Mining Customer-related Data to Enhance Home Delivery in E-commerce: an experimental study

Shenle Pan¹, Yufei Han¹, Bin Qiao¹, Etta Grover-Silva², Vaggelis Giannikas³

¹ MINES ParisTech, PSL Research University, CGS - Centre de gestion scientifique, Paris, France
² MINES ParisTech, PSL Research University, PERSEE - Centre Procédés, Energies Renouvelables et Systèmes Énergétiques, Sophia Antipolis, France
³ Institute for Manufacturing, University of Cambridge, Cambridge, UK

{shenle.pan@mines-paristech.fr, yfhan.hust@gmail.com, bin.qiao@mines-paristech.fr, etta.grover-silva@mines-paristech.fr, v.giannikas@eng.cam.ac.uk}

Abstract. In a B2C e-commerce environment, home delivery service refers to delivering goods from an e-retailer’s storage point to a customer’s home. High rate of failed delivery due to the customer’s absence causes significant loss of logistics efficiency. This paper aims to study innovative solutions to the problem, such as data-related techniques. This paper proposes a methodological approach to use customer-related data to optimize home delivery. The idea is to estimate the attendance probability of a customer via mining his electricity consumption data, in order to improve the success rate of delivery and optimize transportation. Computational experiments reveal that the proposed approach could reduce the total distance from 3% to 20%, and theoretically increase the success rate around 18%-26%. Being an experimental study, this paper demonstrates the effectiveness of data-related techniques or data-based solutions in home delivery problem, and provides a methodological approach to this line of research.

Keywords: City Logistics, Home delivery, Data Mining, Electricity Consumption Data, Capacitated Vehicle Routing Problem with Time Windows.

1 Introduction

In a B2C (business-to-consumer) e-commerce environment, home delivery service involves delivering online purchased goods from the e-retailer’s storage point to the customer’s home [1]. Due to its convenience to customers, in the last ten years home delivery has become a dominant distribution channel of e-commerce, especially for products of high value or high volume like furniture or domestic appliances, or for perishable products like grocery [2]. The particular case requiring the attendance of customer (or receiver in general) at the moment of delivery gives rise to a problem which is known in the literature as Attended Home Delivery Problem (AHDP), see [3]. However, as the attendance of customer is somehow unpredictable in nowadays delivery planning, home delivery usually results in high rate of failures [4]; and, therefore, a number of problems has been raised, e.g., high delivery costs, waste of waiting time (for the customer) and waste of energy for transport etc. Effectively and efficiently tackling AHDP is becoming a key success factor to e-retailers, as well as an important challenge to City Logistics with regard to sustainability [5, 6].

Considering its importance, both practical and academic sides have already studied the AHDP in depth. For example, at the level of reception, some practical solutions are widely employed when a customer is absent, such as leaving the package to someone nearby (neighbors, gatekeeper etc.) or somewhere with security such as mailbox (yard, garage etc.) [1]. In practice, carrier may also plan a second delivery or leave the package to the closest pickup points (e.g., post office, shops, self-service pickup station). These solutions are very practical but their disadvantages are also obvious, e.g., security issues, increased delivery costs and purchaser’s unnecessary moving. At the level of transportation, the AHDP has mainly been addressed as a Vehicle Routing Problem (VRP) - more precisely VRP with time windows (VRPTW), or as a time slot management problem [4]. Reliability and width of the time windows strongly impact the results of routing optimization.

This paper aims to study innovative solutions for AHDP, for example based on data-related techniques. We try to investigate the way to use some customer-related data to improve home delivery efficiency.
an example, this paper experiments on purchaser’s home electricity consumption data. A two-stage approach is proposed. The first stage concerns a data mining process, whose objective is to estimate the purchaser’s attendance probability at a given time windows according to his electricity consumption behaviour. An open dataset from the Irish Commission for Energy Regulation (CER) is employed in this stage with the help of data mining techniques. The outcome, which is a set of time windows with attendance probabilities for each purchaser, serves as input data to the second step where an optimization model for the VRPTW is employed. The model takes into account the purchaser attendance probabilities and aims to establish weekly transportation plans that maximize delivery success rates while minimizing total transportation distances. Being an experimental study, the main contribution of this paper is to demonstrate the effectiveness of data-related techniques or data-based solutions in AHDP, and to provide a methodological approach to this line of research.

