informs
http://pubsonline.informs.org/journal/ijoc

INFORMS JOURNAL ON COMPUTING
Articles in Advance, pp. 1–21
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

Iterative Prediction-and-Optimization for E-Logistics Distribution Network Design

Q:1 Q:2

Q:3 **Junming Liu,[a] Weiwei Chen,[b] Jingyuan Yang,[c] Hui Xiong,[d] Can Chen[d]**

Q:4 [a] Department of Information Systems, City University of Hong Kong, Hong Kong; [b] Department of Supply Chain Management, Rutgers University, New Jersey; [c] Information Systems and Operations Management, George Mason University, Virginia; [d] Management Science and Information System, Rutgers University, New Jersey

Contact: junmiliu@cityu.edu.hk, https://orcid.org/0000-0002-9301-0894 (JL); wchen@business.rutgers.edu, https://orcid.org/0000-0002-7736-3411 (WC); jyang53@gmu.edu, https://orcid.org/0000-0003-4852-2724 (JY); hxiong@rutgers.edu (HX); cc1063@rutgers.edu (CC)

Received: April 8, 2019
Revised: July 6, 2020; January 26, 2021; May 16, 2021
Accepted: June 4, 2021
Published Online in Articles in Advance: ■■■■ ■■, 2021

https://doi.org/10.1287/ijoc.2021.1107

Abstract. The emergence of online retailers has brought new opportunities to the design of their distribution networks. Notably, for online retailers that do not operate offline stores, their target customers are more sensitive to the quality of logistic services, such as delivery speed and reliability. This paper is motivated by a leading online retailer for cosmetic products on Taobao.com that aimed to improve its logistics efficiency by redesigning its centralized distribution network into a multilevel one. The multilevel distribution network consists of a layer of primary facilities to hold stocks from suppliers and transshipment and a layer of secondary facilities to provide last-mile delivery. There are two major challenges of designing such a facility network. First, online customers can respond significantly to the change of logistics efficiency with the redesigned network, thereby rendering the network optimized under the original demand distribution suboptimal. Second, because online retailers have relatively small sales volumes and are very flexible in choosing facility locations, the facility candidate set can be large, causing the facility location optimization challenging to solve. To this end, we propose an iterative prediction-and-optimization strategy for distribution network design. Specifically, we first develop an artificial neural network (ANN) to predict customer demands, factoring in the logistic service quality given the network and the city-level purchasing power based on demographic statistics. Then, a mixed integer linear programming (MILP) model is formulated to choose facility locations with minimum transportation, facility setup, and package processing costs. We further develop an efficient two-stage heuristic for computing high-quality solutions to the MILP model, featuring an agglomerative hierarchical clustering algorithm and an expectation and maximization algorithm. Subsequently, the ANN demand predictor and two-stage heuristic are integrated for iterative network design. Finally, using a real-world data set, we validate the demand prediction accuracy and demonstrate the mutual interdependence between the demand and network design.

Q:5 Q:6 **Summary of Contribution:** We propose an iterative prediction-and-optimization strategy for multilevel distribution network design for e-logistics and evaluate its operational value for online retailers. We address the issue of the interplay between distribution network design and the demand distribution using an iterative framework. Further, combining the idea in operational research and data mining, our paper provides an end-to-end solution that can provide accurate predictions of online sales distribution, subsequently solving large-scale optimization problems for distribution network design problems.

History: Accepted by Ram Ramesh, Area Editor for Knowledge Management and Machine Learning.
Supplemental Material: Data are available at https://doi.org/10.1287/ijoc.2021.1107.

Q:7 **Keywords:** facility location optimization • artificial neural network • heuristic • demand prediction

Q:8 1. Introduction

Q:9 Q:10 The emergence of the online market offers a convenient and efficient way for both customers and retailers. Retailers, especially small and medium ones, can display and sell products on online platforms without setting up offline retail stores, thus avoiding expensive store setup and operational costs. On the other hand, customers can conveniently browse and purchase products online without the hassle of visiting retail stores in person. Upon receiving an order, the retailer ships products directly from its warehouse to the customer.

To attract more online customers, online retailers have been focusing on optimizing the key factors that may affect online purchasing behaviors, including product quality, demographic factors, online shopping experience, and customer services (Bucko et al. 2018, Singh and Rana 2018). Among those key factors, the quality of logistics service is becoming increasingly important for online customers' satisfaction, which in turn significantly impacts their willingness to purchase from a specific online retailer (Lin 2019, Cui et al. 2020). For example, Taobao,[1] the largest online shopping platform in China, has specifically introduced a logistics service score to evaluate online retailers (Yu et al. 2015). Under such fierce competition, online retailers have been striking to shorten the shipping time (from the order validation until the order shipment) and delivery time (from the order shipment until the order delivery). For example, to attract more customers, *Amazon* has provided the two-day Prime Delivery service (Zhu and Liu 2018) and *Taobao* has provided one-day delivery service (Lin 2019). To meet such a high standard on the logistics speed and service quality, online retailers need to improve and optimize their distribution networks, particularly the network structure and facility locations. Typically, the goal of designing/redesigning a distribution network for e-logistics is to minimize the logistics costs while maintaining an acceptably short lead time and fast product delivery (Subramanian et al. 2014, Hübner et al. 2015).

Several major challenges exist for optimizing the distribution network for online retailers. First, it is essential to determine the desired physical distribution of online demand. Because online customers from different locations receive identical product information from an online retailer, one of the key factors that significantly affect online customer demand distribution is the quality of logistics service (Cui et al. 2020). However, the customer demand prediction, with the consideration of logistics service and the effects of a new distribution network, has not been investigated in the literature. Second, the network design is a large-scale optimization problem considering different types of logistics costs and constraints (Melo et al. 2009). The traditional optimization techniques do not scale very well when the number of nodes in the logistics network increases (Ortiz-Astorquiza et al. 2018). More importantly, the demand distribution for online retailers is more sensitive to the distribution network compared with their brick-and-mortar counterparts (Cui et al. 2020). Specifically, the brick-and-mortar stores can maintain a relatively stable customer demand to the stores as long as they retain sufficient inventory levels, regardless of the logistics behind these products. But the demand for online retailers responds more significantly to the logistics service quality, which is dependent on their distribution network. Therefore, as we optimize the location of facilities based on the demand predicted for the current distribution network, the retailers may observe a substantially modified demand distribution, which in turn renders the location of facilities suboptimal. In other words, the demand distribution and the location of facilities for online retailers interact more closely with each other, making the distribution network optimization more iterative and challenging.

A number of recent research have studied customer demand prediction and distribution network optimization for online retailers separately. Most studies focus on improving online customer satisfaction and are based on the correctness and attractiveness of online product descriptions, online retailer reputation, customer service quality, and postsales customer evaluation (Bucko et al. 2018). However, the effect of the logistic services, affected by shipping time, delivery time, and logistics reliability, has not been paid special attention to. A two-level supplier-distribution network was recently designed to provide a faster delivery service with a lower logistics cost and higher warehouse turnover (Chen et al. 2017). In addition, the multilevel distribution network for the product replenishment and transshipment for physical retailers has been studied (Ortiz-Astorquiza et al. 2018). Nonetheless, for the real-world e-logistics, the impact of the distribution network on influencing its logistics service and the online demand distribution, which in turn affects the optimality of the resulting network, has not been investigated in detail.

To address the aforementioned challenges, in this paper, we propose an iterative optimization framework for a multilevel distribution network by leveraging online sales data, logistics data, and multilevel facility location data. Specifically, we first explore the sensitivity of retailer logistics service to the nationwide customer demand distribution by developing an artificial neural network (ANN) demand predictor. Then, we design a multilevel distribution network consisting of a set of primary facilities and secondary facilities. The primary facilities are interconnected with transshipment to ensure a balanced inventory distribution and fast replenishment for the secondary facilities. The transshipment between the primary facilities is necessary to respond to immediate needs from customers and hedge the uncertainty in demand prediction. The secondary facilities are designed for last-mile logistics, aiming to minimize the delivery distance to the regional customers. Then, a mixed integer linear programming (MILP) model is formulated to select the optimal primary and secondary facilities from a large set of facility candidates. We mention that the computational challenge of applying this MILP model stems from not only the interconnected structure of the network (see Section 3.1 for more details) but also the very large size of facility candidates. Specifically, unlike traditional retailers whose hub choices are typically restricted by their store locations and long-term strategies, the online retailers are more flexible and agile in adjusting their network structure and can rent a facility in any location with a reasonable fee and convenient shipping. Therefore, the facility candidates for online retailers have less restrictions and the pool size can be in thousands or larger. For example, our motivating case of a small online retailer had 1,367 facility candidates (see Section 3.1 for more details). Cainiao Network,[2] one of the largest e-logistics provider with a one-day delivery service, has more than 40,000 facility locations. In addition, considering the mutual interdependence between demand distribution and distribution network structure, we integrate the ANN predictor and our network optimization model to optimize the multilevel distribution network iteratively. Finally, we explore a set of real-world sales and logistics data from a leading Chinese online cosmetics retailer to validate the effectiveness and efficiency of our proposed methods.

The remainder of this paper is organized as follows. Section 2 summarizes the related literature. In Section 3, we formulate the problem of demand distribution prediction and multilevel distribution network optimization. Section 4 presents the demand prediction framework, and Section 5 provides the distribution network optimization model. Experimental setup and model performance are reported in Section 6. Finally, Section 7 concludes the paper and discusses the limitations. The research data are provided in the online supplement.

2. Related Work

In this section, we will review the related literature, specifically, in (1) demand forecasting for online products, (2) e-logistics for online retailers, and (3) facility location problems. By comparing with the existing literature, we will summarize our main contributions.

2.1. Online Sales Prediction

A group of researchers has worked on the online demand influential factors discovery. For example, Bucko et al. (2018) evaluated the importance of price, payment method, delivery time, product reviews, and product descriptions on online purchasing behavior. Panagiotelis et al. (2014) explored the nonlinear relationship between the online sales and online browsing-related factors, including website visiting duration and page reviews. Ferreira et al. (2015) implemented machine learning techniques to leverage the price of competing products for new products' demand forecasting in an online environment. Chong et al. (2017) proposed a neural network to examine the online reviews' interplay effects, online promotional strategies, and online sentiments on online sales demand. They further demonstrated that the neural network was efficient in integrating multiple factors to facilitate demand prediction in an online environment. Recently, the macroeconomic indicators were leveraged for online sales foresting (Zhang et al. 2020). However, most research mentioned above focuses on the short-term or long-term demand prediction for a single product. The distribution of online demand and the effects of logistics services have not been examined.

2.2. E-Logistics Service for Online Retailers

The logistics service satisfaction with fast delivery and high physical distribution quality has been discovered to be one of the key factors on customers' purchase satisfaction (Rao et al. 2011, Song et al. 2016). Fernie and Sparks (2018) summarized that the trade-off between the distribution cost and order fulfillment has been one of the most challenging objectives for online retailers. Speranza (2018) discussed that it was essential to consider the demand forecasting to improve logistics services, especially for the emerging e-commerce with variable demand over time. Lu et al. (2020) pointed out that many e-retailing companies are establishing their own distribution networks to improve customers' satisfaction. Recently, the emergence of multisource big data enables a new paradigm for enhancing logistics services (Wang et al. 2016). However, despite the promising prospect of applying the big data analysis to enhance the e-commerce logistics and boost the development of e-commerce (Yu et al. 2017), there are few quantitative analyses on the interactions between the logistics service quality and demand forecasting based on big data, which could be used to improve the accuracy of online demand forecasting.

2.3. Multilevel Facility Location Problem

The facility location has been playing a critical role in the strategic design of supply chain networks (Melo et al. 2009, Subramanian et al. 2014). A recent survey conducted in Ortiz-Astorquiza et al. (2018) discussed the multilevel facility location problem as a rapidly emerging research area, which is a result of the development of production-distribution systems and telecommunication network design. As low inventory costs and high order fulfillment rates are critical to e-logistics, most recent research addressed the importance of last-mile delivery of online orders (Hübner et al. 2016, Lim et al. 2018, Rohmer and Gendron 2020). Chen et al. (2017) designed a two-level supplier-distribution network for online retailers with the consideration of physical distribution of customer demand. Zetina et al. (2019b) proposed a multicommodity network design framework by explicitly considering demand elasticity with respect to routing cost. Ponce et al. (2020) considered a third-party e-logistics service provider to ensure its online sales distribution at various supply chain levels. To tackle the computational issue in multilevel facility location problems, a Lagrangian heuristic was developed featuring a mixed integer programming-based large neighborhood search (Gendron et al. 2016). Benders reformulations for uncapacitated network design problems were studied by enhancing standard decomposition algorithms with the use of variant optimality cuts and the execution of heuristic procedures (Contreras et al. 2011, Fischetti et al. 2017, Ortiz-Astorquiza et al. 2019, Zetina et al. 2019a). The clustering algorithms, which have been proven to scale very well with the size of instances (Xu and Tian 2015), have provided alternative heuristics for solving the optimization problems (Liu et al. 2016). We investigate the potential of solving facility location optimization problems using clustering-based heuristic solutions. Our work also focuses on the design of a multilevel distribution network structure for e-logistics; but, to the best of our knowledge, our focal issue—the interplay between distribution network design and the demand distribution—has not been considered in the literature.

2.4. Main Contributions

This paper contributes to the literature in the following aspects. We propose an iterative prediction-and-optimization strategy for the distribution network design, with the consideration of mutual interdependence between demand distribution and distribution network structures. Specifically, we develop an artificial neural network to predict customer demands, factoring in the logistic service quality and the purchasing power–related demographic statistics. The predictor is further implemented as an influential factor sensitivity analyzer and a demand distribution simulator, which are used in the logistics service sensitivity analysis and iterative distribution network optimization process, respectively.

Moreover, an MILP model is formulated to optimize facility locations in a two-level distribution network with minimum transportation, facility setup, and package processing costs. Although the MILP instances with small to medium sizes can be solved using a commercial MILP solver, large-size instances cannot be solved effectively within a reasonable amount of time for business decision makers. To this end, we develop an efficient two-stage heuristic for solving the facility location problem. The heuristic includes an agglomerative hierarchical clustering algorithm for optimizing the secondary facility locations and an expectation and maximization (E&M) algorithm for optimizing the primary facility locations. Motivated by a real-life distribution network design case study, we conduct numerical studies using real-world data provided by a leading online retailer for cosmetic products on *Taobao*, the largest e-commerce platform in China. Experiments show that the optimization of the distribution network can boost demand, which is further considered for network redesign to reach the optimality of facility locations. Further, the optimal number of primary and secondary facilities and the corresponding distribution network structure are selected using the iterative prediction-and-optimization strategy. This optimization strategy provides an alternative data-driven iterative approach to the traditional "one-off" optimization for facility location problems.

