FRAMEWORK FOR ACCESS HUBS SYSTEM VEHICLE ROUTING WITH STOCHASTIC CUSTOMER DEMAND

Maharshi Dhada, Duncan McFarlane
Institute for Manufacturing, Department of Engineering, University of Cambridge, 17 Charles Babbage Road, Cambridge, UK

Abstract

Purpose: This paper presents a solution framework to a typical yet often encountered problem in last mile internal and external logistics. This is a case where a warehouse has a fleet of vehicles with limited capacities, and caters to customers in its neighbouring region with known locations. The customer demand is continuous, and each demand (package) is associated with weight, earliest start time, and latest arrival time.
Design/ methodology/ approach: This paper presents a solution framework to a typical yet often encountered problem in last mile internal and external logistics. This is a case where a warehouse has a fleet of vehicles with limited capacities, and caters to customers in its neighbouring region with known locations. The customer demand is continuous, and each demand (package) is associated with weight, earliest start time, and latest arrival time.
Findings: The findings presented in this paper are the results obtained using the proposed framework and a synthetic testing dataset approved by globally one of the largest shipping and logistics companies. The results show that the aforementioned problem can be solved using a framework relying on reinforcement learning technique that minimises the overall uncertainty while allocating and routing the fleet vehicles.
Originality/ value: This paper presents a solution framework to address the split delivery vehicle routing problem with stochastic and continuous customer demand. The paper is impactful as this is an often encountered problem in internal and external last mile logistics, and more so because the testing dataset used for obtaining the results presented in this paper is approved by one of the globally largest shipping and logistics service provider.
Keywords: Last mile logistics, vehicle routing, stochastic problem, probabilistic modelling, transportation

1 Introduction

There exist several systems for delivery vehicle routing, aiming to improve efficiency and reduce operational costs for their corresponding applications (Ghiani et al., 2022; Humes, 2016). The most prominent delivery systems include:

1. *Direct routing*, where vehicles deliver goods directly from warehouses to customers, minimizes stops but may increase the total distance travelled.
2. *The access hubs system*, which by contrast to the direct routing system, consolidates goods at central hubs before distributing them, reducing travel distances and optimizing delivery in high-demand areas.
3. *Dynamic routing*, which allows for real-time adjustments based on traffic, weather, or demand, offering flexibility for last-mile delivery.
4. *Multi-depot routing*, that optimises the operations by distributing vehicles across several depots, while crowdsourced delivery leverages independent contractors to offer flexible, cost-effective distribution solutions.

This paper particularly deals with the *access hubs system* for routing delivery vehicles, which is increasingly popular in urban areas. The access hubs system is an advanced centralised logistics approach, designed to enhance the efficiency of the delivery vehicle routing by consolidating goods at centralised hubs before distribution to their corresponding final destinations. The hubs act as intermediate points where goods are aggregated, sorted, and then dispatched to their destinations, making it easier to manage high-demand areas and streamline last-mile delivery. This system plays a crucial role in reducing the complexity of direct deliveries, especially in urban environments and large-scale supply chains. A typical vehicle routing model with hubs can be formulated as a mixed-integer linear programming problem (Faugère et al., 2020, 2022).

This system finds applications in various industries, including e-commerce, retail, and manufacturing. In urban logistics, access hubs allow for efficient distribution in densely populated areas, where direct delivery to each customer may not be practical due to traffic congestion and limited parking. For large-scale supply chains, especially those involving multiple regions, access hubs enable more scalable operations by allowing companies to manage logistics through regional distribution centres. By routing vehicles through hubs, companies can optimise the delivery routes, thereby lowering the transportation costs, reduce delivery times, and minimize environmental impacts through more efficient use of vehicles(Faugère et al., 2020, 2022).

One of the primary advantages of the access hubs system is its scalability. As demand grows, more hubs can be added to the network without significantly increasing operational complexity. Furthermore, the system promotes sustainable logistics by reducing vehicle miles travelled, thus lowering fuel consumption and carbon emissions.

The applications and design of the access hubs system have been extensively explored for urban logistics applications. The literature highlights the efficiency gains from modular hub networks that facilitate the sharing of logistics resources across companies. For example, collaborative logistics hubs can significantly reduce delivery times and costs in e-commerce networks, and improve the sustainability of supply chains by optimizing the use of transportation assets and minimizing empty miles (Crainic et al., 2023). A schematic representation of an access hubs delivery vehicle system is shown in Figure 1. As shown in Figure 1, multiple hubs may sometimes deliver to the same hubs, often making this a split-delivery vehicle routing problem.

