

Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

RiRiShun Logistics: Home Appliance Delivery Data for the 2021 MSOM Data Driven Research Challenge

Xiaolong Guo, Yugang Yu*

Anhui Province Key Laboratory of Contemporary Logistics and Supply Chain, International Institute of Finance, School of Management, University of Science and Technology of China, Hefei 230026 China, gxl@ustc.edu.cn, ygyu@ustc.edu.cn

Gad Allon

Wharton School, University of Pennsylvania, Philadelphia, PA, 19104-6243, USA, gadallon@wharton.upenn.edu

Meiyan Wang

Branding & Marketing Department, Haier Group, 1 Haier Road, Qingdao 266103, Shandong, China, wangmeiy@haier.com

Zhentai Zhang

IT Department, RRS Supply Chain Technology Co., Ltd., 1 Haier Road, Qingdao 266103, Shandong, China, zhangzt@haier.com

To support the 2021 MSOM Data Driven Research Challenge, *RiRiShun Logistics* (a Haier group subsidiary focusing on logistics service for home appliances) provides MSOM members with logistics operational-level data for data-driven research. This paper provides a detailed description of the data associated with over 14 million orders from 149 clients (the consigners) associated with 4.2 million end consumers (the recipients and end-users of the appliances) in China, involving 18 thousand SKUs operated at 103 warehouses. Researchers are welcomed to develop econometric models, data-driven optimization techniques, analytical models, and algorithm designs by using this dataset to address questions suggested by company managers.

Key words: Logistics; Delivery Data; MSOM Society; Data-Driven Research

* Corresponding author: Yugang Yu. Email address: ygyu@ustc.edu.cn

1. Introduction

Online retailers play an important role in home-appliance retailing. According to the home appliance market report for 2019¹, more than 40% of home appliances are sold online, totaling about 367 billion RMB in transaction value in China. Unlike basic consumer products, home appliances are usually bulky and heavy, and their delivery services usually require special equipment and involve special installation procedures. These requirements create many new and challenging problems at both the strategic and operational levels, such as truck loading, routing, integrated delivery and installation, and transshipment based on cloud warehousing.

*RiRiShun Logistics*² (RRS) is the leading logistics enterprise focusing on home appliance delivery and installation in China. As a subsidiary of Haier³, RRS designed the first industry standard for home appliance logistics. RRS provides solutions for transportation, warehousing, distribution, delivery, installation, and after-sale services in the entire supply chain process of home appliances, serving manufacturers, resellers, retail platforms, and other firms.

RRS has built a professional, standardized, open, and smart logistics platform with 7 national central distribution centers (CDCs), 26 regional distribution centers (RDCs), 100 local transfer centers (LTCs), and more than 6000 last-mile hubs. In particular, as shown in Figure 1, RRS's service coverage (mainland China) is divided into seven areas; each area has one CDC, which is responsible for leading the logistics service for the area. The platform has 26 RDCs at the second level (which can be considered a provincial level), and 100 LTCs at the third level (a city cluster level) were established.

Typically, each province has one RDC or CDC that provides the logistics service for the province. However, because the population density of the eastern and southern parts of China is much higher than that in the northern and western parts, the development of e-commerce in the eastern and southern provinces is also more advanced than in the latter areas. As a result, RRS has a much denser logistics network for the former areas. For example, both Jiangsu and Fujian provinces have two RDCs, Zhejiang province has one RDC (besides having one CDC for the eastern China area), and Guangdong province has

¹<http://www.cena.com.cn/special/2020jdscbg/> (in Chinese).

²<https://www.rrswl.com/>.

³<https://www.haierappliances.com/>.

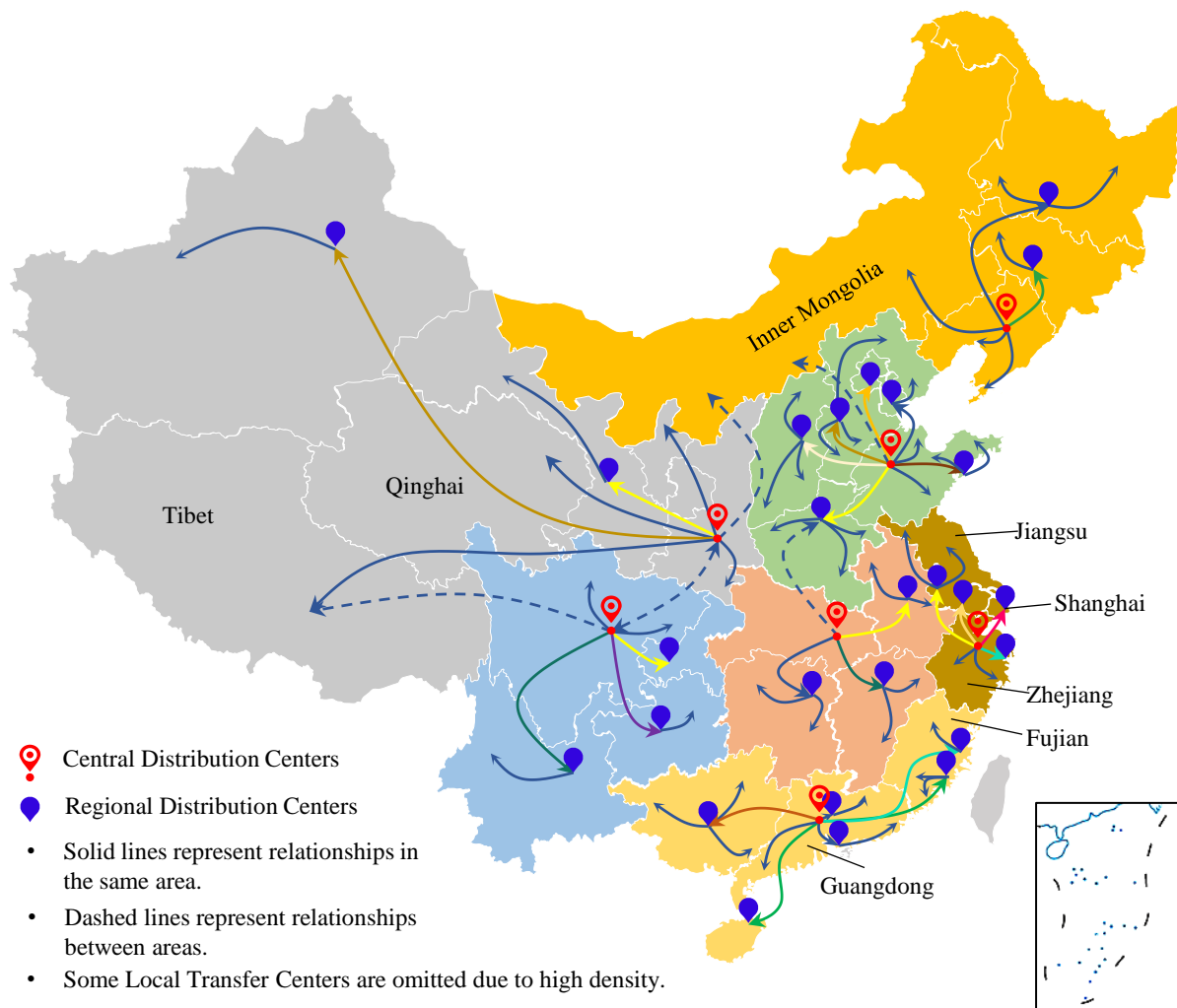


Figure 1 Logistics network of Ririshun Logistics

two RDCs (besides a CDC for the southern China area). In contrast, there are no RDCs or CDCs in lean population regions such as the Tibet Autonomous Region, Inner Mongolia Autonomous Region, or the province of Qinghai, all of which are located in western or northern China. For these regions, logistics demands will be served by the nearest RDCs in neighboring provinces. At the third level, at least one LTC has been established for each province (except the Tibet Autonomous Region) to serve the last-mile hubs. More than 6000 last-mile hubs exist across mainland China, and these are responsible for the last-mile delivery service at the county and town level.