3. Problem Statement

In this section, we first define some preliminaries to be used throughout the rest of the paper and then introduce the problem of multilevel distribution network design considering its influence on demand.

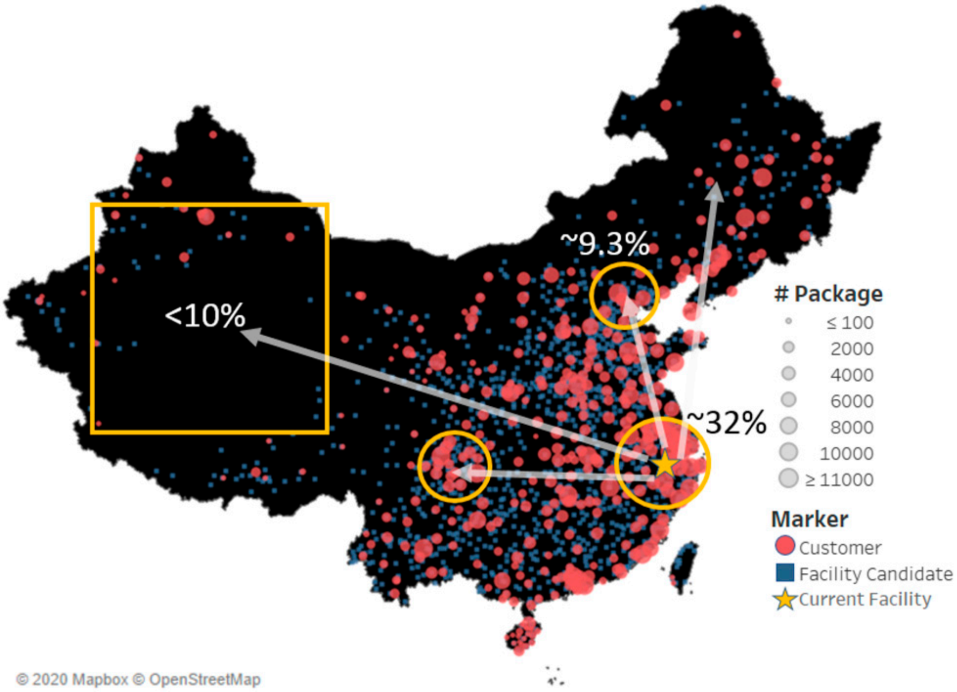
3.1. Problem Description

This research is motivated by a real-life distribution network redesign case from a leading online cosmetic product retailer in China, Xiaoye.[3] To illustrate the scale of this motivating example, Figure 1 presents the current demand distribution of *Xiaoye*, along with its existing centralized facility network. In Figure 1, each red dot represents a city, with its size representing the annual number of packages delivered. The nationwide sales distributions are aggregated into 371 major cities. It is seen that the western China area, with a longer shipping distance, contributed less than 10% of the company's total sales volume. On the other hand, the eastern mega city cluster around the Yangtze River Delta area, with next-day delivery service available, contributed more than 32% of its total sales volume. The current centralized facility network is not efficient, and the demand is low in the areas with long delivery distance and less competitive logistics services. The company's request was to rebuild the centralized facility network into an interconnected multilevel network (to be defined next), so as to enhance the logistics competitiveness and to increase the overall demand. Specifically, a set of 1,367 facility candidates nationwide (represented by the blue squares in Figure 1) are available to be set up as primary or secondary facilities. Note that there is only one supplier considered in this particular case, but the model and algorithms developed in this paper are generalized to support multiple suppliers.

Definition 1 (Multilevel Facility Network). A multilevel facility network consists of two layers: a primary-facility layer and a secondary-facility layer. The primary facilities collect products directly from suppliers whose locations are given. Transshipment between primary facilities is allowed to ensure a balanced inventory distribution and to provide fast replenishment for secondary facilities. Each secondary facility is designed to be located near a group of customers for last-mile delivery, which is key to the one-day or two-day delivery service.

Q: 22

Figure 1. (Color online) Customer Demand Distribution (Dots) and Candidate Facility Locations (Squares)

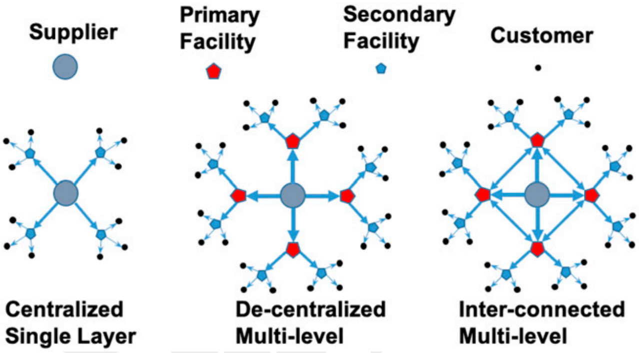


Different from the traditional centralized or decentralized network, which is built for traditional retailers or logistics service providers (Duan and Liao 2013), the multilevel facility network for online retailers requires a nonnegligible amount of transshipment between primary facilities to ensure a balanced inventory distribution and to meet the urgent needs of the secondary facilities (see Figure 2), as requests sent to suppliers are responded to with a longer lead time. We should mention that the transshipment is an operational-level decision used to compensate insufficient inventory at secondary facilities, which is not the same level of decision making as the distribution network design. In this paper, we approximate the incoming transshipment amount to each primary facility as a proportion to the total shipment amount from this primary facility to all the secondary facilities that it serves. It can be considered as a way to deal with the uncertainty in demand at the customer level. The more demand a primary facility serves, the more inventory shortage it may observe and the more incoming transshipment required to meet the urgent demand.

Next, we introduce three different units of logistics cost (per package per mile): C^S for supplier-primary facility transportation, C^P for shipping from primary facilities, and C^K for customer delivery. Usually $C^K > C^P > C^S$, because the transportation cost is discounted for large-volume shipping, whereas the express service for customer door delivery is the most expensive. For example, according to the charge rate of the express and heavy freight service of the leading express companies in China,[4] $C^K = \$12$ per 1,000 packages per mile, $C^P = 0.2C^K$, and $C^S = 0.05C^K$.

In addition to the transportation costs, the multilevel distribution network also incurs fixed setup costs and variable package processing costs (Boysen et al. 2019). Facility rental and package processing costs may vary among different retailers for different types of products. Because this paper considers a single-commodity problem, the fixed rental cost and unit package processing cost only vary per facility type (because of the difference in volume). Particularly, for the online cosmetic product retailer considered in the motivating example above, millions of packages are delivered every year. According to the standard warehousing and packaging service and individualized distribution processing service provided by SF

Figure 2. (Color online) A Sketch of Different Distribution Networks



Express, a primary facility will incur a rental cost of $250,000 per year and a $35 packaging fee per 1,000 packages and a secondary facility will incur a rental cost of $25,000 per year and a $140 package processing fee per 1,000 packages. The full set of facility candidate locations and demand distribution are provided in the online supplement.

In order to evaluate customers' satisfaction, *Taobao* introduced three major postservice evaluation scores: the logistics service quality score, the product description score, and the postpurchase service quality score. In this paper, we focus on the logistics service quality score and its impact on customer demand. Specifically, we extract and define three logistics service-related factors that could contribute to the online customer demand and provide a formal definition as follows.

Definition 2 (Logistics Service Quality).

The logistics service quality is affected by three factors: (1) *shipping time*: the average duration between the time an order is placed and the time the product is shipped; (2) *delivery time*: the average duration between the time a product is shipped and delivered; and (3) *damaged product ratio*: the proportion of damaged products due to improper shipping among all shipped products.

3.2. Iterative Prediction and Optimization

With Definitions 1 and 2, we subsequently define the multilevel distribution network optimization problem with the consideration of mutual interdependence between customer demand and facility network structure. To this end, we define two technical components: (1) demand distribution prediction and (2) facility location optimization. Q:11

- *Demand distribution prediction*: Given a set of customers $\mathcal{K}$ with their mailing addresses, historical demand, logistics service quality–related factors (see Definition 2), and the purchasing power–related demographic statistics at the city level, we develop an ANN predictor to predict the annual customer demand (number of packages delivered to the customers) in different cities. Note that the demand distribution needs to be re-estimated if the shipment and delivery time are changed because of the setup of a multilevel distribution network.
- *Facility location optimization:* Given a set of suppliers ($\mathcal{S}$), customer demand distribution ($\{d_i | i \in \mathcal{K}\}$), and candidates for primary and secondary facilities ($\mathcal{V}^P$ and $\mathcal{V}^Q$, respectively), the facility location optimization problem is formulated to select a set of primary facilities and secondary facilities and package transition paths from suppliers to customers that minimize the total logistics cost in the multilevel facility network defined in Definition 1. We further develop an efficient two-stage heuristic to solve the large-scale instances of this problem, where an agglomerative hierarchical clustering algorithm (Algorithm 1) determines the secondary facility locations and an E&M algorithm (Algorithm 2) determines the primary facility locations.

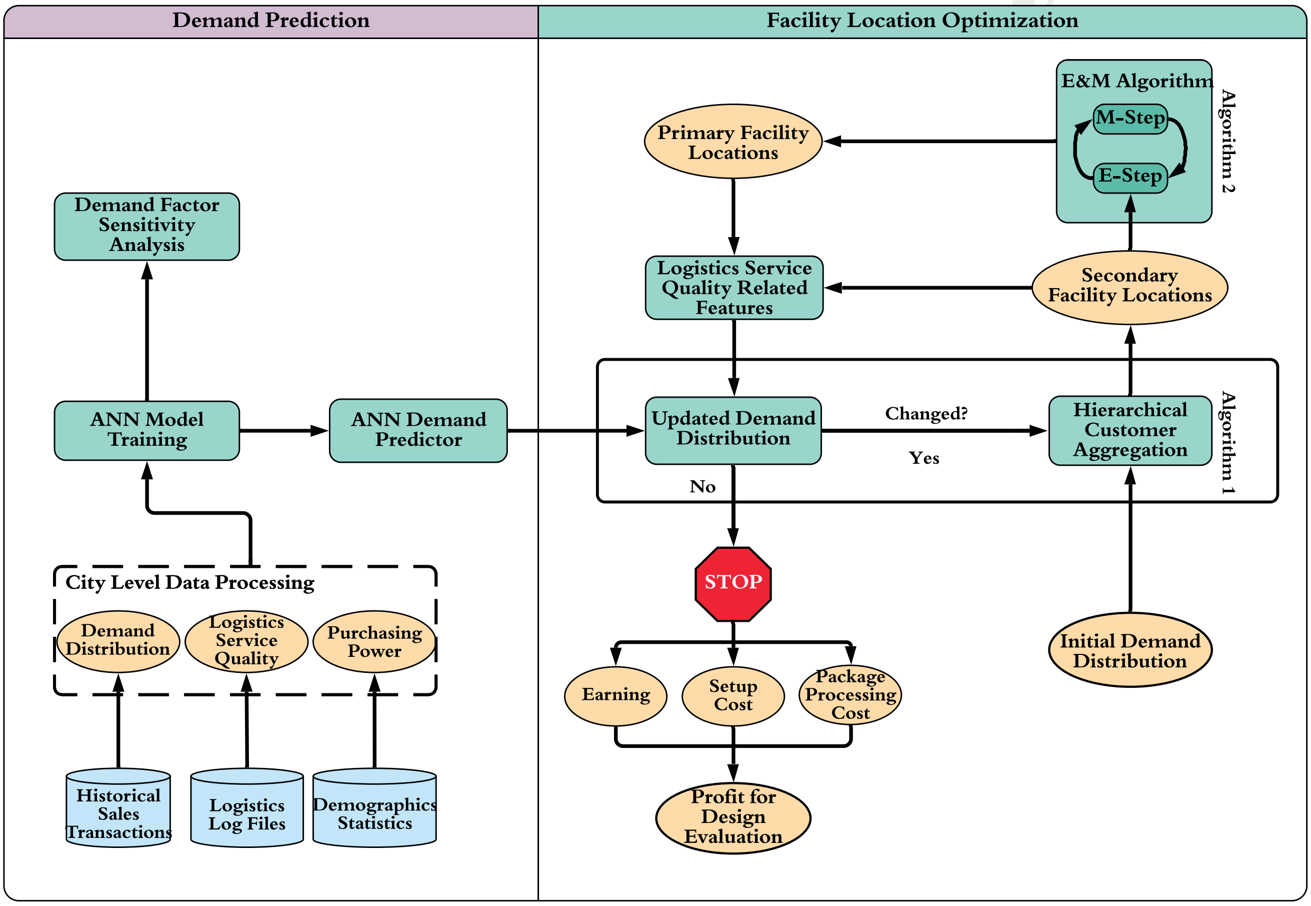
A notable contribution of this paper is the consideration of the mutual interdependence between demand distribution and the distribution network structure. To this end, we propose an iterative process to combine the ANN predictor and the facility location optimizer, until the predicted demand and the optimized facility locations converge. The design workflow is presented in Figure 3. Specifically, we first leverage the historical sales transaction data, logistics log files, and demographics statistics to train the ANN model (see Section 4) for demand prediction. Next, given the current distribution network, the ANN model predicts the demand, which serves as an input to the distribution network optimizer. The two-stage facility location optimization heuristic (see Section 5.2) is run to optimize the distribution network efficiently. Then, based on the optimized distribution network, the delivery time and the logistics service quality are recomputed, which are fed into the ANN model to repredict the demand distribution. After that, a new iteration of facility location optimization will be conducted. Such an iterative predict-and-optimize process continues until the demand distribution and the facility network no longer change between successive iterations.

We further mention that the retailer typically predetermines some parameters for the distribution network, such as the numbers of primary and secondary facilities to be set up. Because the two-stage heuristic can optimize the network design with a relatively short amount of time, it enables the retailer to also experiment with different setup parameters by comparing the total profit obtained. We will demonstrate one such scenario in our numerical experiments in Section 6.