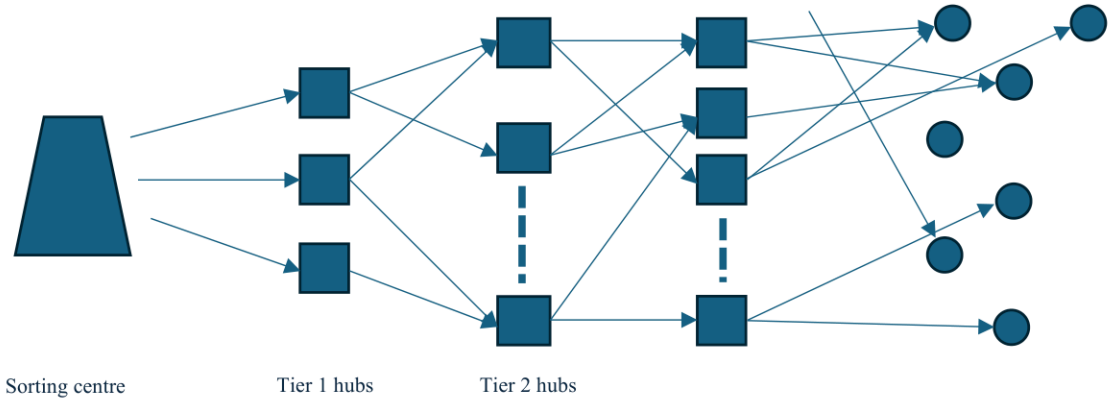

Figure 1. Schematic representation of an access hubs delivery vehicle system

Nonetheless, the access hubs systems are easily prone to propagations of a poor customer demand through the system. This is due to the fact that the access hubs delivery systems are largely serial systems, and therefore the inaccuracies in the customer demand compound as one goes higher up in the system. That is, away from the customer nodes (Liu et al., 2024).

More specifically, over-forecasting the customer demands in an access hubs delivery system leads to excess inventory at hubs, resulting in higher storage costs, inefficient use of transportation resources, and potential product spoilage. Under-forecasting, on the other hand, causes stock shortages at hubs, leading to delayed or incomplete deliveries, dissatisfied customers, and increased transportation costs due to emergency restocking. Both scenarios disrupt the balance of the hub-and-spoke model, reducing the efficiency of vehicle routing and increasing operational complexity. Ultimately, inaccurate forecasting undermines the system's ability to optimize last-mile delivery and meet customer expectations effectively.

This paper presents a basic simulation to demonstrate the impact of over and under forecasting the customer demand, on the overall cost of an access hubs delivery system for static customer demands. This simulation is explain and the results from the simulation are presented in Section 2. Section 3 presents a framework to enable the deployment of an access hubs delivery system by

incorporating the stochastic customer demand, especially for a split-delivery vehicle routing problem with continuous customer demand. Finally Section 4 summarises the key conclusions and future research directions.

2 Simulation to Demonstrate the Effect of Poor Customer Demand on the Access Hubs System

This section demonstrates a basic simulation to demonstrate the impact of a poor customer demand forecast on the overall cost of an access hubs delivery system for static customer demands.

Simulation Setup

A three-tier section of a bigger access hub system setup was considered for the purpose of this simulation, where a local hub tier 1 delivers to *20 local hubs* at the tier 2, and each local hub at tier 2 further delivers to *5 local hubs* each at tier 3. Each lower tier hub's (i.e. the hub next in the direction of the deliveries) demand would be forecasted by their corresponding higher tier hubs that serve them, and this demand would be relayed back to the further higher tier hubs and so on. The deliveries are then dispatched to fulfil the demands at corresponding lower tier hubs, who would then further fulfil their respective demands. A schematic of this is shown below in Figure 2.