Through the logistics network described above, RRS provides both B2B and B2C logistics services for manufacturers, resellers, retail platforms, and others. Since the data offered by RRS is based on B2C services, we shall focus our discussion on the introduction of the

B2C service supply chain in this article. To illustrate RRS’s B2C service, we take a seller on Tmall (an online marketplace of Alibaba) as an example. There are two seller fulfillment policies: (i) fulfillment by *Cainiao* (the logistic platform operator of Alibaba), and (ii) fulfillment by the seller itself. As illustrated in Figure 2, the supply chain process associated with these two policies is as follows: (1) the seller places a replenishment order with the manufacturer (such as Haier). (2) The seller delivers the replenishment information either (i) to *Cainiao*, who will forward this information to RRS if fulfillment by *Cainiao* policy is applied; or (ii) to RRS directly, if fulfillment by the seller is applied. (3) Products are delivered to RRS warehouses according to the replenishment order details and logistics service contracts (details are offered in the following paragraph). (4) Consumers place purchase orders with the seller on the Tmall marketplace, and the seller delivers detailed information (i) to *Cainiao*, who will forward the information to RRS if the order is fulfilled by *Cainiao* (but the actual delivery is provided by RRS); or (ii) to RRS directly if the order is fulfilled by seller and RRS is responsible for completing the delivery. (5) Products are delivered to end consumers through the logistics network by RRS.

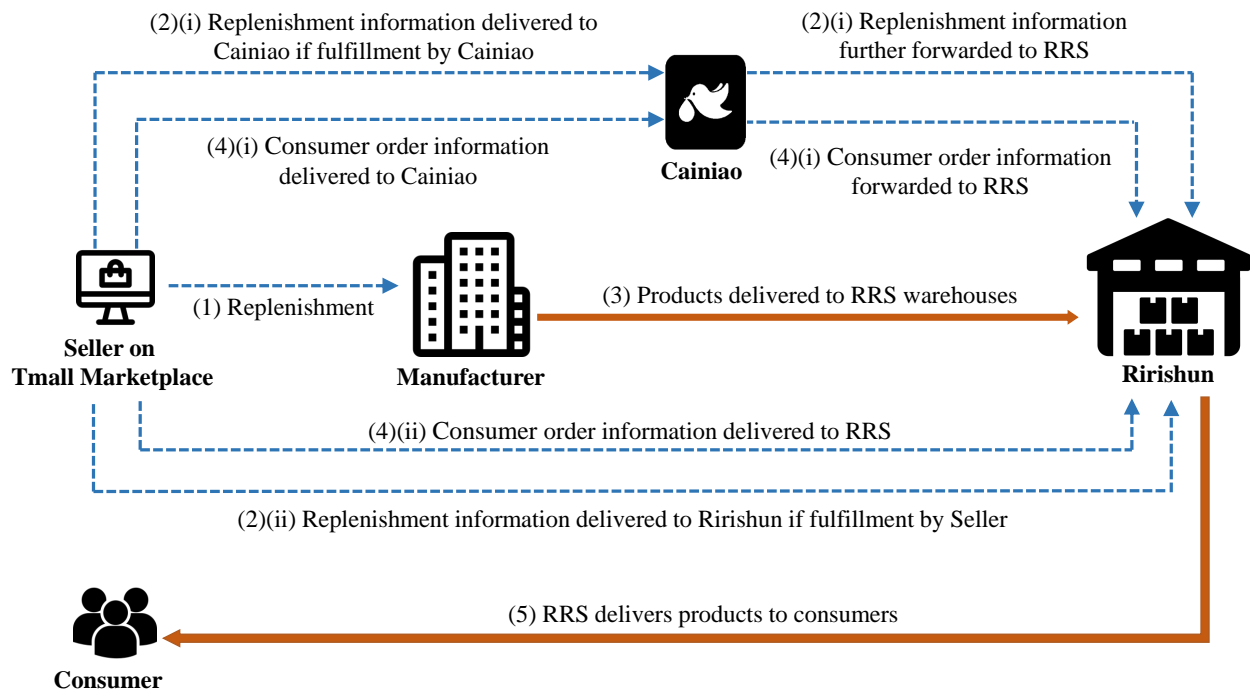


Figure 2 Supply chain structure of B2C logistics service

In the above example, the seller is the client (referred to as the “consigner” in this

article) of RRS if consumer orders are fulfilled by the seller itself. However, *Cainiao* will become the client of RRS if the seller adopts a policy of fulfillment by *Cainiao*. Specifically, in this case, *Cainiao* may forward only the detailed information of products and recipients to RRS, without telling RRS the seller’s information. Similarly, in the case of fulfillment by the seller, it is not necessary to reveal the retail channel to RRS, even though the seller may sell through multiple channels (such as Tmall, JD.com, or Suning). We should note that the manufacturer (such as Haier) can also sell its products directly through the online marketplace, which will make the manufacturer a client of RRS.

As a provider of logistics service, RRS signs contracts with these consigners (i.e., sellers or *Cainiao*) and charges them directly, instead of the end consumers (i.e., the recipients), for providing delivery services. Besides the order delivery cost, the consigners are also required to pay a rental fee for the warehouse space at RRS, a fee which varies across different distribution centers. Hence, the required warehouse space for specific centers (including all CDCs, RDCs, and LTCs) will be included in the service contract, which implies that the inventory decisions are made by the consigners, and RRS only stores and delivers the products on their behalf.

The dataset provided by RRS captures a “full delivery service cycle” that begins when the orders are transferred into the RRS system and ends when the delivery is signed for (if no installation is needed) or the installation is completed (if installation is required), or the consumer refused to sign for the delivery. Detailed information on more than 14 million orders from 149 clients during the period October 2018 to September 2019 are provided in this dataset⁴.

It is worth noting the differences between this dataset and the dataset from the *2018 MSOM Data Driven Research Challenge*⁵. Firstly, different from the *Cainiao* data, which provides delivery details for small and daily packages, this dataset contains delivery details for home appliance products that are usually bulky and heavy. Detailed shape characteristics are given, including length, width, height, volume, and weight of each product, which allows investigation of the scheduling problem for truck loading. Secondly, the detailed logistics network of the company is given in this dataset. The panorama of the network

⁴ The data are extracted from RRS’s database within a specific time window, but due to the effect of time setting and the enormous amount of data, the exact number of orders in the tables varies. For example, the *orders* table contains 14.7 million orders and the *appointment details* table contains 19.3 million orders.

⁵ <https://connect.informs.org/msom/events/datadriven2018>

can be pictured based on the distance information between any two centers in the network. Thirdly, the detailed transfers of products among different centers are provided in the dataset. Considering the products stored in a focal center, it would be interesting to find the optimal cross-docking policy. Finally, our dataset is at a logistics ecosystem level since the clients of RRS include retailers, logistic operators (such as Cainiao), and manufacturers. Researchers are welcome to take advantage of these unique features of the dataset to conduct data-driven research, including developing econometric models, data-driven optimization techniques, analytical models, and intelligent algorithms. The following section provides a detailed description of the dataset, and suggested research questions are given in Section 3.

2. Dataset Description

In this section, we provide a detailed description of the dataset offered by RRS. Certain key identification information such as the user ID, SKU ID, person names, and cell phone numbers are fully anonymized to ensure confidentiality.

Before describing the data, let us recall that the entire delivery cycle begins with the issuing of the order and ends with installation completion (if applicable) or signed/refused-for delivery (if no installation by the delivery service is needed). Firstly, the roles of distribution centers for a specific order can be specified as follows:

- *Origin Center*: the distribution center where the required product of the order is stored. (Each order has one *Origin Center*.)
- *Destination Center*: the distribution center that provides local services associated with the order. An order sent out from a *Destination Center* is delivered to the last-mile hub or directly to the end consumer. Each order has one *Destination Center*, and it is possible that the *Destination Center* is also the *Origin Center* for a specific order.
- *Transfer Center(s)*: If a transfer is required for the products of an order to arrive at its *Destination Center* from its *Origin Center*, then any distribution center (CDC, RDC, or LTC as defined above) that provides the transfer service is defined as a *Transfer Center* for this order. (It should be noted that more than one *Transfer Center* may be required for a specific order.)