4. Demand Distribution Prediction

Recalling that the customers at different locations have identical product descriptions, product evaluation, and convenience level when shopping online, we focus on the effects of demographic factors and logistics service quality on the annual customer demand prediction. Specifically, for each customer destination, we first extract the population, average wage, employment rate at the city level, and calculate the average shipping time, delivery time, damaged product ratio as the online consumption demand features. Then, we construct an artificial neural network (Hassoun et al. 1995), a generalized nonlinear prediction model inspired by the study of human brain recognition system, to leverage the features from different domains. Because of its superiority in learning and modeling

Figure 3. (Color online) Workflow of Iterative Distribution Network Design


Demand Prediction
Facility Location Optimization
Demand Factor Sensitivity Analysis
ANN Model Training
ANN Demand Predictor
City Level Data Processing
Demand Distribution
Logistics Service Quality
Purchasing Power
Historical Sales Transactions
Logistics Log Files
Demographics Statistics
E&M Algorithm
M-Step
E-Step
Algorithm 2
Primary Facility Locations
Logistics Service Quality Related Features
Secondary Facility Locations
Updated Demand Distribution
Changed?
Yes
No
Hierarchical Customer Aggregation
Algorithm 1
STOP
Earning
Setup Cost
Package Processing Cost
Profit for Design Evaluation
Initial Demand Distribution

nonlinear and complex relationships, ANN has been deployed in various disciplines (Abiodun et al. 2018). In our study, ANN can provide a high accuracy on demand prediction as well as a sensitivity analysis, which can help discover the nonlinear relationship between logistics service and demand. More details on prediction accuracy will be discussed in our numerical analysis in Section 6.

Figure 4 represents the architecture of our proposed ANN, including n feature inputs, $M-1$ hidden layers, S_k nodes in the kth hidden layer, and one layer to output the predicted demand. We summarize the details of the ANN model as follows:

4.1. Input Layer

The input layer incorporates the following data:

- Logistics service quality–related features: shipping time, delivery time, and damaged product ratio for each city given the current distribution network
- Purchasing power–related demographic factors: population, average wage, and employment rate for each city

The feature vector is normalized in order to prevent the simulated neurons from being driven too far into saturation.

4.2. Hidden Layer

The input of unit s in hidden layer k is the linear combination of the outputs α^{k-1} of units in layer $k-1$, as shown in Equation (1):

$$\beta^k(s) = \sum_{l=1}^{S_{k-1}} \omega_{ls}^k \alpha^{k-1}(l) + b_s^k, \quad \forall 1 \le s \le S_k, 1 \le k \le M-1, \tag{1}$$

where $\boldsymbol{\alpha}^0$ is the model inputs, ω_{ls}^k is the weight from unit l of layer $k-1$ to unit s of layer k, and b_s^k is the bias of unit s in layer k.

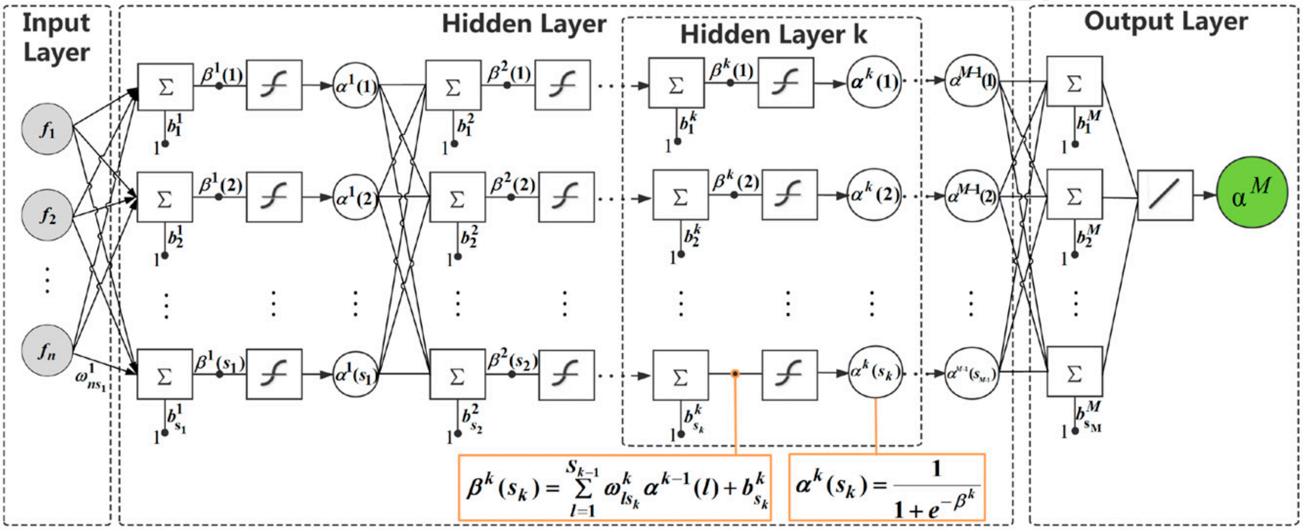
A sigmoid activation function is used to map the input of a neuron to its output. This function is especially advantageous to minimize the computation capacity for training, which is widely used in neural networks (Karlik and Olgac 2011).

$$\alpha^k(s) = \frac{1}{1+e^{-\beta^k}}, \quad \forall 1 \le s \le S_k, 1 \le k \le M-1 \tag{2}$$

4.3. Output Layer

The output layer is a linear layer for the regression problem (Goodfellow et al. 2016, p. 178). Note that the final output at M-th layer, α^M, is the predicted annual demand at the city level. Q: 12

Figure 4. (Color online) Architecture of Artificial Neural Network



4.4. Training Algorithm

The ANN training process aims to build the complex nonlinear relationships between the features and the demand by minimizing the mean squared prediction error. We implement the Levenberg-Marquardt algorithm (Ranganathan 2004), which has been proven to be one of the fastest training algorithms for ANN with a sum of the squared errors objective (Yu and Wilamowski 2011, Mukherjee and Routroy 2012). Moreover, a validation set is used to avoid overfitting during the training process. The model training, validation, and implementation were conducted using Tensorflow 2.0 under the open source Apache 2.0 licensed by Google (Abadi et al. 2016).

In summary, the ANN predictor builds a nonlinear relationship between the inputs and the output, in the form of $\alpha^M = ANN(f_1, f_2, \ldots, f_n)$, where α^M is the demand to be predicted and $f_1, f_2, \ldots, f_n$ are the logistics service quality–related features and purchasing power–related demographic factors at the city level. Each training sample contains the known historical demand at a single city as well as those inputs for the same city, where the logistics service quality–related features depend on the city location and the current distribution network.

For the data that we used for model training and validation, the ANN with two hidden layers ($M = 3$), each of which has 16 neurons ($S_k = 16$ for $1 \le k \le 2$), achieved the highest prediction accuracy and was selected for implementation.

5. Multilevel Facility Location Optimization

In this section, we first formulate a mixed integer linear programming model for the multilevel facility location optimization problem and then propose a two-stage hierarchical optimization algorithm for solving the problem.

5.1. MILP Model

The objective of the facility location optimization problem is to minimize the total logistics cost, including the shipping cost, facility setup cost, and package processing cost. The sets, parameters, and decision variables are defined in Table 1.

Accordingly, we formulate the MILP model as follows:

$$\begin{aligned}\min \ & \sum_{s\in\mathcal{S}}\sum_{i\in\mathcal{V}^P} C^S\delta_{si}x_{si} + \sum_{i\in\mathcal{V}^P}\sum_{j\in\mathcal{V}^Q} C^P\delta_{ij}y_{ij} + \sum_{i\in\mathcal{V}^P}\sum_{i'\in\mathcal{V}^P\setminus\{i\}} C^P\delta_{ii'}\tau_{ii'} \\ & + \sum_{j\in\mathcal{V}^Q}\sum_{k\in\mathcal{K}} C^K\delta_{jk}z_{jk} + \sum_{i\in\mathcal{V}^P}\eta^P_i g_i + \sum_{j\in\mathcal{V}^Q}\eta^Q_j g_j \\ & + \sum_{i\in\mathcal{V}^P}\gamma^P_i\left(\sum_{j\in\mathcal{V}^Q} y_{ij} + \sum_{i'\in\mathcal{V}^P\setminus\{i\}}\tau_{ii'}\right) + \sum_{j\in\mathcal{V}^Q}\gamma^Q_j\sum_{k\in\mathcal{K}} z_{jk}, \end{aligned} \tag{3}$$

$$\text{s.t.} \ \sum_{i\in\mathcal{V}^P} \mathrm{g_i} = \mathrm{r^P}, \tag{4}$$

$$\sum_{j\in\mathcal{V}^Q} g_j = r^Q, \tag{5}$$

$$\sum_{s\in\mathcal{S}} x_{si} \le \sum_{k\in\mathcal{K}} d_k g_i, \quad \forall\, i\in\mathcal{V}^P, \tag{6}$$

$$\sum_{j\in\mathcal{V}^Q} y_{ij} \le \sum_{k\in\mathcal{K}} d_k g_i, \quad \forall\, i\in\mathcal{V}^P, \tag{7}$$

$$\sum_{i\in\mathcal{V}^P} y_{ij} \le \sum_{k\in\mathcal{K}} d_k g_j, \quad \forall\, j\in\mathcal{V}^Q, \tag{8}$$

$$\sum_{k\in\mathcal{K}} z_{jk} \le \sum_{k\in\mathcal{K}} d_k g_j, \quad \forall\, j\in\mathcal{V}^Q, \tag{9}$$

$$\sum_{i'\in\mathcal{V}^P\setminus\{i\}} \tau_{ii'} \le \sum_{k\in\mathcal{K}} d_k g_i, \quad \forall\, i\in\mathcal{V}^P, \tag{10}$$

$$\sum_{i\in\mathcal{V}^P\setminus\{i'\}} \tau_{ii'} \le \sum_{k\in\mathcal{K}} d_k g_{i'}, \quad \forall\, i'\in\mathcal{V}^P, \tag{11}$$

$$\sum_{s\in\mathcal{S}} x_{si} + \sum_{i'\in\mathcal{V}^P\setminus\{i\}} \tau_{i'i} = \sum_{j\in\mathcal{V}^Q} y_{ij} + \sum_{i'\in\mathcal{V}^P\setminus\{i\}} \tau_{ii'} \quad \forall\, i\in\mathcal{V}^P, \tag{12}$$

$$\sum_{i\in\mathcal{V}^P} y_{ij} = \sum_{k\in\mathcal{K}} z_{jk}, \quad \forall\, j\in\mathcal{V}^Q, \tag{13}$$

Q: 23

Table 1. Table of Notations

	Sets	
$\mathcal{V}^P$		Set of primary facility candidates
$\mathcal{V}^Q$		Set of secondary facility candidates
$\mathcal{K}$		Set of customers
$\mathcal{S}$		Set of suppliers
$\mathcal{E}$		Set of edges connecting two nodes in the facility network
	Parameters	
d_k	$k \in \mathcal{K}$	Demand of the kth customer
δ_{ij}	$(i,j) \in \mathcal{E}$	Traveling distance between node i and node j in the facility network
η_i^P, η_j^Q	$i \in \mathcal{V}^P, j \in \mathcal{V}^Q$	Fixed annual rental cost for primary facility i and secondary facility j, respectively
γ_i^P, γ_j^Q	$i \in \mathcal{V}^P, j \in \mathcal{V}^Q$	Unit package processing cost at primary facility i and secondary facility j, respectively
ℓ		Transshipment amount in proportion of total downstream package number
r^P		Number of primary facilities to setup
r^Q		Number of secondary facilities to setup
C^S		Unit transportation cost (per package per mile) initiated from a supplier
C^P		Unit transportation cost (per package per mile) initiated from a primary facility
C^K		Unit delivery cost (per package per mile) to a customer
	Variables	
$x_{si} \in \mathbb{R}_{\geq 0}$	$s \in \mathcal{S}, i \in \mathcal{V}^P$	Number of packages transited from supplier s to primary facility i
$y_{ij} \in \mathbb{R}_{\geq 0}$	$i \in \mathcal{V}^P, j \in \mathcal{V}^Q$	Number of packages transited from primary facility i to secondary facility j
$z_{jk} \in \mathbb{R}_{\geq 0}$	$j \in \mathcal{V}^Q, k \in \mathcal{K}$	Number of packages delivered from secondary facility j to customer k
$\tau_{ii'} \in \mathbb{R}_{\geq 0}$	$i \in \mathcal{V}^P,\ \forall\ i' \in \mathcal{V}^P \backslash \{i\}$	Number of packages transshipped from primary facility i to i'
$g_i \in \{0,1\}$	$i \in \mathcal{V}^P \cup \mathcal{V}^Q$	Binary variable g_i equals 1 if facility i is selected and 0 otherwise

$$\sum_{i' \in \mathcal{V}^P \backslash \{i\}} \tau_{i'i} = \ell \sum_{j \in \mathcal{V}^Q} y_{ij}, \quad \forall\, i \in \mathcal{V}^P, \tag{14}$$

$$\sum_{j \in \mathcal{V}^Q} z_{jk} = d_k, \quad k \in \mathcal{K}. \tag{15}$$

The objective (3) is to minimize the total cost of the multilevel facility network, including the supplier to primary facility transportation cost, primary to secondary facility transportation cost, primary facility transshipment cost, facility setup cost, and package processing cost. Constraints (4) and (5) define the total number of facilities to be selected. Constraints (6)–(11) are the logic constraints indicating that if a facility is not selected, no packages will be transited through it. Constraints (12) and (13) are the package flow conservation constraints. Constraint (14) specifies that a proportion of the total downstream demand for a primary facility comes from the transshipment from other primary facilities because of the urgency of the requests (see Definition 1 and the paragraph that follows for the detailed explanation of transshipment). Q: 13 Constraint (15) indicates that the customer demand is strictly satisfied. Note that the facility location optimization problem is formulated with multiple suppliers. In practice, for online retailers with relatively small annual demand, it is common to have one single supplier for volume discount, which is the case for the real-world case study tested in Section 6.

The small- and medium-size instances of the above MILP model can be solved directly using commercial MILP solvers. In our implementation, we use the Gurobi MILP solver that is one of the industry standards in terms of computational speed and solution quality. However, as mentioned in Section 1, the size of the model can become very large because of the flexibility of online retailers in choosing facility candidates, rendering the model intractable using Gurobi for large-scale instances (e.g., it could take more than one week to solve a sample case with $|\mathcal{V}^P| = 100$ and $|\mathcal{V}^Q| = 1{,}000$; see Section 6 for detailed examples). In addition, the mutual interdependence between the predicted demand and the distribution network requires the facility location optimization problem to be solved iteratively, calling for a much more efficient solution method. We also implemented Lagrangian relaxation (Fisher 2004, Gendron et al. 2016) and Benders decomposition (Contreras et al. 2011), two classic mathematical programming techniques for facility location problems. Both methods were implemented in their generic forms without problem-specific heuristics and had extremely slow convergence and worse performance compared with Gurobi. Therefore, by observing the properties of the practical problem under study, we next develop a two-stage heuristic to solve the large-scale instances of the problem and compare its performance to the Gurobi MILP solver as the benchmark.