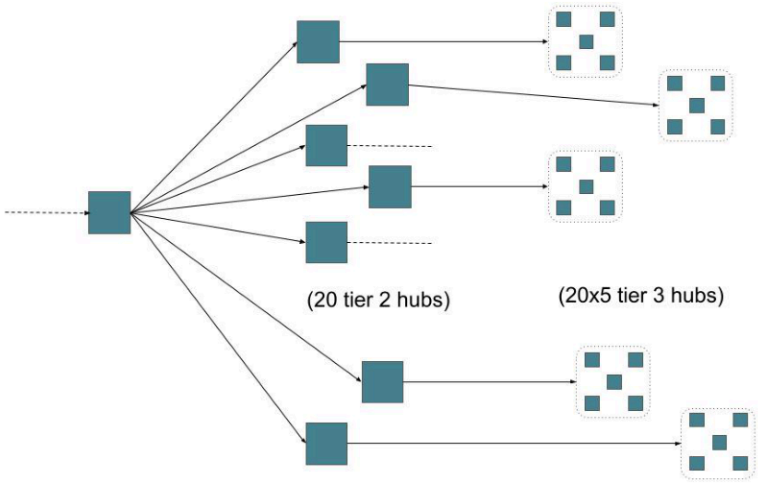

Figure 2. A schematic representation of the simulation, with 20 tier 2 hubs each delivering to 5 tier 3 hubs

For this section of the access hub that was simulated, a single delivery vehicle type with capacity of 50 packages was used for the simulation, and only one package type was considered. There were no time constraints for the deliveries. The demands for the lower tiers were generated as random integers between 10 and 50, and the demands for the higher tiers were the summation of the demands of their corresponding lower tiers. Total 1000 sets of demands for each of the lower tiers were generated, thereby representing 1000 simulation rounds for which the total delivery costs were calculated. Any given hub's demand was capped at 50, which is the capacity of the delivery vehicles. This is to avoid further complications in the routing algorithm. Nonetheless, there exist several heuristics to solve the routing problems where the customer demands are higher than the vehicle capacities (split delivery vehicle routing problems).

Each tier of the system was thus treated as a standard capacitated vehicle routing problem, which was solved using Google OR Tools Python library in this simulation. It should be noted that the goal of this simulation was to demonstrate the impact of a poor forecast in terms of the relative cost with respect to the case of accurate customer demand forecast. Therefore the routing algorithm at each tier does not impact the results significantly, and can be easily replaced with any other vehicle routing heuristic. The distances were kept consistent as the simulation focuses solely on the relative impact of the customer demand forecast of the system. The distance matrices used between the tier 2 to tier 3 hubs is shown below, with D resembling the location of the tier 2 hub and numbers 1 through 5 in bold resembling each of the tier 3 hubs which are being served:

	D	1	2	3	4	5
D	[0,	29,	20,	21,	17,	35]
1	[29,	0,	15,	29,	28,	40]
2	[20,	15,	0,	15,	14,	25]
3	[21,	29,	15,	0,	17,	30]
4	[17,	28,	14,	17,	0,	32]
5	[35,	40,	25,	30,	32,	0]

If the forecasted demand was higher than the actual demand, i.e. the case of over-forecasting, it would still be fulfilled by the previous hub in the system since the over-forecast was relayed to the previous hub earlier, but would end up as excess inventory at that hub in real life. This cost of increased inventory is not considered in this simulation. On the other hand, if the forecasted demand that was relayed to the previous tier hub is lower than the actual demand, the hubs that are served at the lower level would be removed from the route in the order of their decreasing total demands – so that the minimum number of hubs' demands are unfulfilled. In this simulation, the total number of hubs that are not fulfilled are counted at each level for this section of the access hubs, which also include the hubs served by the unfulfilled hub.

In summary, an over-forecast results in an increased travelling cost, whereas an under-forecast results in a number of unfulfilled customer demands. In this simulation, delivery costs associated with only the tier 3 hubs were considered while evaluating the impact of over-forecasting. However, while evaluating the impact of under-forecasting, demand unfulfillment at all the hubs are considered. Schematic representation for the case of an under-forecast is shown in Figure 3.

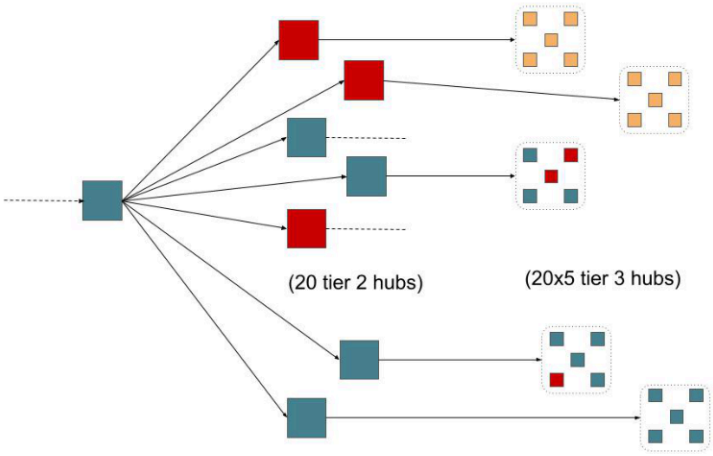

Figure 3. A schematic showing the under-fulfilled nodes in red. It is shown that if a tier 2 node is under-fulfilled, the corresponding tier 3 nodes are impacted as well