Recall the logistic network described in the previous section. All distribution centers are centrally managed by the headquarter of RRS. That is, all the logistic choices, including shipping routes and which center provides the product, are centralized decisions made by

the RRS headquarters to minimize the total delivery cost within the promised delivery time. As a result, the centers just handle the detailed operational instructions from headquarters for each order. As shown in Figure 3, an order will be processed by the system according to its product and delivery information transferred into the system. Then the operational commands will be delivered to the product’s *Origin Center*. According to the processing details, the following cases may occur:

(1) The *Origin Center* is the *Destination Center* when the last-mile delivery is not required: This case happens if the order’s destination is near the *Origin Center* and the order can be delivered (with installation service) directly from the *Origin Center* to the consumer.

(2) The *Origin Center* is the *Destination Center*, but hub delivery is required: In this case, the product will be first shipped to the last-mile hub, and then the hub will complete the delivery, providing installation service if needed.

(3) The *Origin Center* is not the *Destination Center* but *Transfer Centers* are not required: The product will be shipped from the *Origin Center* directly to the *Destination Center* through trunk line transportation, and then the *Destination Center* completes the delivery service of the order.

(4) The *Origin Center* is not the *Destination Center* and transfer service of at least one *Transfer Center* is required: the only difference from case (3) is that the product will require one or more *Transfer Centers*. This case will happen when the *Origin Center* is quite far away from the *Destination Center*.

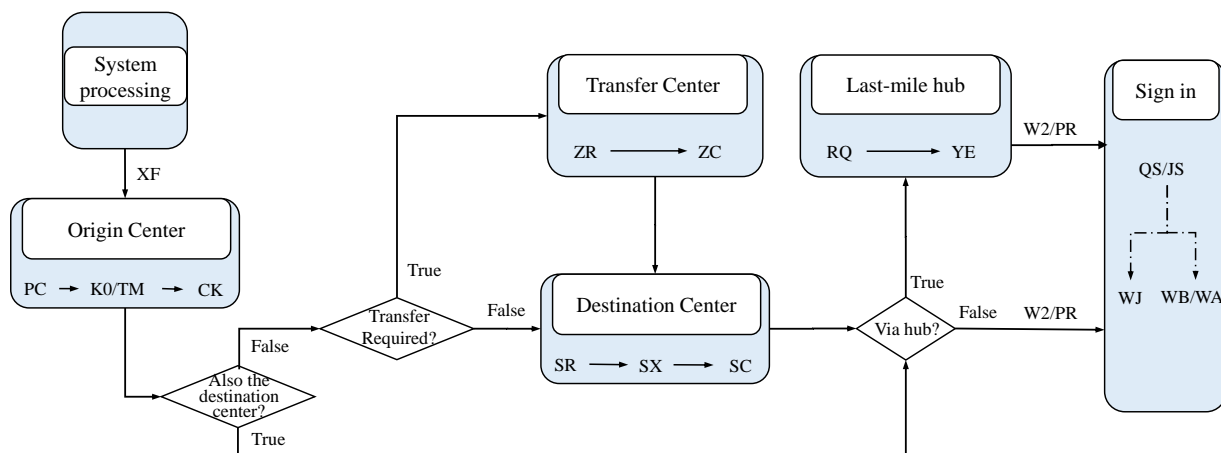


Figure 3 Detailed delivery process for a specific order (Appendix A explains each operation node in this flowchart)

A detailed description for each exact step (called an “operation node” in the data) that occurred during the delivery process is given in Table A.1 in Appendix A. To understand the delivery cycle for orders, RRS provides detailed data for each process and operation node. The dataset consists of the following: (1) orders, (2) SKU details, (3) appointment details, (4) delivery details, and (5) consumer details. In the following subsections, we give detailed descriptions for each of these tables.

2.1. Table: Orders

The *Orders* table (Table 1) provides the information for over 14 million RRS orders from October 2018 to September 2019. The unique identifier for each order is the *order_no*, which allows the order to be identified in all five tables in the dataset. For processing convenience, RRS also assigns each order an identification code for this table, labeled as *rrs_order_id*.

Field	Description	Data type	Sample value
<i>rrs_order_id</i>	RRS database identification	bigint	b9cc5c6cf7775dc4db157701370848f8
<i>order_no</i>	Order unique identifier	varchar	0d0a09d33b1190313a392d619e9d223a
<i>order_date</i>	Order time	datetime	2019-05-27 13:46:00
<i>client_code</i>	Client unique identifier	varchar	RRS18
<i>total_amt</i>	Product quantity in the order	int	2
<i>origin_center_code</i>	Origin center for the order	varchar	RRSZX072
<i>destination_center_code</i>	Destination center of the order	varchar	RRSZX076
<i>delivery_method</i>	Delivery method	int	2
<i>arrive_org_code</i>	Last-mile hub for the order	varchar	rrs_wd_3927
<i>distc_oper_dest</i>	Direct travel distance from the Origin Center to the Destination Center (km)	int	230
<i>distc_dest_org</i>	Direct travel distance from the Destination Center to the Last-mile hub (km)	int	120
<i>system_time</i>	System record create time	timestamp	2019-06-01 00:00:00

Table 1 Description of the Orders table

The 14 million orders cover one year of order time (*order_date*) from 149 order channels (also known as clients, labeled by *client_code*). Here, *client_code* is the unique identifier for the clients of RRS. Clients are not necessarily retailers; a client could also be the platform responsible for delivery for the retailers, such as Cainiao or even a manufacturer such as mi.com. As a result, we can only consider the channels as clients of RRS and cannot consider them as online retailers. Among all the 149 order channels (i.e., clients), “RRS18” is the largest source and contributed about 12 million orders; “RRS23” is the second large

source and contributed more than 1 million orders. Figure 4 shows the distribution of RRS’s clients ordered by the number of orders contributed.