Following this introduction, Section 2 provides a review of related work. Then, Sections 3 and 4 present the two-stage approach, that is, respectively, data mining stage and the transport planning stage. Section 5 presents a numerical study aiming to demonstrate the practicability and performance of the proposed approach. Finally Section 6 concludes this work

2 Related Work

2.1 VRP in Last Mile and Attended Home Delivery

In the context of last mile delivery, the capacitated vehicle routing problem (CVRP) involves the planning of a set of routes for a capacitated vehicle that start from and end with a (retailer’s) warehouse, aiming at delivering the demanded goods to customers’ homes [7]. Particularly when an attended home delivery is planned, the customer is required to be present at the moment of delivery. This case is usually addressed as CVRPTW [7], in which time windows for receiver is introduced [5]. For example [8] study a single-vehicle routing problem with time windows for perishable goods, in which the vehicle may deliver multiple points with multiple routes during its workday. When e-retailers allow purchasers to select the time of delivery at the time of order, as the e-grocers cases studied in [3, 4], the problem is addressed as time slots management problem. The objective is to determine the set of time slots to offer to customers for their selection so as to minimize expected delivery costs while adequately meeting the service requirements [2, 4]. The solution can certainly improve the expected success rate of delivery, yet it also increases the difficulty to transport planning since the time windows proposed are often narrow and overlapped. Contrarily, when the time windows are too wide, customers have to stay at home for hours.

This paper follows the stream of CVRPTW applied in AHDP. Nevertheless, we do not attempt to offer a set of time slots of delivery to purchasers, but to provide an approach to determine the optimal time slot profiles to retailers, in order to optimize the success rate of delivery. The transportation cost and service level to customers can be thus improved. To this end, we experiment data-related techniques and data-based solutions, e.g., analyzing customer’s electricity consumption behavior.

2.2 Data mining in Electricity Consumption Analysis

Building statistical correlation between energy consumption traces and social-economical factors of customers has received a lot of attention lately. The major purpose of the research topic is to find out the underlying patterns of customers’ daily living habits based on their energy usage behaviors, in order to provide a strong base to not only the energy supplying service, but also to other value added services, such as on-line shopping recommendation and targeted advertisement. Generally, the previous research efforts in this domain have two categories.

End-user models are commonly used as an alternative to black-box methods [9-13]. This approach requires information about housing conditions, electrical appliance usage and environmental factors. Such background information is used with energy domain knowledge to disaggregate the daily electricity consumption measurements of a specific user into elementary components, including heating/cooling, water usage, cooking and other behaviors. The disaggregation result is then applied to find out their usage
preferences. The shortcoming of end-user models is that forecasting performance depends heavily on the quality of available information, which makes them sensitive to noise and unable to performed automatically. Econometric methods [10, 12, 13] estimate the relationship between energy consumption profiles and the factors influencing consumption behavior using statistical machine learning approaches, such as support vector regressors, decision trees and so on. Econometric models are built by learning the mapping from pairs of the factors and energy consumption profiles automatically, which is appealing for realistic application deployments. Recently, this category has gained popularity. Most research efforts along this direction focus on estimating the users’ general social economical factors, such as professions, family status, salary levels and so on. They are preferable information for improving quality of on-line purchasing service. To the best of our knowledge, using consumption measurements to estimate residences’ occupancy has remained elusive in the literature.

This paper aims at combining the two research streams - data mining in electricity consumption analysis and VRP - to propose an innovative approach to address the attended home delivery problem. The approach is completely new in the literature, and it can be seen as a contribution to the research stream of data-based solutions to freight transport problem in city logistics [14, 15]. In the next part we present the data mining techniques used to determine customer’s attendance probability.