For the case of over-forecasting

To evaluate the impact of over-forecasting the customer demand, 1000 sets of baseline demands at each of the tier 3 hubs were generated. With total 20 tier 3 hubs, each delivering to 5 further hubs, a single demand set comprises of 20x5 numeric arrays. 1000 such arrays were randomly generated using integers between 10 and 50. Thereby representing 1000 simulation rounds, for which the total delivery costs were calculated. To evaluate the impact of over-forecasting, the demands corresponding to each of the baseline demands were systematically increased in steps of 2%, 5%, 10%, 30%, and 50%. This was to simulate that the forecasted demands are corresponding percentages higher than the actual baseline demand.

Total distance and number of vehicles used for each of these cases were evaluated across the 1000 demand sets, and the distributions over these 1000 sets are presented in Figures 4 and 5, for the distances and the number of vehicles used respectively. It is observed that the delivery costs exponentially increase with the increasing errors in the forecast. Moreover, the number of vehicles required for the deliveries also steeply increase as the error in the demand increases.

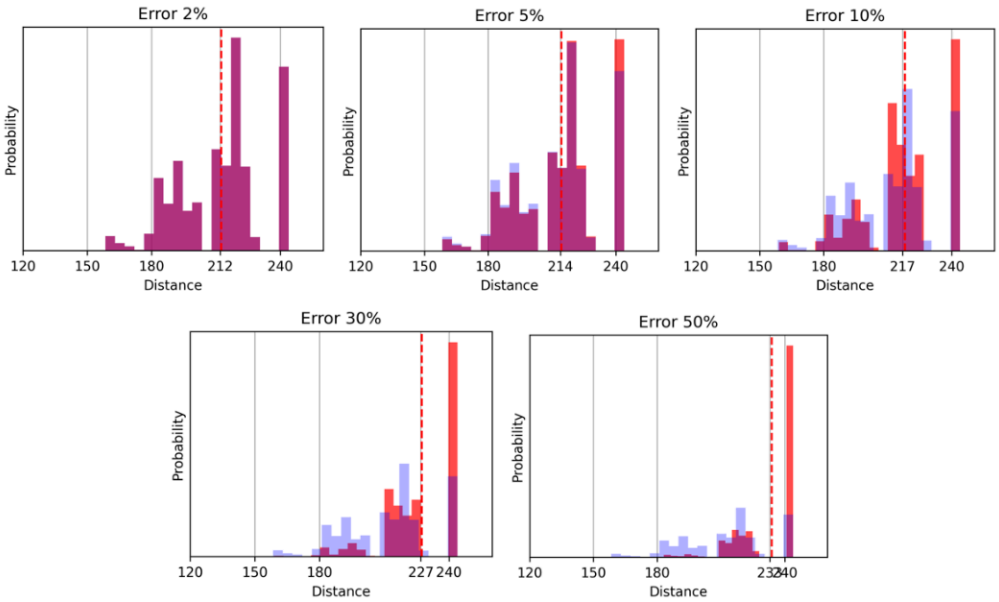

Figure 4. Increasing cost of deliver (shown in red) with respect to the baseline cost (shown in blue) as the forecasted demand is systematically increased