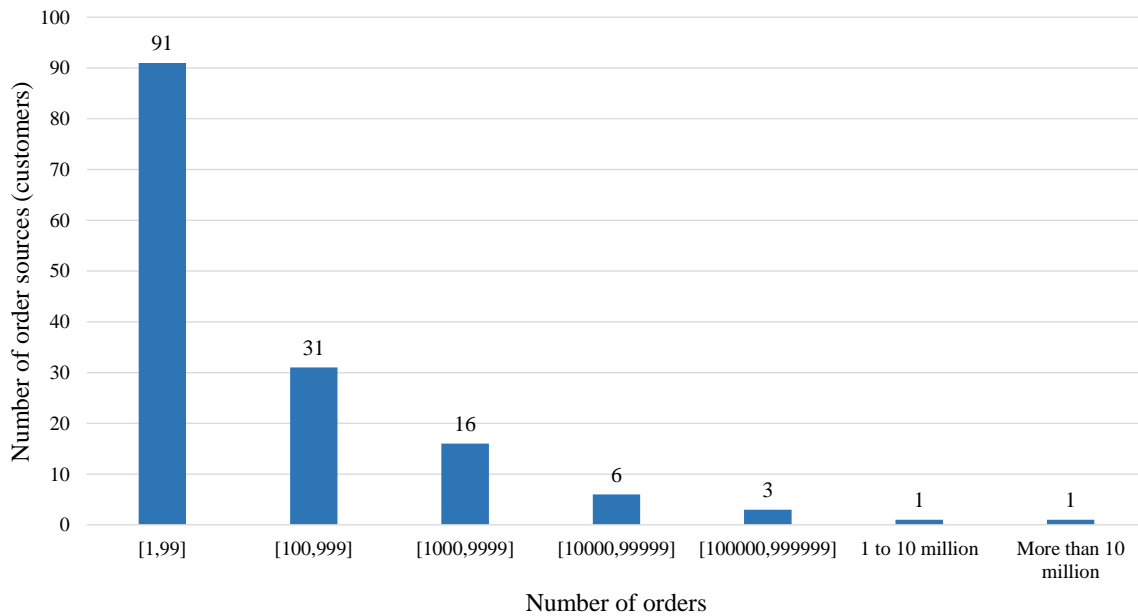
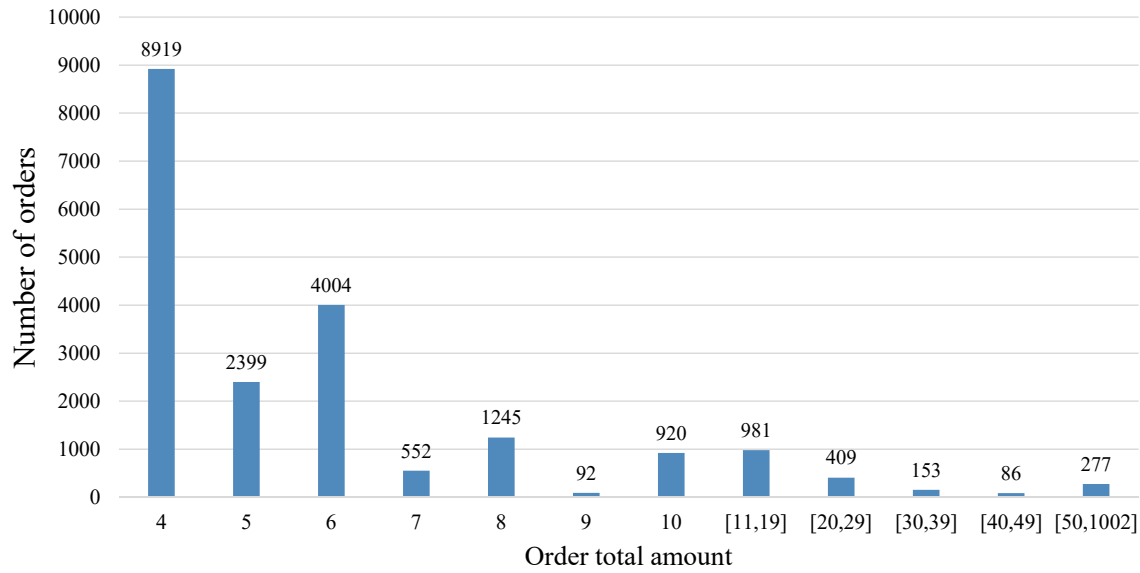


Figure 4 Distribution of order sources

Furthermore, each order may contain more than one product; the quantity ranges from 1 to 1002 (*total_amt*). About 11 million orders contain only one product (i.e., *total_amt*=1), 3 million orders contain two products, and 31 thousand orders contain three products. As a result, only 2,000 orders (0.136% of all orders) contain more than three products. The detailed distribution for these orders is given in Figure 5.

This table also gives the detailed code for the *Origin Center* and *Destination Center* for each order. Theoretically, the logistics cost will be minimized only when the *Origin Center* is exactly the *Destination Center* for each order. However, considering the effect of operating costs (such as reordering cost, warehouse rent, and inventory cost), the distribution centers are practically designed into different levels for various logistics functions. These levels include CDCs at the national level, RDCs at the provincial level, LTCs at the city cluster level, and more than 6000 last-mile hubs at the county and town level. A total of 103 distribution centers (including 7 CDCs, 26 RDCs, and 70 LTCs) are involved in this dataset, and a detailed list of distribution centers for each type is given in Table A.2 in Appendix B. Figure 6 shows the number of orders each center serviced as an *Origin Center* and as a *Destination Center*.



Note: Number of orders with *total_amt*=1, 2, and 3 are 11,280,564, 3,376,571, and 31,585, respectively.

Figure 5 Distribution of order total amount (orders with total amount less than four are excluded)

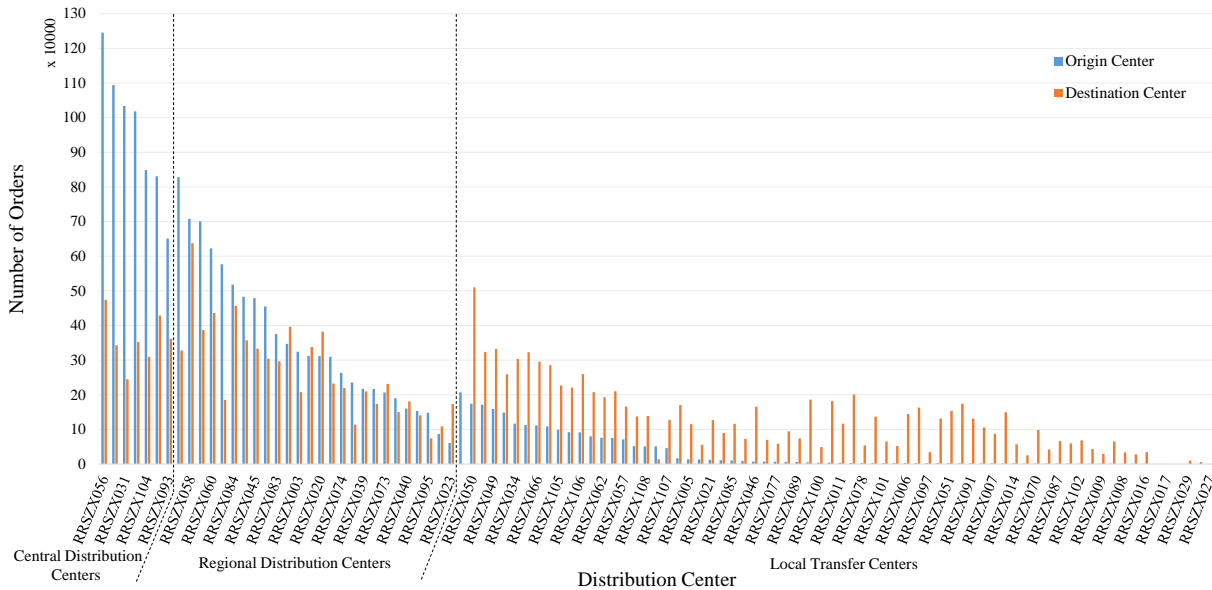


Figure 6 Number of orders served by a center as an Origin Center and as a Destination Center

From the perspective of supply and demand, the number of orders serviced as an *Origin Center* is the supply of the center, and the number of orders serviced as a *Destination Center* is the demand. As we can see, a considerable imbalance exists between supply and demand, and transshipment will occur when products need to be shipped from the *Origin Center* to the *Destination Center*. This phenomenon inspires investigation of the

performance of the distribution network. Operational questions, including transshipment policy and inventory control policy, are all worthy of investigation.

After the products in the order arrive at the *Destination Center*, they can be delivered to the consumer directly from the center (*delivery_method=70*), or distributed to the last-mile hub (labeled as *arrive_org_code*) and then delivered to the consumer through the hub (*delivery_method=2*). About 80% of all the B2C orders are delivered through the hub since home appliance products are normally quite heavy and have a large volume.

Furthermore, the haul distances (travel distance in kilometers) from the *Origin Center* to the *Destination Center* and from the *Destination Center* to the last-mile hub are provided to estimate the delivery cost for each order. Lastly, *system_time* is the timestamp representing the time when the order record was created in the system.

2.2. Table: SKU Details

Table 1 shows that some orders contain more than one product, and they may not be from the same SKU. Consequently, the *SKU Details* table (Table 2) lists the SKU details for each order, separating the order if more than one SKU is ordered, and each SKU is labeled by an *order_item*. Each entry describes the details of a single SKU in an order (*order_no*) and uniquely identified by *rrs_order_detail_id*.