3 Methodology of Mining Electricity Consumption Data

3.1 Data Collection

In our work, we use a publicly available electricity consumption trace data set, named CER ISSDA to simulate and demonstrate the use of estimating occupancy patterns of private households for optimizing shipping plan. The CER ISSDA dataset is collected by the Irish Commission for Energy Regulation (CER) in a smart meter study: it contains electricity consumption data of 4,225 private households and 485 small and medium enterprises (all called consumer hereafter); the trace covers 1.5 years (from July 2009 to December 2010). For each consumer, the daily load curve is sampled every 30 minutes: energy data can be thought of as a series of timestamps and energy readings. In addition to energy data, the dataset includes a series of survey sheets and answers for each consumer, describing their housing condition, occupancy, employment status, income level, social class, appliance usage information and other socio-economic factors. Forty-one survey questions belonging to five categories were selected, that allowed to build consumer profiles based on heating and lighting behavior, hot water and other electrical appliances usage. As a sample for this work, we select 20 private households randomly from the dataset to simulate the shipping network in practical applications. To avoid the impacts of seasonal variation on the consumption profiles, e.g., air conditioning use in summer and heating use in winter changing daily consumption patterns, we only focus on the time period from April 1th to June 30th in 2009 to design the experiments, which in total involves 91 days.

3.2 Mining Model

The detection of occupancy within a residential household is primary based on activity detection. Activity within a household is often linked to electric consumption. Therefore, the variance and magnitude of a residential consumption profile can give indications about whether a household is occupied or not. Therefore peaks and high variation of the consumption profile, these typical active behaviors were labeled in our model and then compared to low variation and low overall consumption periods. The combination of low overall consumption and low consumption profile variation indicates a period of inoccupation. Automated devices within a household can interfere with occupancy detection. High overall consumption due to heating devices left on can also interfere with occupancy detection. Therefore it is necessary to identify patterns for each specific client. For each client, a typical profile variation for occupied and unoccupied periods was established in our model. A threshold of average consumption was also determined. When overall consumption was below the threshold and variation of the profile was relatively low, the period was determined to be unoccupied.

1 http://www.ucd.ie/issda/data/commissionforenergyregulationcer/
Based on the above qualitative analysis, we propose a computational model to estimate occupied periods for a given user, as described in the followings. Hereafter, we use $X_{ij}$ to denote electricity consumption level at the $j$-th time step of $i$-th day for one specific user. Given the context of shipping programming, we especially focus on estimating occupancy period within the time interval ranging from 8:00 am to 8:00 pm. Our estimation approach is defined into 4 successive steps:

1. Indicator estimation: calculating the consumption magnitudes $X_{ij}$ and the absolute values of the first-order difference $D_{ij}^1$ and $D_{ij}^2$, as defined in Eq (1). The three measurements form an indicator cell $(X_{ij}, D_{ij}^1, D_{ij}^2)$, which is used as the feature to detect occupancy patterns. Given all 91 days in the database, we have in total 91 (days) * 12 (delivery hours per day) * 2 (timestamps per hour) indicator cells for each user.

$$D_{ij}^1 = |X_{ij} - X_{ij,1}| \text{ and } D_{ij}^2 = |X_{ij} - X_{ij,2}| \quad (1)$$

2. Label the $L_p$ indicator cells for each user with the most typical active behaviors manually by human energetic experts. Treating occupancy detection as anomaly detection problem, the labeled $L_p$ indicator cells form a reference set to identify whether the residences’ activities are present within a given time period. Based on the labeled indicator cells, we build a linear kernel one-class SVM based detector $F$ [16]. The output $Y_{ij}$ of the SVM detector is 1 or 0, deciding whether the human activity is present at the given time step or not (1 corresponds to the presence of the residence and vice versa).

$$Y_{ij} = F(X_{ij}, D_{ij}^1, D_{ij}^2) \quad (2)$$

3. We apply the built SVM based detector on the 24 time steps between 8:00 am to 8:00 pm for each day besides the labeled $L_p$ time steps. The binary output of the detector is used as an estimated occupancy label of the concerned time steps. As such, for the $i$-th day, we can obtain a 24-dimensional binary valued vector as the estimated occupancy label vector of the concerned day.

4. We accumulate estimated occupancy label vectors derived from the total 91 days. For each day of week (from Monday to Sunday), we calculate the empirical expectation of the occupancy label for each time step as the occupancy probability of the time step $j$ for each day of week. As a result, we generate a probability map $M_{ER}^{24x7}$. Each entry $M_{ij}$ represents the estimated occupancy probability of the time step $(j+16)$ on the given day of week. We denote Monday to Sunday as 0 to 6.