5.2. Two-stage Facility Location Optimization Heuristic

In this section, we develop a heuristic to improve the computational efficiency in solving the facility location problem under study. The heuristic has two stages: (1) an agglomerative hierarchical clustering algorithm for optimizing the secondary facility locations and (2) a distance-weighted expectation and maximization algorithm for optimizing the primary facility locations.

The heuristic scales very well with the size of the problem because (1) the clustering-based algorithm in the first stage has been widely used in data mining problems to group instances with similar patterns and can handle large sets of data efficiently and (2) the E&M algorithm in the second stage operates in a continuous coordinate space to speed up the search for the optimal primary facility locations. We next introduce the two algorithms in sequence to solve the distribution network redesign problem described in Section 3.1. Note that the real-life e-logistics design problem studied in this paper has identical facility setup costs and unit packaging processing costs, so the objective function (3) can be simplified by adding up the last four terms to constants.

5.2.1. Optimizing Secondary Facility Locations. The secondary facilities are built for fast delivery services and thus should be selected near the center of a group of customers. The basic idea of the clustering-based heuristic is that customers with large demands and close geographical locations are more likely to be in the same group of customers served by the same secondary facility.

We first define a similarity measure to quantify how likely two customers belong to the same group served by the same secondary facility. Specifically, given two customers i and j with demands (d_i and d_j) and location coordinates ($\vec{w}_i$ and $\vec{w}_j$), the similarity measure is defined as follows:

$$S(i,j) = \frac{(d_i + d_j)}{2}\exp\left(-\frac{\|\vec{w}_i - \vec{w}_j\|^2}{\sigma^2}\right), \qquad (16)$$
$$\forall i \in \mathcal{K},\ \forall j \in \mathcal{K} \setminus \{i\},$$

where σ is the standard deviation of intercustomer distance within a state/province. For example, we set σ = 100 miles for the real-world case study in Section 6. From Equation (16), the similarity measure increases when the demands become larger and two locations become closer. The similarity decreases significantly as the distance between two customers becomes greater than σ, and thus these two customers are less likely to be served by the same secondary facility regardless of their demand volume.

Next, we describe the clustering-based heuristic in determining secondary facility locations in Algorithm 1. First, we calculate a similarity matrix S whose elements are computed by (16) (step 1–2 in Algorithm 1). Then, targeting on the fourth term in the objective (3) of minimizing the delivery cost from secondary facilities to customers, two customers with the highest similarity are identified and merged into one (steps 3–6), which replaces the two identified customers (steps 7–9). The aforementioned process will iterate until the number of selected secondary facilities reaches the specified number, that is, $|\ \mathcal{N}^Q\ | = r^Q$. The set of centers $\mathcal{W}^Q = \{\vec{w}_i | i \in \mathcal{N}^Q\}$ indicates the locations of selected secondary facilities, and the set of demand $\mathcal{D}^Q = \{d_i | i \in \mathcal{N}^Q\}$ indicates the total number of downstream packages sent from the secondary facilities. Finally, each coordinate in $\mathcal{W}^Q$ is mapped to its closest location in set $\mathcal{V}^Q$ for the final location choice.

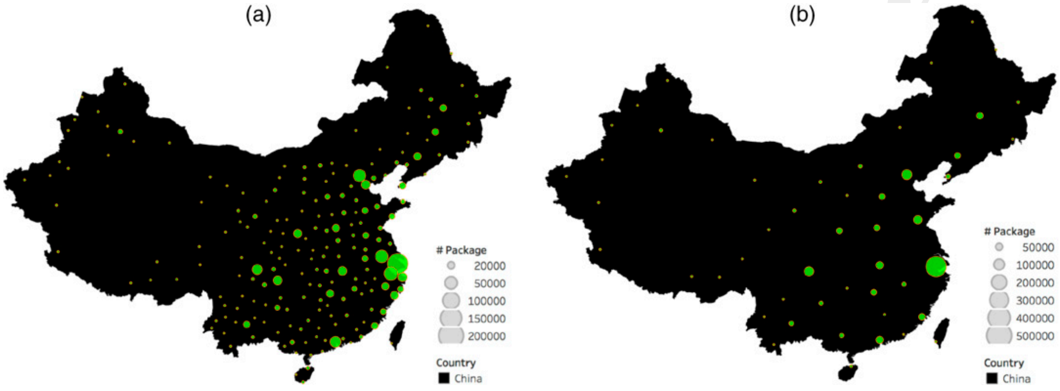
Algorithm 1 Agglomerative Hierarchical Clustering ($\mathcal{W}, \mathcal{D}, r^Q$)

```
Input:
$\mathcal{W} = \{\vec{w}_k | k \in \mathcal{K}\}$: a set of customer coordinates;
$\mathcal{D} = \{d_k | k \in \mathcal{K}\}$: a set of customer demands;
$r^Q$ $(\leq |\ \mathcal{K}\ |)$: target number of secondary facilities
Output:
$\hat{\mathcal{V}}^Q \subseteq \mathcal{V}^Q$: a set of selected secondary facilities where $|\hat{\mathcal{V}}^Q| = r^Q$
Initialization: $\mathcal{N}^Q = \mathcal{K}$, $\mathcal{W}^Q = \mathcal{W}$, $\mathcal{D}^Q = \mathcal{D}$, $\hat{\mathcal{V}}^Q = \emptyset$
While $|\ \mathcal{N}^Q\ | > r^Q$:
1:  Calculate (update) the similarity matrix $S$ using Equation (16):
2:    $S_{ij} = S(i,j),\ \forall\ i \in \mathcal{N}^Q, j \in \mathcal{N}^Q \backslash \{i\}$
3:  Identify and merge two nodes with the maximum similarity:
4:    Find $(i^*, j^*) = \arg\max_{(i,j)}\{S_{ij} : i \in \mathcal{N}^Q, j \in \mathcal{N}^Q, i < j\}$
5:    $\vec{w}_{new} = \frac{d_{i^*}\vec{w}_{i^*} + d_{j^*}\vec{w}_{j^*}}{d_{i^*} + d_{j^*}}$
6:    $d_{new} = d_{i^*} + d_{j^*}$
7:  Replace the two nodes with the merged node:
8:    $\mathcal{N}^Q \leftarrow \mathcal{N}^Q \backslash \{i^*, j^*\}, \mathcal{W}^Q \leftarrow \mathcal{W}^Q \backslash \{\vec{w}_{i^*}, \vec{w}_{j^*}\}, \mathcal{D}^Q \leftarrow \mathcal{D}^Q \backslash \{d_{i^*}, d_{j^*}\}$
9:    $\mathcal{W}^Q \leftarrow \mathcal{W}^Q \cup \{\vec{w}_{new}\}$, $\mathcal{D}^Q \leftarrow \mathcal{D}^Q \cup \{d_{new}\}$
Mapping: For each coordinate $\vec{w} \in \mathcal{W}^Q$, find its closet location $v \in \mathcal{V}^Q \backslash \hat{\mathcal{V}}^Q$ and set $\hat{\mathcal{V}}^Q \leftarrow \hat{\mathcal{V}}^Q \cup \{v\}$ .
```

To illustrate the output of Algorithm 1, Figure 5 presents the optimal secondary facility locations using the demand data shown in Figure 1. Specifically, we implemented Algorithm 1 for two cases with $r^Q = 200$ and $r^Q = 50$. It can be seen from Figure 5 that the agglomerative hierarchical clustering algorithm can aggregate the customers into a localized, small-ranged secondary facility network, where each selected secondary facility will be located in the center of a group of closely located customers with high demand. In general, a smaller secondary facility network results in a larger delivery service area, a higher number of delivery requests, and a longer delivery time.

5.2.2. Optimizing Primary Facility Locations. Given the secondary facility locations, we propose an expectation and maximization algorithm to optimize the primary facility locations. The E&M algorithm is an

Figure 5. (Color online) Distribution of Selected Secondary Facility Locations



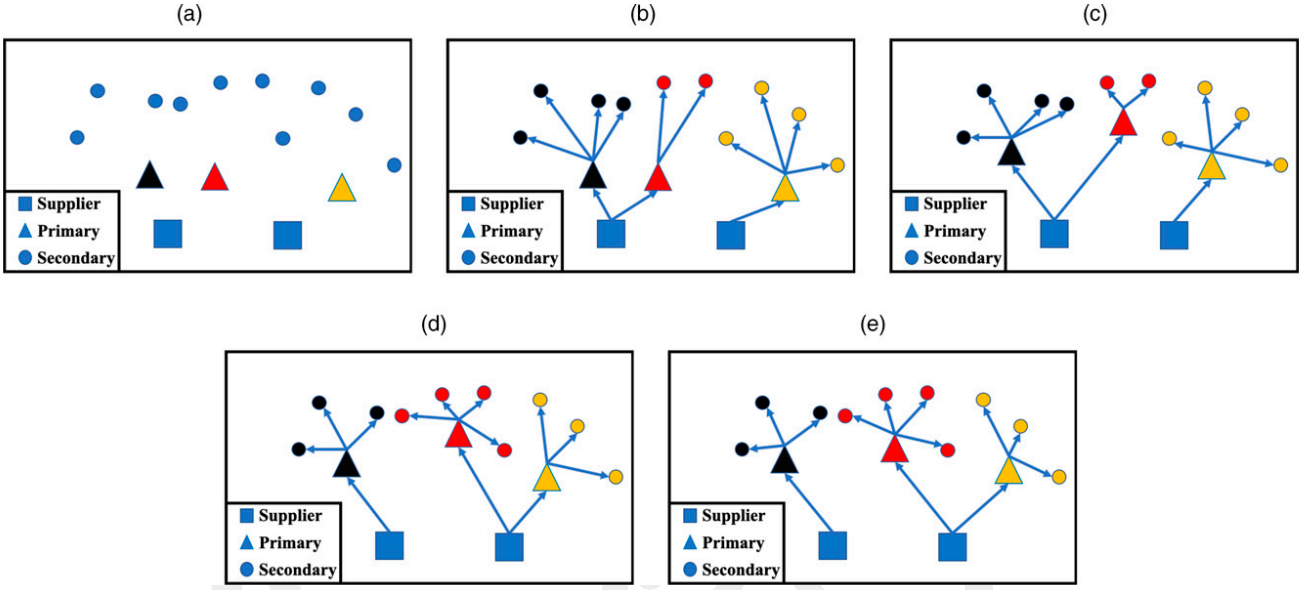
Note. (a) $r^Q = 200$; and (b) $r^Q = 50$.

iterative optimization method to estimate unknown parameters (Do and Batzoglou 2008). Because the proof of the E&M method's convergence has been established (Wu 1983), the algorithm and its variants have been widely deployed in machine learning (Zhang and Suganthan 2016), psychometrics (Bohlmeijer et al. 2011), medical image reconstruction (Masood et al. 2015), and structural engineering (Matarazzo and Pakzad 2016). However, the E&M algorithm has not been fully investigated for facility location problems. Here, we develop an E&M algorithm with closed-form expressions for its E-step and M-step, so as to efficiently optimize the primary facility locations and the resulting network flows.

In a nutshell, starting from an initial set of primary facility locations, the E-step searches for the optimal path from suppliers to the secondary facilities identified by Algorithm 1 and then the M-step reoptimizes the locations of the primary facilities on a continuous coordinate space such that the total cost related to the primary facilities in objective (3) is minimized. This process iterates until convergence where the locations of primary facilities do not change between successive iterations.

To illustrate, Figure 6 provides a simple example showing the iterative process of the E&M algorithm,

Figure 6. (Color online) Illustrative Example of the Iterative Process of the E&M Algorithm



Note. (a) Initialization, (b) E-step of iteration 1, (c) M-step of iteration 1, (d) E-step of iteration 2, and (e) M-step of iteration 2.

Q:14 where the goal is to decide where to locate the primary facilities and which supplier and secondary facilities are assigned to each primary facility. Recall that the locations of suppliers are fixed, and the locations of secondary facilities have been determined by Algorithm 1. In Figure 6(a), a set of initial primary facility locations are randomly selected from the set $\mathcal{V}^P$. We denote the set of selected primary facilities by $\mathcal{N}^P$, Q:15 where $|\mathcal{N}^P| = r^P$. Figure 6, (b) and (c) show the first iteration of the algorithm, where the E-step in Figure 6(b) decides the optimal paths from suppliers to secondary facilities that minimize the total transportation costs and the M-step in Figure 6(c) reoptimizes the locations of the primary facilities given the paths decided in the E-step. Then, this process is repeated in Figure 6, (d) and (e), where the E-step in Figure 6(d) reoptimizes the optimal paths given the primary facilities obtained in Figure 6(c) and the M-step in Figure 6(e) reoptimizes the locations of the primary facilities given the new paths. This process will repeat until the primary facilities and paths stabilize. We now proceed to explain in detail how the E-step and M-step work.