Field	Description	Data type	Sample value
<i>rrs_order_detail_id</i>	RRS database identification	bigint	4a5c8645a0c22ea3a4bd43622a6a6e66
<i>order_no</i>	Order unique identifier	varchar	0d0a09d33b1190313a392d619e9d223a
<i>order_item</i>	SKU sequential number in the order	varchar	1
<i>mat_code</i>	SKU unique identifier	varchar	6e0b602b
<i>order_amt</i>	Product quantity of the SKU	int	1
<i>location</i>	Storage location	varchar	LG001
<i>mat_length</i>	Length of the SKU (<i>cm</i>)	decimal	93
<i>mat_width</i>	Width of the SKU (<i>cm</i>)	decimal	39
<i>mat_height</i>	Height of the SKU (<i>cm</i>)	decimal	29
<i>mat_volume</i>	Volume of the SKU (<i>cm³</i>)	decimal	104231
<i>mat_weight</i>	Weight of the SKU (<i>kg</i>)	decimal	12

Table 2 Description of the SKU details table

To maintain the uniqueness of each SKU, a unique identifier, the *mat_code*, is assigned by RRS. The amount of the required SKU, and the corresponding storage position in its origin center is also given as *order_amt* and *location*. The length, width, height, volume, and weight of the item for each SKU are labeled as *mat_length*, *mat_width*, *mat_height*, *mat_volume*, and *mat_weight*, respectively.

Since the clients of RRS are not necessarily online retailers, we cannot analyze the retail performance of specific retailers. Fortunately, RRS assigns a unique *mat_code* for each SKU in the dataset, which implies that even the same product sold by different clients will have a different *mat_code*. Consequently, each SKU is sold by only one retailer. Therefore, we can analyze the retail performance of the SKUs rather than retailers. From the statistical result shown in Figure 7, for about 63% of all the 18,263 SKUs, fewer than 100 units were sold during the one-year period. Only 665 SKUs had a sales volume of more than 10 thousand, among which 90 had volume more than 50 thousand units, and 28 had volume higher than 100 thousand. The cumulative percentage of sales is shown in Figure 8, with the SKUs ranked according to sales volume.

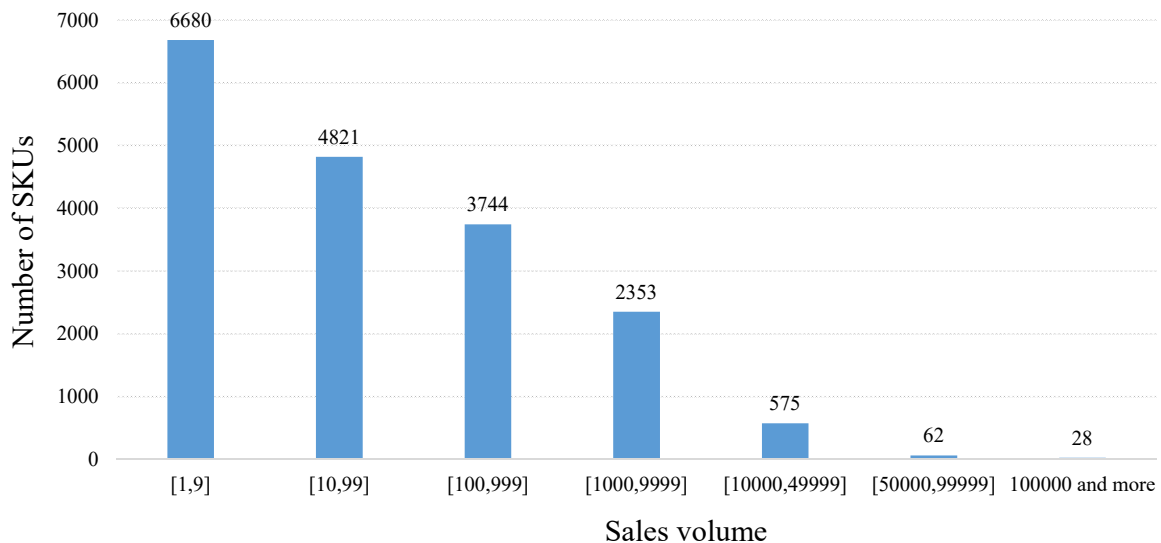


Figure 7 Number of SKUs for different sales volumes

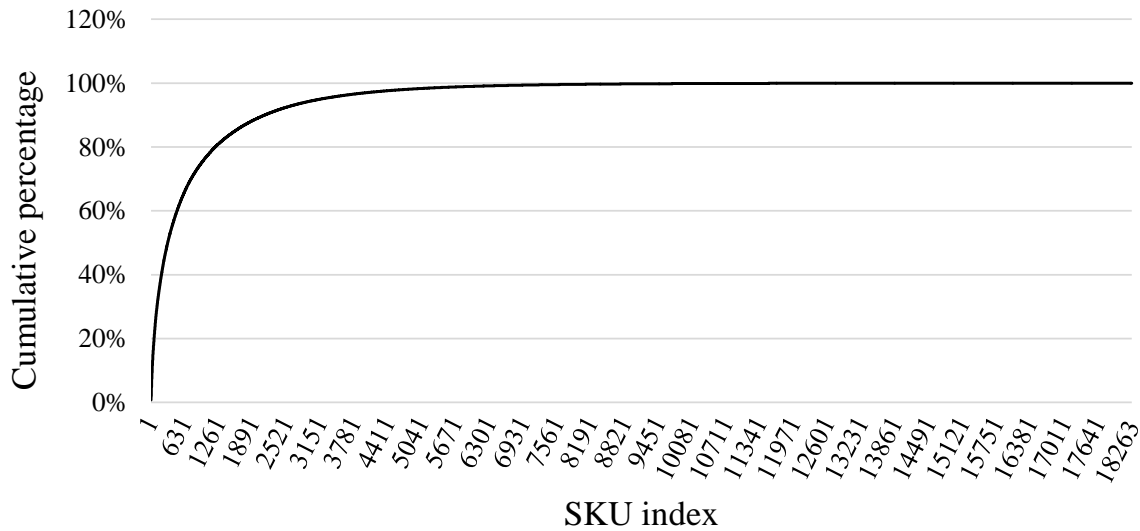


Figure 8 Cumulative percentage of SKU sales

2.3. Table: Appointment Details

The *Appointment Details* table (Table 3) records detailed appointment information for each order. As we mentioned earlier, the exact number of orders in each specific table may vary due to the effect of time setting and the enormous amount of data. This table contains detailed appointment information for over 19 million orders.

Field	Description	Data type	Sample value
rrs_order_extend_id	RRS database identification	bigint	81546ef5b957b810f9389ba42255d3b5
order_no	Order unique identifier	varchar	0d0a09d33b1190313a392d619e9d223a
client_install_date	Installation time requested by the client	datetime	2019-06-04 23:59:59
client_require_date	Delivery time requested by the client	datetime	2019-06-04 23:59:59
oms_aging_name	Effective time zone	int	72
oms_aging_user_date	Estimated latest delivery time according to its effective time zone	timestamp	2019-06-04 23:59:59
appointment_date	Delivery time appointed with consumer	timestamp	2019-06-04 23:59:59

Table 3 Description of the appointment details table

Each entry, uniquely identified by *rrs_order_extend_id*, corresponds to an order (*order_no*). The delivery time and installation time requested by the client, labeled by *client_install_date* and *client_require_date*, are collected when the order is imported into

RRS's logistics system. (Note that these two requests are not placed by the end-user consumer but rather by one of the 149 clients of RRS). Each order falls into an area subject to a certain promised delivery time, given as *oms_aging_name*, is defined as the maximum time required by RRS to deliver the product to the consumer. The area is measured by hours, and the value of *oms_aging_name* can be 12, 24, 48, 72, 96.... Correspondingly, the estimated latest delivery time can be further obtained and is labeled as *oms_aging_user_date*. Finally, *appointment_date* is the appointed delivery time RRS's client has set with the consumer. In most circumstances, the client does not provide the delivery and installation requirements, in which case the *client_install_date* and *client_require_date* will be filled in by the system according to the estimated delivery time (*oms_aging_user_date*).

