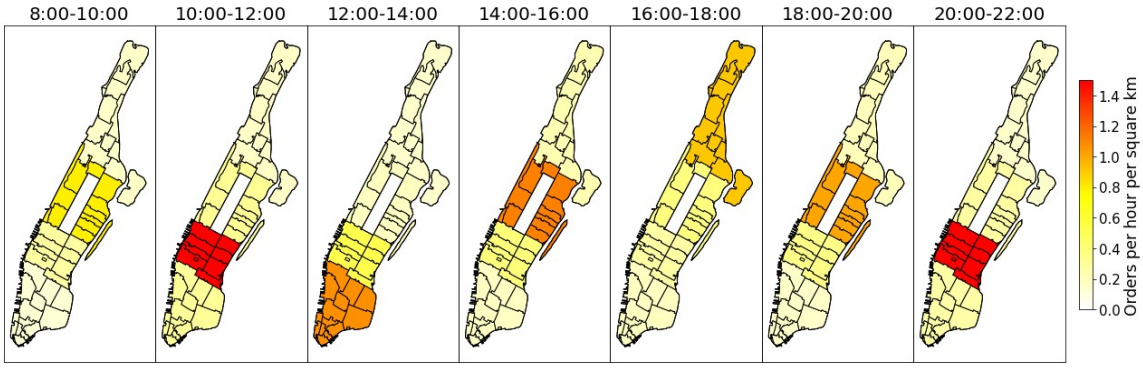


**Figure 3:** Variable demand case

### 4.2. Travel Cases

The highly responsive network design is also highly dependent on the time it takes for road users, such as couriers and mobile depots, to travel around Manhattan. To capture this travel time information, we introduce the concept of a travel case. A travel case fully defines the time it takes to travel around a particular city on a particular day, and is a function of the road user type, trip start location, trip end location, and time of day. Travel times can vary enormously day-to-day and multiple travel cases can be created to capture travel times on different days; e.g., travel times in Manhattan on UN day, a notoriously bad day, versus travel times on a typical weekday.

To obtain realistic travel times, we query the Google Distance Matrix API (Google, 2017) which returns the travel time between an origin and destination pair at a particular time, using a particular travel mode.

## 5. Key Findings

Some preliminary key findings arise from our model-based case study analysis:

1. Highly responsive orders are more expensive than next-day delivery orders, but consolidation reduces this price differential.

2. Highly responsive networks generally have a significantly worse congestion and emissions footprint relative to next-day delivery networks, but, under certain operational modes, they have a smaller impact.

3. Increased levels of flexibility in highly responsive network designs results in better overall performance.

4. Consolidation and bicycles as last-mile couriers are critical operational parameters for a sustainable highly responsive network.

5. The congestion charge policy changes the strategic design of the flexible-facility highly responsive network, but has a marginal effect on the network's congestion footprint and greatly increases the total network cost.

In summary, our study suggests that an urban freight policy aimed at reducing congestion, and also emissions, should focus on incentivizing companies to use bicycles as last-mile couriers and improving their operations such that they can consolidate customer orders. Imposing a simple congestion charge on urban freight operations does most likely not lead to the desired improvements.

**References**


Clarke, G. & Wright, J. W. (1964). Scheduling of Vehicles from a Central Depot to a Number of Delivery Points. *Operations Research*, *12*(4), 568–581

Crainic, T. G., Ricciardi, N., & Storchi, G. (2004). Advanced freight transportation systems for congested urban areas. *Transportation Research Part C: Emerging Technologies*, *12*(2), 119–137

Crepy, M., Poullet, J., Moshref-Javadi, M., & Winkenbach, M. (2019). *Quantitative Modeling of Alternative Last-mile Delivery Systems*. MIT CTL Working Paper.

Google. (2017). Google Distance Matrix API. Retrieved from https://developers.google.com/maps/documentation/distance-matrix/

Gurobi Optimization, L. (2019). Gurobi Optimizer Reference Manual. Retrieved from http : //www.gurobi.com

Howard, R. (2014). Same-Day Delivery: A Checklist for Retailers Seeking an Antidote to Amazon: Part One. Retrieved from https://parcelindustry.com/article-4003-Same-Day-Delivery-A-Checklist-for-Retailers-Seeking-an-Antidote-to-Amazon-Part-One.html

Keller, M., Hausberger, S., Matzer, C., Wüthrich, P., & Notter, B. (2017). *HBEFA Version 3.3 Hintergrundbericht* (tech. rep. No. April). MKC Consulting GmbH.

Lavenir, X. (2019). *"The Strategic Design and Environmental Footprint of Highly Responsive Urban Distribution Networks"*, Master's Thesis, Massachusetts Institute of Technology, 2019.

Lavenir, X. and Winkenbach, M. (2019). A simulation-based optimization approach to the design and performance evaluation of sustainable highly responsive urban logistics services. *MIT CTL Working Paper.*

Savelsbergh, M. & Van Woensel, T. (2016). 50th Anniversary Invited Article—City Logistics: Challenges and Opportunities. *Transportation Science*, *50*(2), 579–590

SimPy. (2013). SimPy Reference Manual. Retrieved from https://simpy.readthedocs.io/en/latest/contents.html

United Nations. (2018). *World Urbanization Prospects: The 2018 Revision*. United Nations Population Division.

Verlinde, S., Macharis, C., Milan, L., & Kin, B. (2014). Does a Mobile Depot Make Urban Deliveries Faster, More Sustainable and More Economically Viable: Results of a Pilot Test in Brussels. *Transportation Research Procedia*, *4*, 361–373