In the E-step, given the predetermined primary facilities $\mathcal{N}^P$, we search for the optimal paths from the suppliers to the secondary facilities, such that the sum of the first two terms in objective (3) is minimized. Because the set of secondary facilities $\mathcal{N}^Q$ and their downstream customers have been determined in Algorithm 1, the demand at each secondary facility is known. Therefore, given a secondary facility $j \in \mathcal{N}^Q$, minimizing the total transportation cost from suppliers to primary facilities and from primary facilities to j is essential to find the shortest path from a supplier to j going through one of the primary facilities. That is, for each $j \in \mathcal{N}^P$, we can construct a shortest path problem, where the set of nodes is $V_j = \mathcal{S} \cup \mathcal{N}^P \cup \{j\}$, the set of edges is $E_j = \mathcal{S} \times \mathcal{N}^P \cup \mathcal{N}^P \times \{j\}$, and the weight on each edge is $f_{si} = C^S \|\vec{w}_s - \vec{w}_i\|$ for $s \in \mathcal{S}, i \in \mathcal{N}^P$ and $f_{ij} = C^P \|\vec{w}_i - \vec{w}_j\|$ for $i \in \mathcal{N}^P$. Thus, the optimal path is decided by solving the following shortest path problem:

$$(s_j^* \in \mathcal{S}, i_j^* \in \mathcal{N}^P) \leftarrow SP(V_j, E_j, f), \quad \forall j \in \mathcal{N}^Q,$$

where s_j^* and i_j^* represent the optimal supplier and primary facility to serve the secondary facility j with the minimal transportation cost. Each shortest path problem $SP(V_j, E_j, f)$ can be solved independently and parallelly using an efficient Dijkstra's algorithm (Rardin and Rardin 1998, p. 440). Once the path is determined, we can subsequently calculate the flow on each edge, denoted by $\{\hat{x}_{si} | s \in \mathcal{S}, i \in \mathcal{N}^P\}$, $\{\hat{\tau}_{ii'} | i \in \mathcal{N}^P, i' \in \mathcal{N}^P \backslash \{i\}\}$, and $\{\hat{y}_{ij} | i \in \mathcal{N}^P, j \in \mathcal{N}^Q\}$. Note that the transshipment into primary facility i comes from its closely neighbor i', and $\hat{\tau}_{i'i} = \ell \sum_{j \in \mathcal{N}^Q} \hat{y}_{ij}$ as specified in Equation (14). Q:16

In the M-step, given the predetermined transportation paths and flow $\{\hat{x}_{si}\}$, $\{\hat{\tau}_{ii'}\}$, $\{\hat{y}_{ij}\}$, we reoptimize the primary facility coordinates $\{\vec{w}_i | i \in \mathcal{N}^P\}$, such that the sum of the first three terms in objective (3) is minimized. For a simpler derivation, we approximate the δ parameter between two nodes in objective (3) with the square of their Euclidean distance and rewrite the objective as follows:

$$\min_{\{\vec{w}_i | i \in \mathcal{N}^P\}} \mathcal{G} = \sum_{s \in \mathcal{S}} \sum_{i \in \mathcal{N}^P} C^S \hat{x}_{si} \|\vec{w}_i - \vec{w}_s\|^2 + \sum_{i \in \mathcal{N}^P} \sum_{i' \in \mathcal{N}^P \backslash \{i\}} C^P \hat{\tau}_{ii'} \|\vec{w}_i - \vec{w}_{i'}\|^2 + \sum_{i \in \mathcal{N}^P} \sum_{j \in \mathcal{N}^Q} C^P \hat{y}_{ij} \|\vec{w}_i - \vec{w}_j\|^2. \tag{17}$$

The following proposition provides a closed-form expression for calculating the optimal primary facility locations $\{\vec{w}_i | i \in \mathcal{N}^P\}$ that minimize function $\mathcal{G}$.

Proposition 1. *Given the flows in the network, $\{\hat{x}_{si} | s \in \mathcal{S}, i \in \mathcal{N}^P\}$, $\{\hat{\tau}_{ii'} | i \in \mathcal{N}^P, i' \in \mathcal{N}^P \backslash \{i\}\}$, and $\{\hat{y}_{ij} | i \in \mathcal{N}^P, j \in \mathcal{N}^Q\}$, the coordinates of primary facilities $\{\vec{w}_i = (\mu_i, \nu_i) | i \in \mathcal{N}^P\}$ that minimize the total transportation cost $\mathcal{G}$ in Equation (17) are given by*

$$\mu_i = \sum_{u \in \mathcal{N}^P} [\mathbf{T}^{-1}]_{iu} \mathbf{A}_u, \quad \nu_i = \sum_{u \in \mathcal{N}^P} [\mathbf{T}^{-1}]_{iu} \mathbf{B}_u, \quad \forall i \in \mathcal{N}^P, \tag{18}$$

*where **T** is an invertible $(r^P \times r^P)$-matrix with its element specified by*

$$\mathbf{T}_{iu} = \begin{cases} C^S \sum_{s \in \mathcal{S}} \hat{x}_{si} + C^P \sum_{i' \in \mathcal{N}^P \backslash \{i\}} (\hat{\tau}_{ii'} + \hat{\tau}_{i'i}) + C^P \sum_{j \in \mathcal{N}_Q} \hat{y}_{ij} & \text{if} u = i \\ -C^P (\hat{\tau}_{iu} + \hat{\tau}_{ui}) & \text{if} u \neq i \end{cases}, \forall i, u \in \mathcal{N}^P \tag{19}$$

*and **A** and **B** are $(r^P \times 1)$-vector with its element specified respectively by*

$$\mathbf{A}_i = C^S \sum_{s \in \mathcal{S}} \hat{x}_{si} \mu_s + C^P \sum_{j \in \mathcal{N}^Q} \hat{y}_{ij} \mu_j, \quad \forall i \in \mathcal{N}^P, \tag{20}$$

$$\mathbf{B}_i = C^S \sum_{s \in \mathcal{S}} \hat{x}_{si} \nu_s + C^P \sum_{j \in \mathcal{N}^Q} \hat{y}_{ij} \nu_j, \quad \forall i \in \mathcal{N}^P. \tag{21}$$

Proof of Proposition 1. We first consider the latitude μ_i of the coordinate $\vec{w}_i$ for all primary facility candidates $i \in \mathcal{N}^P$. Because the objective function $\mathcal{G}$ in (17) is a convex function of μ_i, the optimality of $\mathcal{G}$ is achieved when $\partial \mathcal{G} / \partial \mu_i = 0$ for $i \in \mathcal{N}^P$. Breaking down the expression of $\partial \mathcal{G} / \partial \mu_i$, we have a system of equations as $\mathbf{T} \boldsymbol{\mu} = \mathbf{A}$, where $\mathbf{T}$ and $\mathbf{A}$ are given by Equations (19) and (20), respectively, and $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_{r^P}]^\top$.

Further, the coefficient matrix $\mathbf{T}$ is a strictly diagonally dominant matrix, which satisfies $|\mathbf{T}_{ii}| > \sum_{u \in \mathcal{N}^P \backslash \{i\}} |\mathbf{T}_{iu}|$ for all $i \in \mathcal{N}^P$. From the Levy-Desplanques theorem, a strictly diagonally dominant matrix is nonsingular (Cheney and Kincaid 2009, p. 654). Therefore, $\mathbf{T}$ is invertible and the optimal solution $\boldsymbol{\mu} = \mathbf{T}^{-1}\mathbf{A}$ in Equation (18) exists, where $\mu_i = \sum_{u \in \mathcal{N}^P} [\mathbf{T}^{-1}]_{iu} \mathbf{A}_u$.

The second part of Equation (18) can be obtained following a similar proof by considering the longitude v_i of the coordinates $\vec{w}_i$ for all primary facility candidates $i \in \mathcal{N}^P$. □

The E-step and M-step repeat until the algorithm converges when the optimal primary facility locations and network flows become stable. Algorithm 2 describes the two-step iterative E&M algorithm by summarizing the above results. Note that each coordinate in $\mathcal{W}^P$ is mapped to its closest location in set $\mathcal{V}^P$ for the final location choice.

Algorithm 2 E&M Algorithm ($\mathcal{W}^S, \mathcal{W}^Q, r^P$)

```
Input:
  $\mathcal{W}^S = \{\vec{w}_s \mid s \in \mathcal{S}\}$: a set of supplier location coordinates;
  $\mathcal{W}^Q = \{\vec{w}_j \mid j \in \mathcal{N}^Q\}$: a set of selected secondary facility locations;
  $r^P$: target number of primary facilities
Output:
  1. $\hat{\mathcal{V}}^P \subseteq \mathcal{V}^P$: a set of selected primary facilities where $|\hat{\mathcal{V}}^P| = r^P$
  2. $\mathcal{F} = \{\hat{x}_{si} | s \in \mathcal{S}, i \in \mathcal{N}^P\} \cup \{\hat{\tau}_{ii'} | i \in \mathcal{N}^P, i' \in \mathcal{N}^P \backslash \{i\}\} \cup \{\hat{y}_{ij} | i \in \mathcal{N}^P, j \in \mathcal{N}^Q\}$: flows in the network
Initialization:
  Randomly select $r^P$ initial primary facilities $\{\vec{w}_i \mid i \in \mathcal{N}^P\}$ from the candidate set $\mathcal{V}^P$
Repeat Until $\mathcal{W}^P$ and $\mathcal{F}$ do not change:
  E-step: optimize paths and determine network flows:
  Solve the shortest path problem for each secondary facility $j \in \mathcal{N}^P$:
  $(s_j^* \in \mathcal{S}, i_j^* \in \mathcal{N}^P) \longleftarrow SP(V_j, E_j, f)$
  Based on the optimal paths, aggregate the network flows:
  $\mathcal{F} \longleftarrow \{\hat{x}_{si} | s \in \mathcal{S}, i \in \mathcal{N}^P\} \cup \{\hat{\tau}_{ii'} | i \in \mathcal{N}^P, i' \in \mathcal{N}^P \backslash \{i\}\} \cup \{\hat{y}_{ij} | i \in \mathcal{N}^P, j \in \mathcal{N}^Q\}$
  M-step: optimize primary facility locations
  Update primary facility coordinates $\mathcal{W}^P \longleftarrow \{\vec{w}_i \mid i \in \mathcal{N}^P\}$ using Proposition 1
Mapping: For each coordinate $\vec{w} \in \mathcal{W}^P$, find its closest location $v \in \mathcal{V}^P \backslash \hat{\mathcal{V}}^P$, and set $\hat{\mathcal{V}}^P \longleftarrow \hat{\mathcal{V}}^P \cup \{v\}$.
```

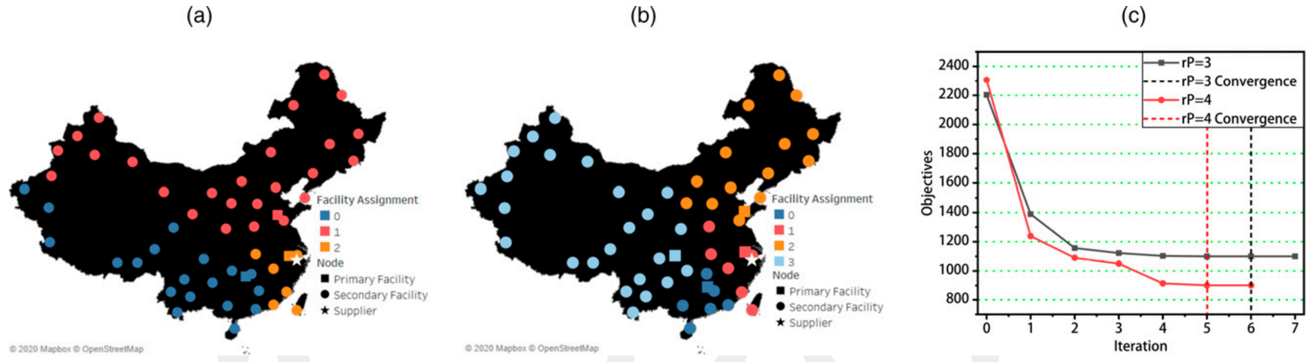
Figure 7 presents two examples of the resulting optimal primary facility locations (represented by squares) and the attendant secondary facilities that they serve (circles with the same color as the assigned primary facility). Note that the final locations of primary facilities are driven by not only the assigned secondary facility locations but also other primary facility locations and the supplier locations, which reflect the objective of minimizing transportation cost. Furthermore, Figure 7(c) plots the convergence progresses for the two examples, which stop in six and five iterations, respectively.

5.3. Iterative Multilevel Distribution Network Design

After we optimize the multilevel distribution network, the quality of logistics services, especially the delivery time, may change significantly compared with the network before optimization. As a result, the customer demand distribution will change and, therefore, the facility locations need to be reoptimized.

To consider the mutual interdependence between the demand distribution and distribution network structure, we combine the ANN predictor and the two-stage heuristic into an iterative distribution network design framework, as shown in Figure 3 in Section 3. Specifically, Algorithm 3 implements the workflow in Figure 3. We first leverage the historical sales transaction data, logistics log files, and demographics statistics to extract demand influential factors. These factors are used to train the ANN model in Section 4, which is later used for demand prediction Q:17

Figure 7. (Color online) Examples of Primary Facility Locations and Secondary Facility Assignments



Note. (a) r^P=3, r^Q=50; (b) r^P=4, r^Q=50; and (c) convergence progress.

once a new distribution network is selected. Then, the agglomerative hierarchical clustering algorithm is used to optimize the secondary facility locations, which are then used in the E&M algorithm to optimize the primary facility locations and the attendant network flows. With the optimized distribution network, the ANN model is used to repredict the demand distribution. If the demand distribution is different from the previous prediction, the distribution network is reoptimized using Algorithms 1 and 2; otherwise, the iterative process terminates and the network is used as the final output for implementation. We mention that, for small- and medium-size instances, steps 1–4 can be replaced by solving the MILP model (3)–(15) using Gurobi directly. Finally, real-life case studies (see Section 6.2) will show that the iterative network design process can substantially boost the annual demand and total profit of online retailers.

Algorithm 3 Iterative Prediction and Optimization $(\mathcal{W}^S, \mathcal{W}, \mathcal{D}, r^P, r^Q)$

```
Input:
  $\mathcal{W}^S = \{\vec{w}_s \mid s \in \mathcal{S}\}$: a set of supplier location coordinates;
  $\mathcal{W} = \{\vec{w}_i \mid i \in \mathcal{K}\}$: a set of customer coordinates;
  $\mathcal{D} = \{d_i \mid i \in \mathcal{K}\}$: a set of customer demand;
  $r^P$, $r^Q$: target numbers of selected primary and secondary facilities
Output:
  1. $\hat{\mathcal{V}}^P$: optimal primary facility locations;
  2. $\hat{\mathcal{V}}^Q$: optimal secondary facility locations;
  3. $\mathcal{F}$: optimal flows in the network
Repeat Until demand distribution $\mathcal{D}$ does not change:
  1: Optimize secondary facility locations:
  2:     $\hat{\mathcal{V}}^Q \leftarrow$ Agglomerative Hierarchical Clustering $(\mathcal{W}, \mathcal{D}, r^Q)$
  3: Optimize primary facility locations and network flows:
  4:     $(\hat{\mathcal{V}}^P, \mathcal{F}) \leftarrow$ E&MAlgorithm $(\mathcal{W}^S, \mathcal{W}^Q, r^P)$
  5: Predict customer demand:
  6:     $\mathcal{D} \leftarrow ANN(\hat{\mathcal{V}}^P, \hat{\mathcal{V}}^Q)$
```

6. Numerical Experiments

In this section, we demonstrate the efficiency and effectiveness of the proposed methods with extensive experiments using one-year real-world data from *Xiaoye*, a leading e-commerce company on *Taobao*. Summary statistics of the e-commerce company's logistics service data and purchasing power–related demographic data are presented in Table 2. More than one million customers from 371 big cities in China are included. All experiments were conducted on a server with a 2X 10-core Intel Xeon Gold 5215 Processor, 1 TB RAM, and 10X 2080Ti GPUs.