When we look at the delivery performance of RRS, delivery speed is an important criterion. As shown in Figure 9, 99.30% of all the orders were to be delivered within one week, and about 48% of those orders were scheduled to be delivered within 24 hours.

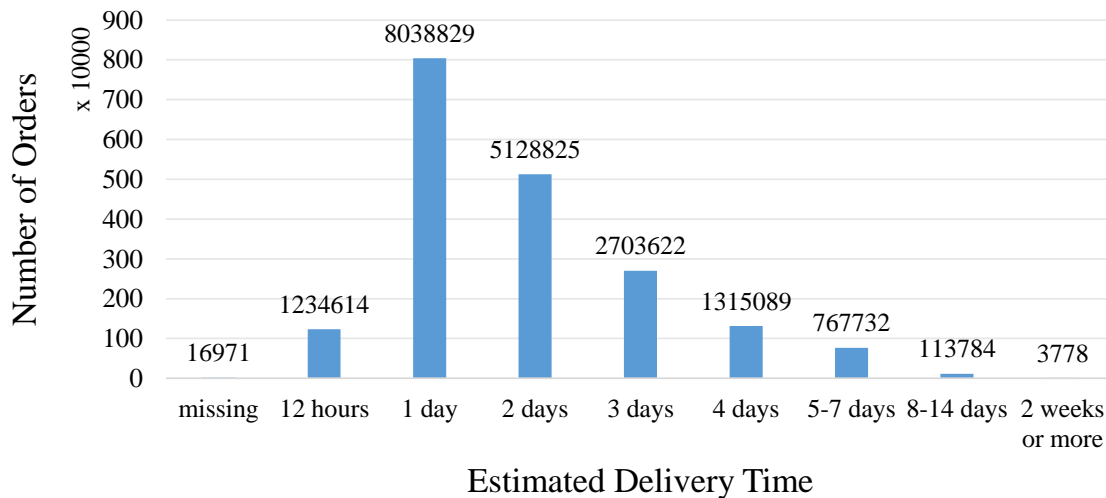


Figure 9 Distribution of orders with different estimated delivery time

2.4. Table: Delivery Details

The *Delivery Details* table (Table 4) provides the detailed delivery information of each order (*order_no*), containing all operation nodes in the entire logistics distribution process (Figure 3). The unique RRS identification code of each operation node is labeled as *rrs_pool_node_info_id* in this table⁶. The location at the distribution center level (CDC,

⁶ Detailed descriptions of the operation node codes are given in Table A.1.

RDC, or LTC) where the node action takes place is labeled as the *operation_center_code*; this location could be the *Origin Center*, *Transfer Center*, or *Destination Center* of the order. Because the operation nodes indicate the product’s arrival or departure, the origin (which could be the *Origin Center*, *Transfer Center*, *Destination Center*, or the last-mile hub of the order) and destination (which could be the *Transfer Center*, *Destination Center*, the last-mile hub, or detailed GB code) for the particular operation are labeled by *orig_code* and *dest_code*, respectively (*NULL* for other operation nodes). More specifically, each operation node starts or ends a part of the delivery process, so the *orig_code* is the start location of this part, and the *dest_code* is the part’s destination. Take “*QS*” (signed) as an example; it ends the delivery from the last-mile hub to the consumer. Therefore, the *orig_code* and *dest_code* are the last-mile hub and the location of the consumer (i.e., GB code in the dataset), respectively. The time when the operation occurs is recorded as *node_operation_date*.

Field	Description	Data type	Sample value
rrs_pool_node_info_id	RRS database identification	bigint	01a967a7bda6a071e7b4f71275e102aa
order_no	Order unique identifier	varchar	0d0a09d33b1190313a392d619e9d223a
operation_center_code	Operation distribution center	varchar	RRSZX076
orig_code	Origin code of each node	varchar	rrs_wd_3927
dest_code	Destination code of each node	varchar	GB00264
node_code	Code of the operation node	varchar	QS
node_operation_date	Time when the node occurs	datetime	2019-06-04 18:32:13

Table 4 Description of the delivery detail table

To deliver a better understanding of the dataset, we provide detailed statistical results. The amount of data in this table is quite large (about 18 GB), so we choose the part of the table containing information for orders issued in the first half of June 2019 to perform a sample analysis. A total of 1,463,018 orders are involved in this sample, and 1,346,188 of them have a whole delivery cycle (with the starting node ‘*XF*’ and the finishing node ‘*QS*’). Figure 10 shows the distribution of these orders with respect to the time spent completing the order. As we can see, 98% of the orders were finished within one week, and more than half of the orders were delivered successfully in two days.

When we recall the estimated delivery time given in Table 3, we can compare the actual time spent and its estimation in Figure 11⁷. The table shows that more than 72% orders

⁷ One order was excluded because its promised delivery time (*oms_aging_name*) is missing.

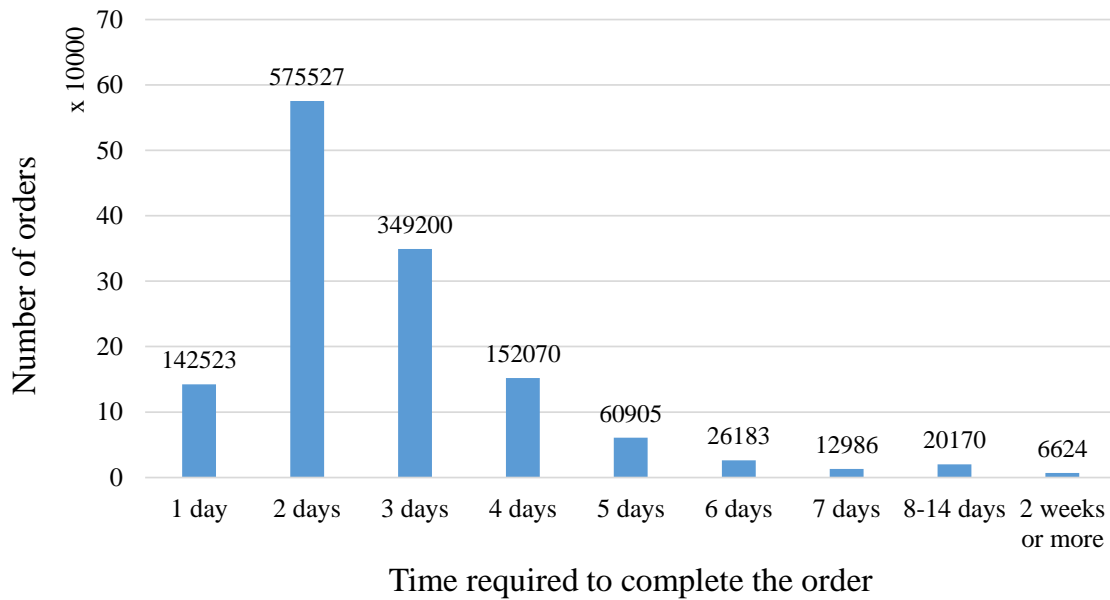


Figure 10 Distribution of orders with respect to time spent to complete the order