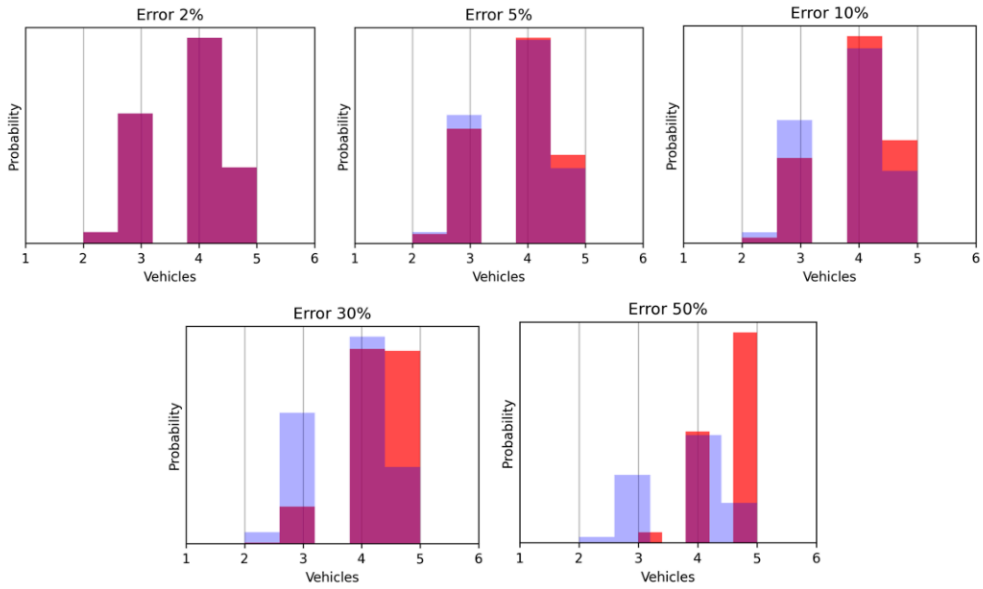

Figure 5. Increasing number of vehicles needed (shown in red) with respect to the baseline number of vehicles needed (shown in blue) as the forecasted demand is systematically increased

For the case of under-forecasting

To evaluate the impact of under-forecasting, 1000 sets of baseline demands at each of the tier 3 hubs were generated. This was similar to the case of over-forecasting simulation above. The forecasted demands for tier 2 hubs were the cumulative of the demands at their corresponding tier 3 hubs. To evaluate the impact of under-forecasting the demand, the baseline demand was systematically decreased in steps of 2%, 5%, 10%, 30%, and 50%. This was to simulate the fact that the forecasted demand was corresponding percentage lower than the actual baseline demand.

For each of the cases of under-forecasting, the number of under-fulfilled hubs were calculated. The hub chosen to be under-fulfilled would be the one with maximum demand, so that the number of under-fulfilled hubs are minimised. The actual impact of under-fulfilled hubs was also calculated based on the number of hubs the under-fulfilled hub delivers to. For example, the impact of one

under-fulfilled tier 2 hub results in under-fulfilment of five tier 3 hubs in this case. Table 1 presents the average number of tier 2 under-fulfilled hubs, resulting impact in the tier 3 hubs, and the number of under-fulfilled tier 3 hubs for each of the cases of under-forecasting the demand across 1000 simulation rounds. It can be observed in Table 1 that under-forecasting leads to a large number of under-fulfilled hubs, which is especially true for the case of an access hubs delivery system where the demand fulfilments of the hubs rely on the hubs before.

Table 1. Average number of under-fulfilled hubs at tiers 2 and 3

Under-forecast percentage	Tier 2 hubs under-fulfilled	Impact of the tier 2 nodes	Tier 3 hubs under-fulfilled
2	5	25	20
5	5	25	20
10	10	50	20
30	27	135	32
50	46	230	45

3 Proposed Framework for an access hubs system with continuous and stochastic demand

In Section 2 it is shown that an access hubs delivery system critically relies on demand forecast. Given the serial nature of the access hubs system, errors in demand can easily propagate through the logistics network leading to inefficient deliveries. This is especially challenging when the delivery network deals with a variety of old and new products, a fleet with a variety of vehicles associated with their own costs and capacities, and when the customer demand is continuous. A continuous customer demand refers to the case where the demand is received at every timestep, which means the vehicles must be loaded and routed continuously.

In all the above cases, it is critical to forecast the customer demands accurately. However for the particular cases of *cold-start* problems, for example when a new product is introduced or a new region is to be served where there exists limited historical data, the proposed statistical hierarchical modelling framework is deemed useful to *borrow* information from other *similar* components of the system. For example, a similar component for a newly introduced product can be a similar other product in the same category or in a complementary category, a similar component for a new region could be a similar other region already being served in terms of the demography, wealth, social behaviour, etc. Furthermore, the framework uses a Bayesian approach, allowing prior knowledge from higher levels to inform forecasts for the new hub, while gradually updating predictions as data accumulates. This ensures robust initial forecasts, improving hub performance from the beginning (Gelman et al., 1995).