6.1. Numerical Results for Demand Prediction

6.1.1. Baselines and Metrics. To evaluate the demand prediction accuracy, we compare the ANN demand predictor with the following commonly used nonlinear predictors:

- *Random forest* (RF) (Liaw et al. 2002): Random forest is built upon an ensemble of decision trees and an equally weighted voting mechanism.
- *Gradient boosting trees regressor* (GBR) (Friedman 2001): Gradient boosting trees regressor is built in a stagewise fashion, which combines decision trees iteratively.
- *AdaBoost regressor* (AR) (Solomatine and Shrestha 2004): AdaBoost regressor uses sequential and weighted stumps to produce predictive outputs.
- *Decision tree regressor* (DTR) (Loh 2014): Decision tree regressor produces continuous predicted values by dividing the observations at thresholds where the sum of squared residuals can be minimized for each decision node.
- *K-nearest neighbors regressor* (KNR) (Hastie and Tibshirani 1996): K-nearest neighbors regressor is applied to estimate the target based on its similarity to the neighbors in the feature space.
- *Support vector regressor* (SVR) (Drucker et al. 1997): The model of support vector regressor family depends on a subset of the training data, which ignores any training data close to the model prediction. The SVR with radial basis function kernel is chosen as a baseline algorithm.

We adopt two widely used measures for performance comparison, namely, the error rate (*ER*) and root mean squared logarithmic error (*RMLSE*), which are formally defined as follows:

$$ER = \frac{\sum_{k \in \mathcal{K}_t} \mid \hat{d}_k - d_k \mid}{\sum_{k \in \mathcal{K}_t} d_k},$$

$$RMLSE = \sqrt{\frac{1}{\mid \mathcal{K}_t \mid} \sum_{k \in \mathcal{K}_t} (log(\hat{d}_k + 1) - log(d_k + 1))^2},$$

where d_k is the ground truth value for the demand of customer k, whereas $\hat{d}_k$ is the predicted counterpart. $\mathcal{K}_t (\subset \mathcal{K})$ represents the testing set. The ER metric provides an overall percentage error, which is a good evaluation metric when the demand varies significantly among different customers. The RMLSE provides a small misprediction penalty when there are customers with extremely large demand. The research data for the ANN model training and evaluation are provided in the online supplement.

Table 2. Summary Statistics of the Data Sets

Data source		Min	Max	Average
Logistics service	Shipping time	<1 hour	3 days	10.8 hours
	Delivery time	<1 hour	18 days	51 hours
	Damage ratio	0.13%	2.15%	0.33%
Purchasing power	Population	0.573 million	14.269 million	4.786 million
	Average wage	$437	$5,121	$3,514
	Employment rate	68.30%	98.52%	91.86%

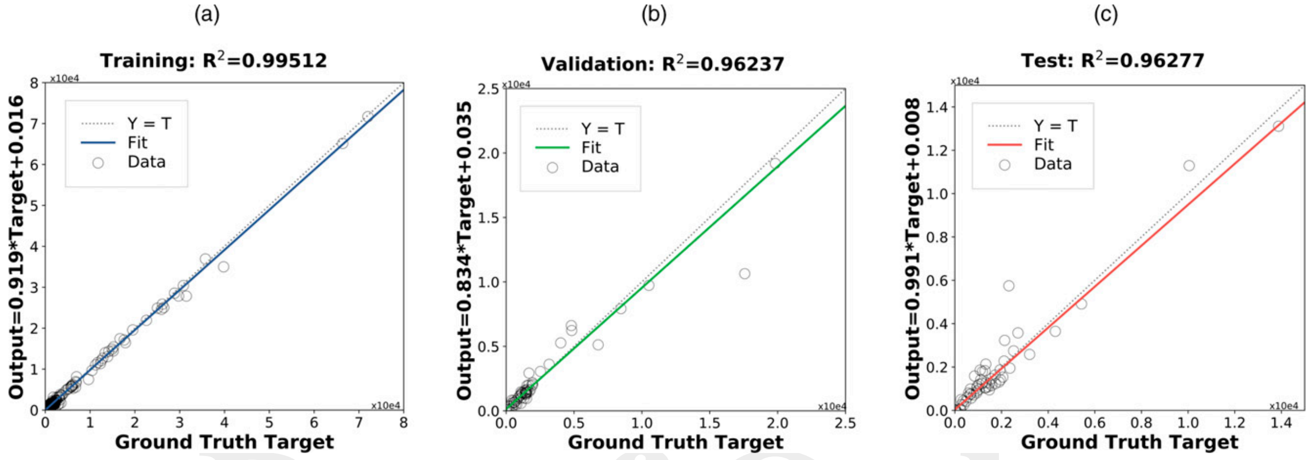
6.1.2. Performance Comparison. The original instance set with 371 cities is randomly divided into three subsets: a training set, a validation set, and a testing set. The training subset has 260 instances and is used to optimize the ANN parameters. The second subset has 55 instances that are used to monitor the validation error in each training epoch. If the validation error continues to grow for three consecutive epochs, the training process is stopped and the ANN reaching the minimum validation error is selected. The rest of the instances are used as the testing data.

Figure 8 presents the scatter plot of predicted demand against its ground truth value. Ideally, the scattered points should be close to the regression line of $y = x$. From Figure 8(a) and Figure 8(b), we can see that the ANN achieves a good training result and avoids overfitting. Moreover, the fitting result of the regression line in Figure 8(c) achieves an R^2 (R-squared, the coefficient of determination) of 0.96277 and a slope of 0.991, indicating that the predicted values are close to the ground truth values in the testing set.

Next, we assess the performance of the proposed ANN model by comparing with the baseline methods using the same training-testing set. Figure 9(a) summarizes the performance comparison in terms of ER and RMLSE metrics defined above. As seen, the proposed ANN model achieves the lowest ER of 22.24% and RMLSE of 0.7149, which outperforms the second best baseline random forest regressor with an ER of 26.27% and an RMLSE of 1.0670. The proposed ANN model outperforms the boosting and ensemble methods, such as RF, GBR, and AR, which shows its ability to implicitly discover complex nonlinear relationships between demand and the influential factors by detecting all possible interactions between predictor variables. The well-trained ANN model with a high prediction accuracy can be further implemented as an influential factor sensitivity analyzer and a demand distribution simulator, which are used in the following sensitivity analysis and iterative distribution network optimization process.

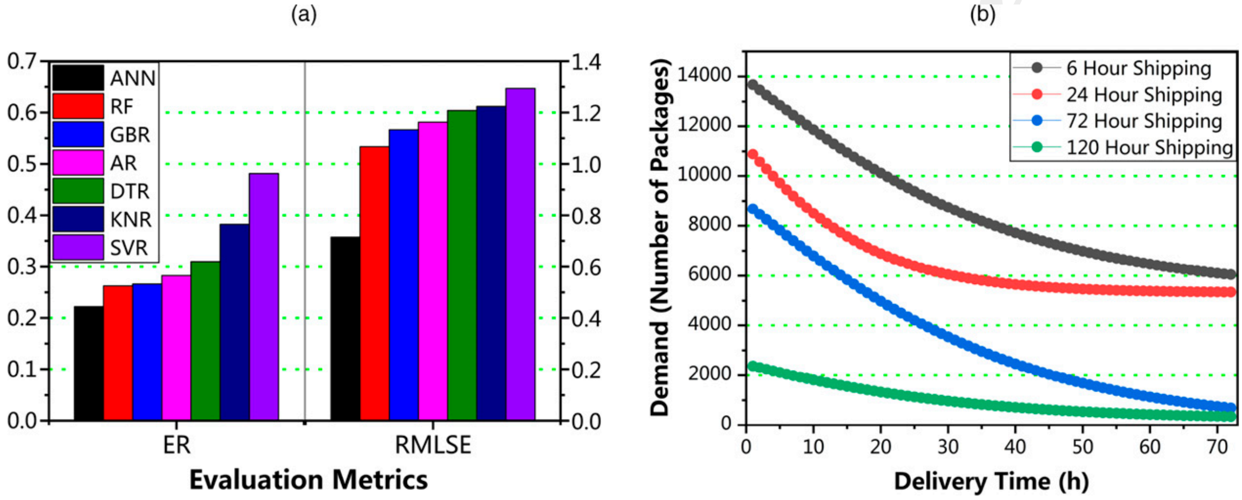
6.1.3. Sensitivity Analysis. Because the motivation of an interconnected distribution network is to offer fast shipping and delivery services, we analyze how these two factors affect the customer demand based on our optimized ANN model. By keeping other factors fixed at their average values, we vary the delivery time from 1–72 hours and the shipping time from 6–120 hours. Figure 9(b) illustrates the sensitivity of the

Figure 8. (Color online) Scatter Plots of Predicted Results from ANN vs. Ground Truth Values



Note. (a) Training, (b) validation, and (c) testing.

Figure 9. (Color online) Performance Comparison of Prediction and Factor Sensitivity Analysis



Note. (a) Overall performance comparison and (b) factor sensitivity analysis.

simulated customer demand with respect to these two factors. With varying delivery time in hours as the horizontal axis, we use different colors to represent varying shipping time periods, namely, six-hour shipping (black), one-day shipping (red), three-day shipping (blue), and five-day shipping (green). As seen from the simulated results in Figure 9(b), for a fixed shipping time, the customer demand decreases with a longer delivery time. Similarly, for a fixed delivery time, the demand decreases with a longer shipping time. The green line in Figure 9(b) indicates that if it takes 120 hours for the retailer to ship a product, the demand will be quite low (mostly lower than 2,000) regardless of the delivery time. For the fast shipping (6-hour or 24-hour shipping), the customer demand remains relatively high regardless of the delivery time. As the delivery time is mainly decided by the facility locations, it is important for online sellers to optimize the facility locations for fast logistics services and consequently higher demand. The insights based on the sensitivity analysis are helpful for the facility network design: a fast delivery requires a short facility-customer delivery distance, whereas a quick shipping time requires a fast supply lead time, which is achieved by a short supplier-facility shipping distance and a well-designed interfacility inventory transshipment distance for a balanced inventory level. In other words, a well-designed multilevel facility network can boost online customer demand.

6.2. Numerical Results for Facility Location Optimization

Next, we show the numerical results for the facility location optimization. Specifically, we first present the results showing the efficacy of the proposed two-stage heuristic for solving the facility location problem under study. Then, we show the benefits of the iterative facility network design process.

6.2.1. Performance of the Two-stage Heuristic. To test the performance of the two-stage heuristic developed in Section 5.2 on the *static* network design problem (without considering the iterative process between demand prediction and facility location optimization), two benchmark algorithms are chosen. The first benchmark is to solve the MILP Model (3)–(15) to optimality using the Gurobi MILP solver version 9.0.2 (Linux 64), with the presolve and heuristics options turned on. The second benchmark is a genetic algorithm combined with linear programming (LP), which has been shown effective in solving many large-scale facility location problems (Alp et al. 2003, Michalewicz 2013, Liu et al. 2015). Specifically, the genetic algorithm implemented here performs crossover and mutation operations on chromosomes coded for binary variables g_i in the MILP model and the remaining model with g_i fixed is an LP model and is solved using the Gurobi LP solver. There are two segments of chromosomes representing the primary facility selection and secondary facility selection, respectively. The population size is set to be 100. Uniform crossover is used to generate offsprings. The mutation operation randomly flips 5% genes coded "0" (indicating nonselection), and the corresponding numbers of genes coded "1" (indicating selection) are flipped for the two segments of chromosomes to satisfy Constraints (4) and (5). The algorithm will terminate if no better individual is generated within three successive generations or the evolution reaches 100 generations.

All the test cases were generated based on the real-life data mentioned in Section 3.1. The computational data are provided in the online supplement. To this end, we first generate small- to medium-size instances, where Gurobi can solve them to optimality within five hours. Specifically, the sizes of the primary and secondary facility candidates are set to 50 and 500, respectively, that is, $|\mathcal{V}^P| = 50$ and $|\mathcal{V}^Q| = 500$. We then vary the numbers of target primary and secondary facilities, where r^P is varied between 2 and 8 in an increment of 1 and r^Q is varied between 50 and 200 in an increment of 50.

Table 3 shows the numerical results of the heuristic compared with the two benchmarks for all cases. Recall that Gurobi solves the MILP problem to optimality and the attendant results are the baselines. It is seen that in general the computational time (CT) of Gurobi increases as r^P and r^Q increase. It takes the Gurobi MILP solver more than three hours to solve the case with 8 target primary facilities and 200 target secondary facilities. Although the genetic algorithm reduces the computational time for solving each case, it sacrifices the solution quality with an optimality gap (compared with the Gurobi baseline) ranging between 13.89% and 49.17%. In contrast, the two-stage heuristic takes less than 30 seconds to find solutions, with an optimality gap ranging between 1.21% and 11.31%, mostly within 6%.