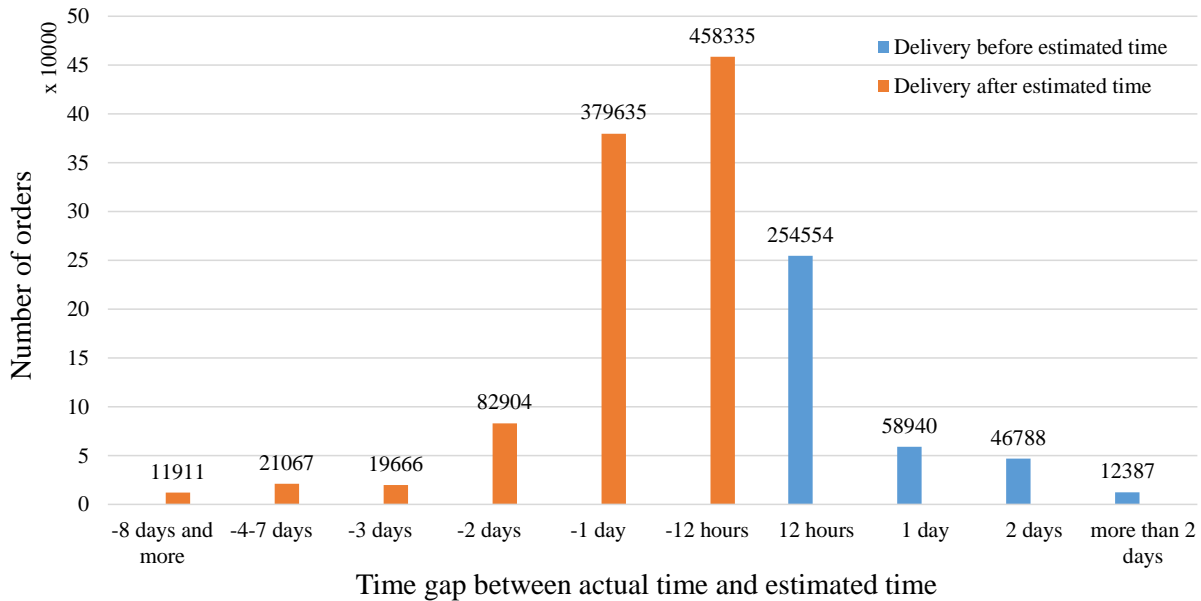


Figure 11 Delivery performance compared with estimated delivery time

were delayed, taking longer than the estimated delivery time, although only about 10% were delayed by more than one day. In practice, consumers are more sensitive to delays when they are given a specific appointment time rather than a general estimated time of delivery. Among the orders in this sample, only 842,397 have an appointed delivery time confirmed

with the consumers. By comparing the actual delivery time and the appointment, as shown in Figure 12, only 4,276 orders (0.52%) were delayed enough to miss the appointment. This result implies that the company performs at a very high level.

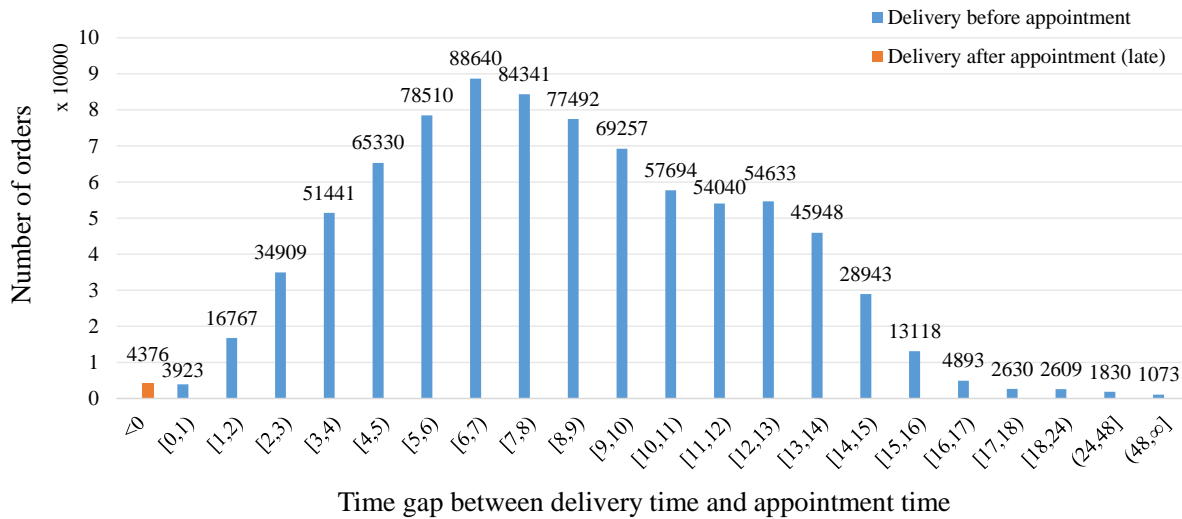


Figure 12 Delivery performance with respect to appointments

Finally, the *Delivery Details* table also provides the installation information of orders. Among all the home appliance products, most of them (more than 80%) do not require installation service. These products, such as refrigerators, washing machines, and other small products, can be manually installed by consumers themselves. Therefore, only 123,010 orders (about 10% in the selected part of the table) have installation information in the dataset. In particular, 16,165 (13% of 123,010) of them are installed by RRS (with operation node “WA”, which means “Installation completed by RRS”) because RRS only provides installation service for products produced by *Haier*. Other products which require installation but are not produced by *Haier* have operation node “WB” (meaning “Not installed”). It is of interest to investigate the behavior of the installers and consumers based on the time used to complete the installation (the time gap between operation nodes “QS” and “WA”). Unfortunately, among the 16,165 orders installed by RRS, 15,934 were completed within five minutes after “QS”. This results from the deliverymen also being the installers of the product, and they send the signing information together with the installation information to reduce the number of operations. Nevertheless, research on finding more driving forces of this behavior and how to incentivize the workers to upload the information promptly are also of interest to company managers.

2.5. Table: Consumer Details

The *consumer details* table (Table 5) provides detailed information about the consumers (i.e., the recipients) of each order (*order_no*), where *rrs_order_person_info_id* is the RRS identification code used in the table. As mentioned before, the number of orders contained in this table is more than 19 million due to the effect of time setting and the enormous amount of data. This table covers 4,223,372 consumers. Each consumer’s name and detailed address information is provided with anonymization: *person_name*, *person_province*, *person_city*, *person_area*, and *person_town*. The three-grade GBcode (China’s standard unique identification codes for counties) and four-grade GBcode (China’s standard unique identification codes for towns) are also given as *person_three_gbcode* and *person_four_gbcode*. Although the postal code of each person may also be provided (*person_post_code*), most of these are missing.

Field	Description	Data type	Sample value
<i>rrs_order_person_info_id</i>	RRS database identification	bigint	0bf5b524f47a23a20441f63e42d87114
<i>order_no</i>	Order unique identifier	varchar	0d0a09d33b1190313a392d619e9d223a
<i>person_name</i>	Person name	varchar	eafb22a5ebef6d950bbc7f6cd9f1bfd4
<i>person_province</i>	Province	varchar	72537f5d52e1a9edd15a41654887e3f5
<i>person_city</i>	City	varchar	fbf1a9249b9ffec15347f4783a643a9
<i>person_area</i>	County	varchar	183170b80bf4f5d175ff5b3047096b08
<i>person_town</i>	Town	varchar	ef30debda3196a13dc4b28b97250de56
<i>person_three_gbcode</i>	Third GB code	varchar	GB00264
<i>person_four_gbcode</i>	Fourth GB code	varchar	GB07194
<i>person_post_code</i>	Post code	varchar	123456

Table 5 Description of the consumer detail table

One consumer may purchase more than once during the time window covered by the data. As shown in Figure 13, more than 40% of consumers have more than two orders, and more than 10% of consumers have six or more orders. Intuitively, only rarely would one person place orders for more than ten home appliances in a one-year time window. However, as we can see from Figure 13, 206,263 consumers (4.88%) have 11 and more orders, and 39,651 of them have more than 50 orders during the year. Investigating the purchasing behavior of these consumers will be interesting and helpful for home appliance manufacturers.