According to the algorithmic description, we define the occupancy detection procedure as an one-class anomaly detection problem in our work. Compared with varied inactive consumption behaviors, typical active behavior patterns indicating residences’ occupancy is easier to identify manually for human experts. Therefore, the expected occupancy detector is built to describe the common characteristics of the active consumption profiles and expected to differentiate the active profiles from inactive ones at the same time. This is a typical one-class learning problem in machine learning theory [16] and one-class SVM perfectly matches with this goal. In previous research, it is popularly used to detect a concerned events from backgrounds given only limited number of samples belong to the event are available. The indicator cell designed in this work covers both the instant measurement of the electricity consumption level and the dynamical information about the consumption variation, represented by the two first-order differences. The feature design is based on the theoretical analysis of active consumption profiles. High instant consumption level and high variance of consumption measurement are strong indicators of active behaviors, thus the chosen three features are able to sensitively indicate the potential human activities existing behind the consumption measurements. Finally, the binary output of the occupancy detector can be noisy due to the hard threshold intrinsic. Furthermore, to insert occupancy estimate as a constraint into the shipping programming problem, we need to smooth the binary decision into a soft, continuously valued confidence of residences’ occupancy for each specific time step. As a result, we use empirical expectation as the estimate of the underlying true occupancy probability.

### 3.3 Mining Results

We finally obtained the results of time slots with inoccupation probability by consumer, as the examples shown in Figure 1. For example, overall consumption was below the threshold and variation of the profile was relatively low, the period was determined to be unoccupied. Among the 20 consumers randomly selected, 5 are excluded in this study because the variation of their consumption curve is not significant.
(which are meaningless in this study), e.g., inoccupation probability always > 90%. Moreover, we fix the daily working hour from 8h-17h (common delivery hours in France). The consumer', i.e., customers', attendance probability is then the inverse of the inoccupation probability (knowing that both staying out or inactive at home are considered as inoccupation here). The results show that most consumers have similar inoccupation probability curve during weekdays but Saturday. The figure shows only the average of weekdays’ probability to compare with Saturday (Sunday is not considered for delivery).

Figure 1. Heat map of time slots with inoccupation probability during working hour (Weekdays vs Saturday)

From the figure, we may easily observe that during a day some time windows are relatively more favorable for home delivery and such time windows are different to each consumer. Accordingly we may define a best delivery time window profile for every consumer.

4 VRPTW and Formulation

At the level of transport planning, the problem in this paper is a typical CVRPTW [8, 17, 18], while considering the customers’ attendance probability. We propose a Mixed Integer Linear Programming (MILP) model for the CVRPTW which follows the guidelines of [7] and incorporates the classical constraints to enforce time windows that can be found in [8]. The MILP can be described as follows. Given a set of customers $V = \{1, 2, \ldots, n\}$ with known demands of $q_i$ for any $i \in V$, we have a fleet of homogeneous vehicles of capacity $Q$ to delivery those demands, from a depot noted as 0 to customers. The directed graph can be thus noted as $G = (V+, A)$, where $V+ = V \cup \{0\}$ is the set of nodes and $A$ is the set of arcs. Each arc $(i, j) \in A$ is associated with a travel time $t_{ij} > 0$ and a distance $d_{ij} > 0$. Each customer $i \in V$ is associated with a service time $s_i$ and a time windows $[a_i, b_i]$ that presents respectively the earliest and latest time at which the service must begin at $i$. The objective is to minimize the total distance traveled to serve all customers while satisfying the capacity and time window constraints.

Min \[ \sum_{i \in V} \sum_{j \in V^+} d_{ij} x_{ij} \] \hspace{1cm} (3)

s.t. \[ \sum_{j \in V^+} x_{ij} = 1, \quad i \in V \] \hspace{1cm} (4)

\[ \sum_{i \in V} x_{ia} - \sum_{j \in V^+} x_{ij} = 0, \quad h \in V \] \hspace{1cm} (5)

\[ q_i \leq u_j \leq Q, \quad i \in V \] \hspace{1cm} (6)

\[ u_i - u_j + Q x_{ij} \leq Q - q_j, \quad i \in V, j \in V \] \hspace{1cm} (7)

\[ t_i + (t_j + s_j) x_{y} - N(1-x_y) \leq t_j, \quad i \in V, j \in V \] \hspace{1cm} (8)

\[ a_i \leq t_i \leq b_j, i \in V \] \hspace{1cm} (9)

\[ x_{ij} \in \{0,1\}, \quad i \in V^+, j \in V^+ \] \hspace{1cm} (10)
With decision variables: \( x_{ij} \) is 1 if arc \((i,j)\) is included in any route, 0 otherwise; \( i \) indicates the start time of service on every node \( i \).