Statistical hierarchical models, or multi-level models, have been conceptualised long ago and find applications in diverse applications such as infrastructure asset management, predictive maintenance, election results forecasting, or even burglary forecasting in urban areas (Bull et al., 2023; Davies & Bishop, 2013; Dhada et al., 2022; Gelman et al., 1995)! A hierarchical model is characterised by each individual (in this case a customer or a hub) being modelled with a forecasting algorithm, whereby the parameters of this forecasting algorithm are sampled from a higher level distribution. This higher level distribution is shared by *similar* components of the logistics network, as discussed above. In practice, when there is a lack of training data, the forecasts rely on the higher level distributions which represent in fact the prior information about the forecasts from similar other components.

By leveraging both granular and aggregated data, a hierarchical model can account for local variations while aligning with broader trends. This helps in more accurate predictions for each hub by sharing information across regions. The model reduces noise in smaller datasets, corrects biases, and incorporates temporal, spatial, and external factors, ultimately improving the forecast's precision and robustness, leading to more efficient resource allocation and delivery scheduling. For example, in the absence of sufficient historical data for a new hub, a hierarchical model leverages data from established hubs at higher levels in the hierarchy (regional or system-wide). This shared information helps in estimating demand patterns by recognising similarities in customer behaviours, geographic proximity, hub characteristics, etc. Figure 6 represents an example of a basic hierarchical model for a

newly introduced product, that relies on historical information from other similar products for improved forecasts.

4 Conclusions and Future Research Directions

This paper presented the impact of an inaccurate forecast for an access hubs delivery system. Demand forecast is especially important for an access hubs system due to its largely sequential or serial nature. In Figures 4 and 5 it is observed that the delivery costs steeply increase when there is an over-forecast of the demand, whereas an under-forecast renders several hubs/ customers under-fulfilled. The challenge with the demand under-forecasting is especially true for an access hubs system due to the sequential nature, where the hubs that are in the subsequent tiers are also impacted if a higher up hub is under-fulfilled. This is shown in Figure 3.

Moreover, a framework relying on statistical hierarchical modelling is also proposed in this paper as a solution for the case of continuous stochastic customer demand, which especially presents a real challenge for the access hubs system in the presence of sparse historical data.

There exist several future research directions stemming from this work, which include:

1. Conducting a sophisticated analysis of the impact of demand forecasting on the logistics network type. For example, defining the criticality of hubs based on the number of other hubs it is connected to, and studying the most important features that determine the impacts of an inaccurate forecast
2. Expanding beyond a standard capacitated vehicle routing problem, and into more realistic split delivery vehicle routing problems within the access hubs delivery system
3. Deploying the proposed hierarchical model framework for the case of an access hubs system in presence of a continuous and stochastic customer demand

5 Acknowledgements

This research was sponsored by the SF Express Research Fellowship in Logistics.

References

- Bull, L. A., Di Francesco, D., Dhada, M., Steinert, O., Lindgren, T., Parlikad, A. K., Duncan, A. B., & Girolami, M. (2023). Hierarchical Bayesian modeling for knowledge transfer across engineering fleets via multitask learning. *Computer-Aided Civil and Infrastructure Engineering*, *38*(7), 821–848.
- Crainic, T. G., Klibi, W., & Montreuil, B. (2023). Hyperconnected city logistics: a conceptual framework. In *Handbook on City Logistics and Urban Freight* (pp. 398–421). Edward Elgar Publishing.
- Davies, T. P., & Bishop, S. R. (2013). Modelling patterns of burglary on street networks. *Crime Science*, *2*, 1–14.
- Dhada, M., Bull, L. A., Girolami, M., & Parlikad, A. K. (2022). Modelling the Times-to-Failures Using a Statistical Hierarchical Model. *European Workshop on Structural Health Monitoring*, 985–994.
- Faugère, L., Klibi, W., White III, C., & Montreuil, B. (2022). Dynamic pooled capacity deployment for urban parcel logistics. *European Journal of Operational Research*, *303*(2), 650–667.
- Faugère, L., White III, C., & Montreuil, B. (2020). Mobile access hub deployment for urban parcel logistics. *Sustainability*, *12*(17), 7213.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.
- Ghiani, G., Laporte, G., & Musmanno, R. (2022). *Introduction to Logistics Systems Management: With Microsoft Excel and Python Examples*. John Wiley & Sons.
- Humes, E. (2016). Door to door: the magnificent, maddening, mysterious world of transportation. *(No Title)*.
- Liu, X., Li, J., & Montreuil, B. (2024). Logistics Hub Capacity Deployment in Hyperconnected Transportation Network Under Uncertainty. *ArXiv Preprint ArXiv:2402.06227*.