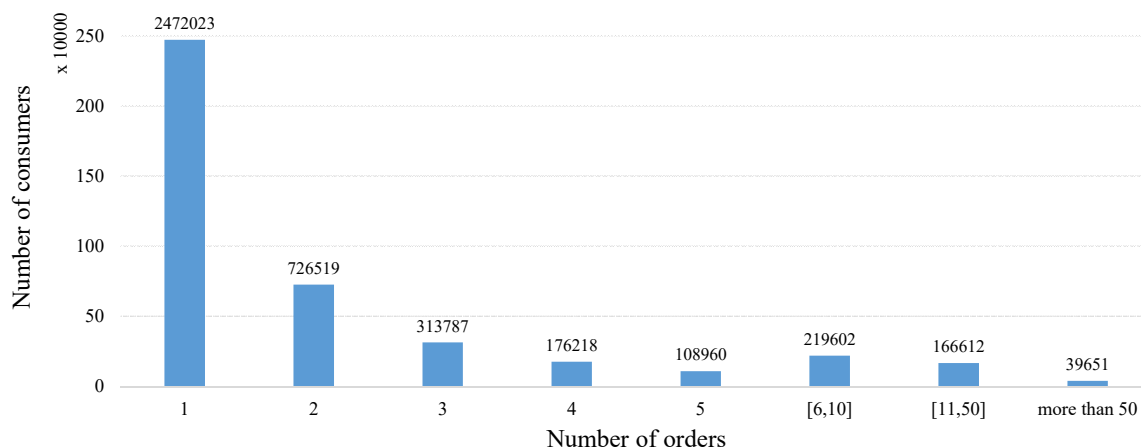


Figure 13 Distribution of consumers with respect to order frequency

2.6. Tables: Client Details and Distance Information

In this subsection, we provide two tables containing important information which will help researchers use the dataset better. The first one, the *Client Details* table (Table 6), provides the type information of the 149 clients, i.e., logistics operators (such as *Cainiao*), retailers, or manufacturers. Therein, 6 clients are logistics operators ($client_type=1$), 72 clients are retailers ($client_type=2$), and the remaining 71 are manufacturers ($client_type=3$).

Field	Description	Data type	Sample value
client_code	Client unique identifier	varchar	RRS18
client_type	The type of the client	int	1

Table 6 Description of the client detail table

The other table (*Distance Information*) provides a travel distance matrix covering all the 103 centers. Travel distance between any two centers in the network is estimated by navigation data based on the geo-location of each center, and we believe this information will give a more visualized picture of the network structure of RRS. It should be noted that the data may have slight differences from the travel distance information for specific orders provided in the *Orders* table (Table 1). This is because there are usually multiple routing suggestions between any two centers, and we choose the first option when carrying out the navigation estimation.

3. Conclusion and Suggested Research Directions

After the success of the last two *M&SOM Data-Driven Research Competitions* supported by *Cainiao Network*⁸ and *JD.com*⁹, *RiRiShun Logistics* decided to offer a detailed logistics dataset to support the *2021 M&SOM Data-Driven Research Competition*. This article provides a detailed description of this dataset. Covering the entire delivery cycle, delivery details are given for over 14 million orders from 149 RRS clients with more than 18 thousand SKUs that passed through 103 distribution centers.

Details for the involved 4.2 million consumers are also offered. A limitation of the dataset is that detailed information about the retailers involved in these orders is missing, which was unavoidable since the RRS clients are normally online retail platforms that provide an online marketplace and are responsible for delivery on behalf of their retailers. Fortunately, a unique identification code is issued for each SKU of each client, which allows analysis of the retail performance for a specific SKU, though not a specific retailer.

Given the dataset, potential research questions about econometric models, data-driven optimization techniques, analytical models, and intelligent algorithms are all welcomed, regarding either the suggested research directions or research questions arising from scholars' other interests. Specifically, studies on the following problems are highly recommended by RRS managers:

(1) On the strategic level, how should the cooperation contracts for RRS's clients (i.e., consigners) be designed in view of different delivery requirements? Given the order details, the demand for a client can be forecast, so how should this information be utilized in optimizing warehouse design and negotiating service contracts?

(2) With cloud storage and cloud distribution technology, are all transshipments among individual distribution centers necessary? If yes, what is the optimal transshipment policy for each individual center? Further, what is the best way to manage each center's inventory, considering the effect of transshipment to retailers?

(3) In the current decision-making system of the whole ecosystem (i.e., the RRS company), all logistical choices, including shipping routes and which center to collect the product, are centralized decisions by the RRS headquarters to minimize cost. Then, from the perspective of optimizing the whole ecosystem, how should the logistics cost and revenue be allocated for each center involved in a specific order?

⁸ <https://connect.informs.org/msom/events/datadriven2018>

⁹ <https://connect.informs.org/msom/events/datadriven2020>

(4) Given the considerable cost of reassignments and multi-distribution of home appliances, how can effective incentives be designed to increase the delivery success rate?

(5) Cross-dock warehousing optimization: The transportation modes and vehicles for trunk line transportation, urban logistics, and last-mile delivery are different. Optimizing the cross-dock warehousing operations between different transportation modes is of interest, including but not limited to warehouse design, storage policies, and standard pallet design.

(6) What is the optimal way to design the distribution network and determine the service area of each distribution center?

(7) Since e-commerce eliminates the geographical constraint for e-retailers, the end-user consumers may be located across the country. Is it beneficial for these clients (consigners or the e-retailers) to keep a pooled inventory at a variety of distribution centers, considering the effects of transportation cost and lost sales?

4. Sample Data and Python Code

To provide better understanding, a Python notebook with runnable sample code and a sample dataset with 500 detailed orders are provided to facilitate reviewing and understanding of the dataset. The detailed Python code and a preview are provided in the Online Appendix as supplemental files.

Acknowledgments

The authors sincerely thank Editor Christopher S. Tang, the associate editor, and three anonymous reviewers for their valuable and helpful suggestions.

Funding

This work was supported by the National Natural Science Foundation of China [Grants 72091215/72091210, 71871207, 71991464/71991460, and 71921001].

Appendix A: Description of the Detailed Operation Notes

Node Code	Description	Node Code	Description
CK	Product has left the origin center	TE	Vehicles on the way
CM	Door dismantlement is required for delivering to house	TL	Transferred to a third-party express company
DH	Product lost during delivery	TM	Bar code
ER	Abnormal operation	TW	Special service required (e.g., taken upstairs by manual labor (elevator unavailable), lifting)
JS	The consumer refused to sign for the order	UX	Transferred to <i>SF Express</i>
K0	Print delivery information	W2	Product has left the hub to the consumer
LF	Intercept failed	WA	Installation completed by RRS
LJ	Intercepted successfully	WB	Product not Installed
PC	Vehicle assigned for the order	WJ	Order status changed from signed to rejected
PR	Delivered from destination center to consumer directly	XF	Order issued and delivered to the origin center
QS	Signed: consumer received the product successfully	XFK	Payment request delivered
QY	Reminder for delivery according to the estimated time	XSK	Collection request delivered
RQ	Arrived at hub	Y8	Delivery appointment is made with consumer
SC	Product has left the destination center	YA	Order delayed
SM	System scan	YE	Appointment modified
SR	Received by the destination center	ZC	Product has left the transfer center
SX	Order issued from the destination center	ZR	Received by the transfer center

Table A.1 Description of the Detailed Operation Nodes

Appendix B: List of national central distribution centers (CDCs), regional distribution centers (RDCs), and local transfer centers (LTCs)

Center Type	Center codes
National Central Distribution Centers (CDCs)	RRSZX031, RRSZX042, RRSZX056, RRSZX072, RRSZX081, RRSZX093, RRSZX104
Regional Distribution Centers (RDCs)	RRSZX001, RRSZX002, RRSZX003, RRSZX004, RRSZX020, RRSZX022, RRSZX023, RRSZX033, RRSZX038, RRSZX039, RRSZX040, RRSZX041, RRSZX043, RRSZX044, RRSZX045, RRSZX058, RRSZX059, RRSZX060, RRSZX073, RRSZX074, RRSZX082, RRSZX083, RRSZX084, RRSZX094, RRSZX095, RRSZX110
Local Transfer Centers (LTCs)	All the other 70 centers not listed as CDCs or RDCs. Therein, “RRSZX037” is closed in May 2019.

Table A.2 List of CDCs, RDCs, and LTCs