In the model, (4)-(5) guarantee that every customer is visited exactly once and that every route begins from the depot and ends at the depot. Eq (6)-(7) ensure that the total demand on every route will not exceed the vehicle capacity \( Q \), in particular \( u_i \) is a variable indicating the accumulative total demand on customer \( i \). Eq (8)-(9) ensure feasibility of the time schedule on every route, with \( N \) a large number.

5 Numerical Study

5.1 Input Data and Assumptions

A numerical study is developed in this part to demonstrate the performance of the proposed model. The 15 consumers presented in Figure 1 are considered as customers in the case and every customer is associated with an inoccupation probability profile. We use a two-dimension \((x,y)\) plan to simulate a city of 18*18 km\(^2\), in which the 15 customers are randomly located, see Figure 2. In particular \((0,0)\) represents e-retailer’s storage point for the city.

<table>
<thead>
<tr>
<th>Customer</th>
<th>x</th>
<th>y</th>
<th>qi</th>
<th>ai</th>
<th>bi</th>
<th>( u_i )</th>
<th>( a_i )</th>
<th>( b_i )</th>
<th>Inoccupation probability</th>
<th>WD / Saturday</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>12307</td>
<td>6596</td>
<td>1</td>
<td>60</td>
<td>120</td>
<td></td>
<td>60</td>
<td>120</td>
<td>26%</td>
<td>23%</td>
</tr>
<tr>
<td>C2</td>
<td>2377</td>
<td>6646</td>
<td>1</td>
<td>300</td>
<td>360</td>
<td></td>
<td>480</td>
<td>540</td>
<td>75%</td>
<td>S</td>
</tr>
<tr>
<td>C3</td>
<td>13009</td>
<td>12331</td>
<td>1</td>
<td>300</td>
<td>360</td>
<td></td>
<td>300</td>
<td>360</td>
<td>59%</td>
<td>S</td>
</tr>
<tr>
<td>C4</td>
<td>1986</td>
<td>10763</td>
<td>1</td>
<td>0</td>
<td>60</td>
<td></td>
<td>0</td>
<td>60</td>
<td>36%</td>
<td>WD</td>
</tr>
<tr>
<td>C5</td>
<td>2115</td>
<td>14209</td>
<td>1</td>
<td>480</td>
<td>540</td>
<td></td>
<td>480</td>
<td>540</td>
<td>52%</td>
<td>S</td>
</tr>
<tr>
<td>C6</td>
<td>11533</td>
<td>6618</td>
<td>1</td>
<td>480</td>
<td>540</td>
<td></td>
<td>300</td>
<td>360</td>
<td>81%</td>
<td>S</td>
</tr>
<tr>
<td>C7</td>
<td>5919</td>
<td>3709</td>
<td>1</td>
<td>480</td>
<td>540</td>
<td></td>
<td>480</td>
<td>540</td>
<td>24%</td>
<td>WD</td>
</tr>
<tr>
<td>C8</td>
<td>11769</td>
<td>1560</td>
<td>1</td>
<td>60</td>
<td>120</td>
<td></td>
<td>60</td>
<td>120</td>
<td>65%</td>
<td>WD</td>
</tr>
<tr>
<td>C9</td>
<td>13484</td>
<td>13895</td>
<td>1</td>
<td>0</td>
<td>60</td>
<td></td>
<td>120</td>
<td>180</td>
<td>55%</td>
<td>S</td>
</tr>
<tr>
<td>C10</td>
<td>10497</td>
<td>3702</td>
<td>1</td>
<td>60</td>
<td>120</td>
<td></td>
<td>120</td>
<td>180</td>
<td>37%</td>
<td>S</td>
</tr>
<tr>
<td>C11</td>
<td>13321</td>
<td>6989</td>
<td>1</td>
<td>0</td>
<td>60</td>
<td></td>
<td>60</td>
<td>120</td>
<td>40%</td>
<td>S</td>
</tr>
<tr>
<td>C12</td>
<td>4227</td>
<td>9932</td>
<td>1</td>
<td>480</td>
<td>540</td>
<td></td>
<td>480</td>
<td>540</td>
<td>11%</td>
<td>WD</td>
</tr>
<tr>
<td>C13</td>
<td>13229</td>
<td>4121</td>
<td>1</td>
<td>180</td>
<td>240</td>
<td></td>
<td>180</td>
<td>240</td>
<td>23%</td>
<td>WD</td>
</tr>
<tr>
<td>C14</td>
<td>17471</td>
<td>11555</td>
<td>1</td>
<td>0</td>
<td>60</td>
<td></td>
<td>240</td>
<td>300</td>
<td>45%</td>
<td>S</td>
</tr>
<tr>
<td>C15</td>
<td>15605</td>
<td>8721</td>
<td>1</td>
<td>240</td>
<td>300</td>
<td></td>
<td>300</td>
<td>360</td>
<td>52%</td>
<td>S</td>
</tr>
</tbody>
</table>