Given that the problem under study is a design problem, it could be argued that it is worth the wait to solve the problem to optimality using Gurobi. But we point out that the value of the heuristic lies in its scalability as the size of the facility candidate pool becomes larger, which is often the case for online retailers. To this end, we provide several examples when Gurobi struggled to solve the static MILP model in days or weeks. Specifically, Table 4 shows how the computational time of Gurobi (with 0% optimality gap) increases exponentially as we increase $|\mathcal{V}^P|$ from 50 to 80 and 100 and $|\mathcal{V}^Q|$ from 500 to 800 and 1,000. It is also seen that Gurobi fails to solve the case with $|\mathcal{V}^P| = 100$, $|\mathcal{V}^Q| = 1{,}000$, $r^P = 8$, and $r^Q = 200$ to optimality within eight days. In comparison, the computational times for the heuristic for all cases remain under 40 seconds with optimality gaps within 7%. To further test how quickly Gurobi can find a solution as good as the one found by the heuristic, we use the objective obtained by the heuristic as a stopping criterion

Table 3. Experimental Results for Small- to Medium-Size Instances

		Gurobi MILP static		Genetic algorithm static			Two-stage heuristic static			Two-stage heuristic iterative			
r^P	r^Q	Obj ($K)	CT (s)	Obj ($K)	Gap	CT (s)	Obj ($K)	Gap	CT (s)	Obj ($K)	Setup ($K)	Profit ($K)	CT (s)
2	50	2,329	1,101	2,759	18.46%	850	2,528	8.55%	19	4,469	2,186	30,710	233
2	100	2,152	1,469	2,568	19.33%	856	2,314	7.50%	18	3,781	3,439	30,385	198
2	150	2,094	2,030	2,696	28.75%	814	2,119	1.21%	18	3,496	4,699	30,299	211
2	200	2,069	1,624	2,612	26.24%	905	2,157	4.23%	20	3,298	5,965	30,624	183
3	50	2,175	2,841	2,663	22.44%	913	2,303	5.87%	20	4,060	2,436	30,869	221
3	100	1,998	1,563	2,559	28.08%	1,152	2,028	1.49%	18	3,352	3,689	30,564	205
3	150	1,939	4,033	2,727	40.64%	1,332	2,105	8.55%	18	3,062	4,949	30,483	214
3	200	1,915	1,801	2,175	13.58%	1,301	1,980	3.40%	22	2,823	6,215	30,849	186
4	50	2,044	1,687	2,427	18.74%	1,078	2,114	3.46%	22	3,658	2,686	**31,021**	205
4	100	1,868	2,492	2,406	28.80%	1,059	1,887	4.18%	19	2,994	3,939	30,672	233
4	150	1,810	1,780	2,257	24.70%	1,493	1,860	2.77%	20	2,679	5,199	30,616	232
4	200	1,786	2,058	2,362	32.25%	1,584	1,857	3.99%	23	2,472	6,465	30,950	206
5	50	1,989	5,234	2,734	37.46%	1,608	2,036	2.38%	17	3,448	2,936	30,981	248
5	100	1,812	7,672	2,503	38.13%	1,773	1,887	4.18%	17	2,744	4,189	30,672	243
5	150	1,754	5,654	2,042	16.42%	1,547	1,810	3.17%	19	2,453	5,449	30,592	222
5	200	1,731	5,581	2,465	42.40%	1,334	1,838	6.21%	23	2,213	6,715	30,959	208
6	50	1,941	5,265	2,593	33.59%	1,037	2,003	3.20%	19	3,282	3,186	30,897	216
6	100	1,765	4,641	2,399	35.92%	1,806	1,939	9.84%	16	2,596	4,439	30,570	219
6	150	1,707	6,185	2,525	47.92%	1,330	1,815	6.31%	23	2,284	5,699	30,511	238
6	200	1,684	6,056	2,405	42.81%	1,405	1,850	9.83%	25	2,070	6,965	30,852	322
7	50	1,916	6,681	2,300	20.04%	2,073	2,017	5.28%	22	3,177	3,434	30,557	306
7	100	1,740	6,896	2,238	28.62%	2,214	1,929	10.86%	18	2,481	4,689	30,435	311
7	150	1,682	14,397	2,430	44.47%	3,989	1,814	7.83%	24	2,200	5,949	30,345	284
7	200	1,659	9,471	2,470	48.88%	2,348	1,694	2.10%	24	1,928	7,215	30,744	287
8	50	1,893	10,260	2,597	37.19%	4,011	2,019	6.67%	26	3,016	3,686	30,663	336
8	100	1,718	8,102	2,617	52.33%	3,054	1,912	11.31%	27	2,386	4,939	30,280	343
8	150	1,659	9,179	2,346	41.41%	3,551	1,794	8.18%	27	2,122	6,199	30,173	320
8	200	1,636	10,734	2,324	42.05%	3,098	1,715	4.87%	26	1,830	7,465	30,592	308

Note. CT, computational time.

Table 4. Experimental Results of the Static Problem for Different Sizes of Instances

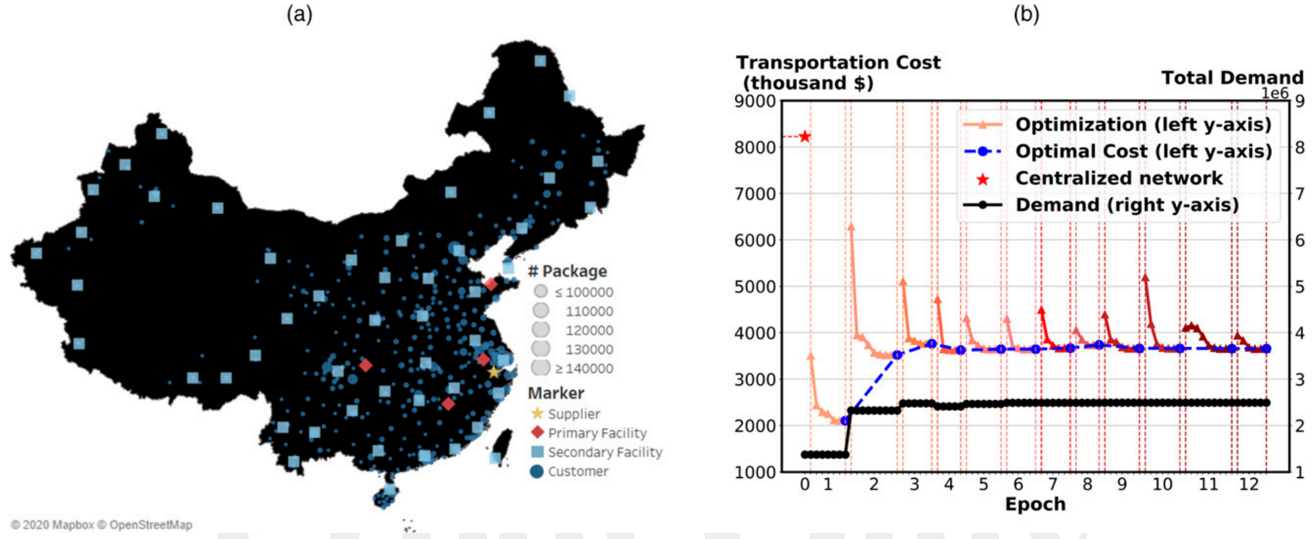
				Static Gurobi MILP (0% optimality gap)		Static two-stage heuristic			Static Gurobi MILP (BestObjStop)		
r^P	r^Q	$\lvert\mathcal{V}^P\rvert$	$\lvert\mathcal{V}^Q\rvert$	Obj ($K)	CT (s)	Obj ($K)	Gap	CT (s)	Incumbents ($K)	Gap	CT (s)
4	50	50	500	2,044	1,687	2,114	3.46%	22	2,080	1.79%	916
4	50	80	800	1,984	43,354	2,092	5.45%	27	2,083	5.01%	1,017
4	50	100	1,000	1,951	167,511	2,084	6.81%	25	2,062	5.71%	36,448
8	200	50	500	1,636	10,734	1,715	4.87%	26	1,695	3.62%	113
8	200	80	800	1,500	518,414	1,591	6.03%	31	1,565	4.33%	4,301
8	200	100	1,000	–	>691,200	1,582	–	38	1,561	–	20,677

for Gurobi. That is, when Gurobi finds an incumbent that is no worse than the objective obtained by the heuristic, the solver will stop. The performance of Gurobi under such an implementation is reported in column "Static Gurobi MILP (BestObjStop)" of Table 4. Although the computational times reduce compared with Gurobi with 0% optimality gap, the time can still be quite long and increases dramatically as the problem size increases. For example, for the case with $\lvert\mathcal{V}^P\rvert = 100$, $\lvert\mathcal{V}^Q\rvert = 1{,}000$, $r^P = 4$, and $r^Q = 50$, it takes more than 10 hours to reach the objective found by the heuristic within 25 seconds. Note that such computational time advantage of the heuristic is even more substantial when we conduct the iterative prediction-and-optimization process, to be demonstrated next. Therefore, we claim that the heuristic reaches a good balance between the solution quality and the computational time and should be deployed for large-size instances.

6.2.2. Results for Iterative Prediction-and-Optimization.

Recall that the demand distribution may change once the facility network is optimized, rendering the obtained facility network suboptimal. Therefore, we need to iteratively reoptimize the facility locations with the updated demand prediction. To this end, for each test case in Table 3, we iteratively estimate the demand distribution and optimize the multilevel facility network using Algorithm 3 until convergence. The results are shown in the last column "Two-stage heuristic iterative" of Table 3. It is seen that Algorithm 3 scales very well and most cases converge within six minutes. In Table 3, we also report the facility setup and package processing cost (the "Setup ($K)" column) and the estimated profit (the "Profit ($K)" column) for each case. The revenue is calculated using an average price spread (gap between selling price and purchase price) of $15 per package; the cost consists of the transportation cost, facility setup cost, and package processing cost. It is seen that, although more facilities can reduce the transportation distance and boost total demand, it may not necessarily increase the profit as the setup cost becomes higher. In this case study, the distribution network with 4 primary facilities and 50 secondary facilities achieves the highest total profit, which is shown in Figure 10(a).

Figure 10. (Color online) Experimental Result of the Iterative Multilevel Facility Location Optimization



Note. (a) Optimal facility network ($r^P = 4$, $r^Q = 50$) and (b) convergence progress.

Table 5. Experimental Results of the Iterative Problem for Different Sizes of Instances

				Iterative Gurobi MILP (0% optimality gap)		Iterative Gurobi MILP (5% optimality gap)			Iterative two-stage heuristic		
r^P	r^Q	$\lvert\mathcal{V}^P\rvert$	$\lvert\mathcal{V}^Q\rvert$	Profit ($K)	CT (s)	Profit ($K)	Gap	CT (s)	Profit ($K)	Gap	CT (s)
4	50	50	500	31,512	9,565	31,012	1.59%	2,004	31,021	1.56%	205
4	50	80	800	32,557	206,589	32,152	1.24%	17,759	31,785	2.37%	211
4	50	100	1,000	–	>1,382,400	32,881	–	483,217	32,147	2.23%*	208
8	200	50	500	31,002	48,160	30,514	1.57%	4,632	30,592	1.32%	233
8	200	80	800	–	>1,382,400	30,619	–	33,204	30,601	0.06%*	283
8	200	100	1,000	–	>1,382,400	30,924	–	440,350	30,885	0.13%*	280

*Indicates that the gap is computed by comparing to the result using Gurobi with 5% optimality gap.

With respect to the convergence of the iterative process, Figure 10(b) shows the convergence progress of Algorithm 3 for the case in Figure 10(a). As a baseline, the original centralized distribution network has a transportation cost of $8,225K and a total demand (number of packages) of 1.37 million. Algorithm 3 starts with the initial demand distribution and runs the two-stage heuristic to find the resulting optimal facility locations. Within this iteration (epoch 1), the curve with triangles in light-red color shows the convergence process of the heuristic for optimizing the facility locations. During epoch 1, the demand distribution remains as the initial demand of 1.37 million. With the optimized facility location at the end of epoch 1, the ANN model repredict the demand with the improved logistics service. It is seen that, after we set up a multilevel distribution network, the demand is boosted to 2.32 million. Then, the facility locations are optimized again in epoch 2 given the newly predicted demand. This iterative process continues until the demand does not change between successive epochs. In this particular case, the algorithm stops at epoch 12. As a final result, the demand is predicted to increase to 2.49 million, approximately 81.2% more than the initial demand under the centralized distribution network.

To further illustrate how the heuristic performs compared with Gurobi in the iterative network design, we ran the iterative process using three methods: Gurobi with 0% optimality gap, Gurobi with 5% optimality gap, and the two-stage heuristic. Table 5 shows the performance of three methods in terms of the objective and the computational time. For three out of six cases, Gurobi with 0% optimality gap achieves a converged result within 16 days and the corresponding profit serves as the optimality baseline for each case. For Gurobi with 5% optimality gap, all six cases converge within 16 days, with the two largest cases tested ($|\mathcal{V}^P| = 100$ and $|\mathcal{V}^Q| = 1{,}000$) converge in more than five days. On the other hand, the heuristic solves all six cases within five minutes. It is worth noting that, for the three cases with the Gurobi optimality baseline, the profits achieved by the converged network using the heuristic are only 1.56%, 2.37%, and 1.32% from the Gurobi baselines. For the three cases without Gurobi optimality baselines, the profits obtained using the heuristic are compared with those obtained using Gurobi with a 5% optimality gap. The gaps for these cases are 2.23%, 0.06%, and 0.13%, respectively, all achieved within only a small fraction of time using Gurobi. Note further that the heuristic achieves better results than Gurobi with a 5% optimality gap for the two cases with $|\mathcal{V}^P| = 50$ and $|\mathcal{V}^Q| = 500$. Given that the values of $|\mathcal{V}^P|$ and $|\mathcal{V}^Q|$ can be even larger in practical problems, the heuristic is the preferred option for solving those large cases.

In summary, using the heuristic in the iterative prediction-and-optimization framework substantially improves the scalability of the solution approach, without sacrificing the solution quality by much. We recommend the use of Gurobi when the problem size (in terms of $|\mathcal{V}^P|$ and $|\mathcal{V}^Q|$) is small or medium and recommend the use of the two-stage heuristic when the problem size becomes large.

7. Conclusion

In this paper, we develop a data-driven iterative optimization framework for the multilevel distribution network design problem for online retailers. Specifically, we first propose an artificial neural network with exogenous demographic factors and logistics service-related factors for predicting customer demand distribution. The demand distribution is subsequently used for the facility location optimization with the objective of minimizing the total facility setup cost, package processing cost, and transportation cost, including supplier to primary facility transportation cost, interprimary facility transshipment cost, primary to secondary facility transportation cost, and customer delivery cost. To efficiently optimize the large-scale multilevel facility locations with a given demand distribution, we further propose a two-stage heuristic based on an agglomerative hierarchical clustering algorithm and an expectation and maximization algorithm. Furthermore, because a new facility network may affect demand distribution, we propose an iterative process to optimize the distribution network considering the mutual interdependence between demand distribution and facility locations. Extensive experiments using a real-world data set from a leading online retailer on Taobao demonstrated the accuracy of the proposed demand predictor and the effectiveness of the proposed facility location optimization heuristic. Finally, the results show that the iterative prediction and optimization process is superior to the one-off facility location optimization based on static demand distribution.

This work has some limitations that require future research. First, this problem is originated from a real-world case; but some of the problem settings can be further generalized. We do not consider the transportation conditions, which may be important factors for some online retailers, especially for perishable products with special transportation requirements. Another study can be performed to examine the collaborations of different online sellers that are willing to share the same facility network. With multiple online sellers sharing one facility network, the demand predictor and optimization heuristic developed in this paper need to be revised to consider the collaboration and competition among these sellers. Further, although the two-stage heuristic has been shown effective in solving large-scale facility location problems under study, it may be worthwhile to investigate alternative heuristics under the mathematical programming framework, such as Lagrangian relaxation and Benders decomposition approaches with problem-specific heuristics. Finally, the facility location problem considered in this paper is a deterministic problem. Its stochastic counterpart can be studied considering the demand uncertainty and other supply chain disruptions.

Acknowledgments

The authors are grateful to the area editor, the anonymous associate editor, and the two anonymous reviewers for their constructive comments and suggestions that led to a substantially improved paper.

Endnotes

[1] See https://en.wikipedia.org/wiki/Taobao.

[2] See https://en.wikipedia.org/wiki/Cainiao.

[3] See https://xiaoye.world.tmall.com/.

[4] See https://www.sf-express.com.

References

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, et al. (2016) Tensorflow: A system for large-scale machine learning. *Proc.* 12th USENIX Conf. Oper. Systems Design Implementation (OSDI 16) (USENIX Association, Berkeley, CA), 265–283.

Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H (2018) State-of-the-art in artificial neural network applications: A survey. *Heliyon* 4(11): e00938 .

Alp O, Erkut E, Drezner Z (2003) An efficient genetic algorithm for the p-median problem. *Ann. Oper. Res.* 122(1-4):21–42.

Bohlmeijer E, Ten Klooster PM, Fledderus M, Veehof M, Baer R (2011) Psychometric properties of the five facet mindfulness questionnaire in depressed adults and development of a short form. *Assessment* 18(3):308–320.

Boysen N, de Koster R, Weidinger F (2019) Warehousing in the e-commerce era: A survey. *Eur. J. Oper. Res.* 277(2):396–411.

Bucko J, Kakalejčík L, Ferencová M (2018) Online shopping: Factors that affect consumer purchasing behaviour. *Cogent Bus. Management* 5(1):1–15.

Chen C, Liu J, Li Q, Wang Y, Xiong H, Wu S (2017) Warehouse site selection for online retailers in inter-connected warehouse networks. *Proc. 2017 IEEE Internat. Conf. Data Mining (ICDM)* (IEEE, Piscataway, NJ), 805–810.

Cheney W, Kincaid D (2009) *Linear Algebra: Theory and Applications* (Jones & Bartlett Learning, Burlington, MA).

Chong AYL, Ch'ng E, Liu MJ, Li B (2017) Predicting consumer product demands via big data: The roles of online promotional marketing and online reviews. *Internat. J. Production Res.* 55(17): 5142–5156.

Contreras I, Cordeau JF, Laporte G (2011) Benders decomposition for large-scale uncapacitated hub location. *Oper. Res.* 59(6): 1477–1490.

Cui R, Li M, Li Q (2020) Value of high-quality logistics: Evidence from a clash between sf express and Alibaba. *Management Sci.* 66(9):3879–3902.

Do CB, Batzoglou S (2008) What is the expectation maximization algorithm? *Nature Biotechnology* 26(8):897–899.

Drucker H, Burges CJ, Kaufman L, Smola AJ, Vapnik V (1997) Support vector regression machines. Jordan MI, Petsche T, eds. *Proc. 9th Internat. Conf. Neural Inform. Processing Systems* (MIT Press, Cambridge, MA), 155–161.

Duan Q, Liao TW (2013) Optimization of replenishment policies for decentralized and centralized capacitated supply chains under various demands. *Internat. J. Production Econom.* 142(1):194–204.

Fernie J, Sparks L (2018) *Logistics and Retail Management: Emerging Issues and New Challenges in the Retail Supply Chain* (Kogan Page Publishers, London).

Ferreira KJ, Lee BHA, Simchi-Levi D (2015) Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing Service Oper. Management* 18(1):69–88.

Fischetti M, Ljubić I, Sinnl M (2017) Redesigning benders decomposition for large-scale facility location. *Management Sci.* 63(7): 2146–2162.

Fisher ML (2004) The Lagrangian relaxation method for solving integer programming problems. *Management Sci.* 50(12, supplement):1861–1871.

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Ann. Statist.* 29(5):1189–1232.

Gendron B, Khuong PV, Semet F (2016) A Lagrangian-based branch-and-bound algorithm for the two-level uncapacitated facility location problem with single-assignment constraints. *Transportation Sci.* 50(4):1286–1299.

Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning* (MIT Press, Cambridge, MA).

Hassoun MH (1995) *Fundamentals of Artificial Neural Networks* (MIT Press, Cambridge, MA).

Hastie T, Tibshirani R (1996) Discriminant adaptive nearest neighbor classification and regression. Touretzky D, Mozer M, Hasselmo M, eds. Advances in Neural Information Processing Systems 8 (MIT Press, Cambridge, MA), 409–415.

Hübner A, Holzapfel A, Kuhn H (2015) Operations management in multi-channel retailing: An exploratory study. *Oper. Management Res.* 8(3):84–100.

Hübner A, Kuhn H, Wollenburg J (2016) Last mile fulfilment and distribution in omni-channel grocery retailing: A strategic planning framework. *Internat. J. Retail Distribution Management* 44(3): 228–247.

Karlik B, Olgac AV (2011) Performance analysis of various activation functions in generalized MLP architectures of neural networks. *Internat. J. Artificial Intelligence Expert Systems* 1(4):111–122.

Liaw A, Wiener M (2002) Classification and regression by random forest. *R News* 2(3):18–22.

Lim SFW, Jin X, Srai JS (2018) Consumer-driven e-commerce: A literature review, design framework, and research agenda on last-mile logistics models. *Internat. J. Physical Distribution Logist. Management* 48(3):308–332.

Lin Y (2019) E-urbanism: E-commerce, migration, and the transformation of Taobao villages in urban China. *Cities* 91:202–212.

Liu J, Sun L, Chen W, Xiong H (2016) Rebalancing bike sharing systems: A multi-source data smart optimization. Proc. 22nd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining (ACM, New York), 1005–1014.

Liu J, Li Q, Qu M, Chen W, Yang J, Xiong H, Zhong H, Fu Y (2015) Station site optimization in bike sharing systems. 2015 IEEE Internat. Conf. Data Mining (IEEE, Piscataway, NJ), 883–888.

Loh WY (2014) Classification and regression tree methods. *Wiley StatsRef: Statistics Reference Online.* Q:20

Lu Y, Liu Z, Ma L (2020) Competing through logistics management: Studies on e-retailing in China. *Supply Chain and Logistics Management: Concepts, Methodologies, Tools, and Applications* (IGI Global, Hershey, PA), 1075–1094.

Masood S, Sharif M, Masood A, Yasmin M, Raza M (2015) A survey on medical image segmentation. *Current Medical Imaging* 11(1): 3–14.

Matarazzo TJ, Pakzad SN (2016) Stride for structural identification using expectation maximization: Iterative output-only method for modal identification. *J. Engrg. Mechanics* 142(4). Q:21

Melo MT, Nickel S, Saldanha-Da-Gama F (2009) Facility location and supply chain management–A review. *Eur. J. Oper. Res.* 196(2):401–412.

Michalewicz Z (2013) *Genetic Algorithms+ Data Structures= Evolution Programs* (Springer Science & Business Media, Berlin).

Mukherjee I, Routroy S (2012) Comparing the performance of neural networks developed by using Levenberg–Marquardt and quasi-newton with the gradient descent algorithm for modelling a multiple response grinding process. *Expert Systems Appl.* 39(3):2397–2407.

Ortiz-Astorquiza C, Contreras I, Laporte G (2018) Multi-level facility location problems. *Eur J. Oper. Res.* 267(3):791–805.

Ortiz-Astorquiza C, Contreras I, Laporte G (2019) An exact algorithm for multilevel uncapacitated facility location. *Transportation Sci.* 53(4):1085–1106.

Panagiotelis A, Smith MS, Danaher PJ (2014) From Amazon to Apple: Modeling online retail sales, purchase incidence, and visit behavior. *J. Bus. Econom. Statist.* 32(1):14–29.

Ponce D, Contreras I, Laporte G (2020) E-commerce shipping through a third-party supply chain. *Transportation Res. Part E: Logist. Transportation Rev.* 140:101970.

Ranganathan A (2004) The Levenberg-Marquardt algorithm. *Tutorial LM Algorithm* 11(1):101–110.

Rao S, Goldsby TJ, Griffis SE, Iyengar D (2011) Electronic logistics service quality (e-lsq): Its impact on the customer's purchase satisfaction and retention. *J. Bus. Logist.* 32(2):167–179.

Rardin RL, Rardin RL (1998) *Optimization in Operations Research*, vol. 166 (Prentice Hall, Upper Saddle River, NJ).

Rohmer S, Gendron B (2020) *A guide to parcel lockers in last mile distribution–Highlighting challenges and opportunities from an OR perspective* (CIRRELT, Montreal).

Singh S, Rana R (2018) Effect of demographic factors on consumers' perception of online shopping. *Global J. Management Bus. Res.* 18(6):6–E.

Solomatine DP, Shrestha DL (2004) Adaboost.RT: A boosting algorithm for regression problems. 2004 IEEE Internat. Joint Conf. Neural Networks, vol. 2 (IEEE, Piscataway, NJ), 1163–1168.

Song G, Zhan Y, Guo Y (2016) The effectiveness of online shopping characteristics and logistics service on satisfaction. 13th Internat. Conf. Service Systems Service Management (ICSSSM) (IEEE, Piscataway, NJ), 1–6.

Speranza MG (2018) Trends in transportation and logistics. *Eur. J. Oper. Res.* 264(3):830–836.

Subramanian N, Gunasekaran A, Yu J, Cheng J, Ning K (2014) Customer satisfaction and competitiveness in the Chinese e-retailing: Structural equation modeling (SEM) approach to identify the role of quality factors. *Expert Systems Appl.* 41(1):69–80.

Wang G, Gunasekaran A, Ngai EW, Papadopoulos T (2016) Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *Internat. J. Production Econom.* 176:98–110.

Wu CJ (1983) On the convergence properties of the EM algorithm. *Ann. Statist.* 11(1):95–103.

Xu D, Tian Y (2015) A comprehensive survey of clustering algorithms. *Ann. Data Sci.* 2(2):165–193.

Yu H, Wilamowski BM (2011) Levenberg-Marquardt training. Wilamowski BM, Irwin JD, eds. *The Industrial Electronics Handbook* (CRC Press, Boca Raton), 12-1–12-16.

Yu J, Subramanian N, Ning K, Edwards D (2015) Product delivery service provider selection and customer satisfaction in the era of internet of things: A Chinese e-retailers' perspective. *Internat. J. Production Econom.* 159:104–116.

Yu Y, Wang X, Zhong RY, Huang G (2017) E-commerce logistics in supply chain management: Implementations and future perspective in furniture industry. *Indust. Management Data Systems* 117(10):2263–2286.

Zetina CA, Contreras I, Cordeau JF (2019a) Exact algorithms based on benders decomposition for multicommodity uncapacitated fixed-charge network design. *Comput. Oper. Res.* 111:311–324.

Zetina CA, Contreras I, Cordeau JF (2019b) Profit-oriented fixed-charge network design with elastic demand. *Transportation Res. Part B: Methodological* 127:1–19.

Zhang L, Suganthan PN (2016) A survey of randomized algorithms for training neural networks. *Inform. Sci.* 364:146–155.

Zhang C, Tian YX, Fan ZP, Liu Y, Fan LW (2020) Product sales forecasting using macroeconomic indicators and online reviews: A method combining prospect theory and sentiment analysis. *Soft Comput.* 24:6213–6226.

Zhu F, Liu Q (2018) Competing with complementors: An empirical look at Amazon. com. *Strategic Management J.* 39(10):2618–2642.

NUMBER 1 OF 2

AUTHOR QUERIES

DATE 7/29/2021

JOB NAME IJOC

ARTICLE 20211107

QUERIES FOR AUTHOR Liu et al.

THIS QUERY FORM MUST BE RETURNED WITH ALL PROOFS FOR CORRECTIONS

Q: 1_Your color figures appear in color in your page proof to simulate their presentation in color online, but they will be converted to grayscale in the print version of your article because you have not requested color processing for print. Please (a) check your figures to confirm that they will be sufficiently clear in grayscale presentation in the print version of your article and (b) modify your figure legends as necessary to remove any references to specific colors in the figure.

Q: 2_Please provide all funding information, if applicable, including the names of the institutions as well as any grant numbers associated with the funding, or confirm that no funding was received.

Q: 3_Please confirm that the article title, author names, affiliations, and email addresses are set correctly. If applicable, please provide author ORCID numbers.

Q: 4_Please provide the city and postal codes for affiliations "b," "c," and "d" and the city name for affiliation "a."

Q: 5_Please confirm that the added term "strategy" preserves your intent in "We propose an iterative prediction-and-optimization strategy…"

Q: 6_Both "E-logistics" and "e-logistics" were used in the article. For consistency, "e-logistics" has been retained throughout the article text. Please confirm that the edit preserves your intent.

Q: 7_Please confirm that keywords are correct as set.

Q: 8_Please confirm that heading levels are correct as set.

Q: 9_Please verify that all displayed equations and in-text math notations are set correctly.

Q: 10_Per INFORMS style, all variables should be italic and all vectors should be bold. Please confirm that all terms have been formatted properly throughout.

NUMBER 2 OF 2

Q: 11_Per journal style, the term “above” is not used to refer to numbered parts of the article. Please confirm that the change of “the above definitions” to “Definitions 1 and 2” preserves your intent.

Q: 12_Page numbers are only required for direct quotes. Please clarify if a direct quote accompanies “Goodfellow et al. 2016” and in reference citations elsewhere in the article where reference page numbers are provided but no apparent direct quote.

Q: 13_Please confirm/correct the change of “paragraph followed” to “paragraph that follows” in the sentence beginning “Constraint (14) specifies...”

Q: 14_Please confirm/correct the change of “decide where the primary facilities locate” to “decide where to locate the primary facilities.”

Q: 15_Please confirm that the change of “subfigure (a),” Subfigures (b) and (c),” and so forth to “Figure 6(a),” Figure 6, (b) and (c),” and so forth, preserves your intent.

Q: 16_Please confirm/correct the change of “closely neighbor” to “closest neighbor” in the sentence beginning “Note that the...”

Q: 17_Please confirm that “extract demand influential factors” can be changed to “extract influential demand factors.”

Q: 18_Per journal style, sentences do not begin with mathematical notation. Please recast the sentence beginning “ represents the...”

Q: 19_Please clarify if “K” represents “thousand” in “$8,225K.” If so, can it be removed, as “thousand” is already expressed in numbers here?

Q: 20_For the Loh (2014) reference. please provide the URL address that links to this article and your complete accessed date (month/day/year).

Q: 21_Please provide the page numbers for the Matarazzo, Pakzad (2016) reference.

Q: 22_Please confirm edits to the figure titles and legends regarding color.

Q: 23_Please add a heading to the second and third columns in Table 1.