Table 1. Input data of customers

![Figure 2. Geolocation of e-retailer’s storage point and customers](image-url)
The objective of the case is to establish delivery routes that satisfy all delivery requests, while minimize total distance generated by the first-round delivery and, in case of failure, the rescheduled delivery. Several assumptions are made here. (1) Every customer has one delivery request of size 1 at a week and the truck capacity is 5. (2) All customers will accept the proposed delivery time windows. The time windows is set to 1 hour and the associated inoccupation probability is the average of that in Figure 1, e.g., 8h-8h30 and 8h30-9h for 8h-9h. (3) Failed delivery due to the absence of customer will be rescheduled as a direct delivery from the e-retailer’s storage point to the customer in the next day. Accordingly, the distance generated by the rescheduling is equal to round trip distance * inoccupation probability for each customer. In other words, we do not consider another VRPTW for the failed deliveries due to the service constraint and the lack of knowledge on new deliveries. (4) Truck’s speed is set to 20 km/h in city; and (5) Service time per customer is set to 5 minutes, in other words, \(0.3t_{ij} = d_{ij}\) and \(S_i=5\).

5.2 Scenarios and Results

Five scenarios are designed in the study, as shown Table 2. S0 serves as baseline scenario that does not consider time windows in transport plan. In other words, it is a classic VRP optimizing transport distance. S1 and S2 take into account the customers’ optimal time windows profile only in weekdays. S3 and S4 consider both weekdays and Saturday so that customers are divided in two clusters: one for the customers with highest attendance probability appeared on weekdays and another one for the others. All time windows are presented in Table 1. Since the optimal time windows can be dispersive in a day, in S2 and S4 we deliberately added the constraint (12) to limit the waiting time in-between two successive customers to 60 minutes. Constraints (8) and (12) define that, when \(x_{ij}=1\),

\[ t_i + (t_{ij} + s_j)x_{ij} - N(1-x_{ij}) \geq t_j - 60, \quad i \in V, j \in V \]  

(12)

All scenarios were run using GUSEK on ThinkPad T440 with 4 GB of RAM. For all scenarios, every single computation process costs about one minute, except S0 that costs nearly 7 hours. The results are summarized in Table 2 and Table 3. Theoretically S0 provides the shortest routes without considering time windows and failure probability, i.e., the first-round delivery. However, this routing plan caused a low rate of successful delivery of 37% and, as a result, generated the highest total distance resulted from a significant number of rescheduled deliveries. In terms of total distance, S3 considering Saturday delivery performs the best, reducing 20% KM comparing to S0, thanks to a higher rate of successful delivery of 63%. This is because 9 of the 15 customers have lower inoccupation probability on Saturday. As shown by S2 vs S1, or S4 vs S3, the constraint (12) that aims to limit each waiting time made total distance increased. However, without this constraint, the waiting time in-between two successive customers can be longer than 3 hours in S1 and S3 (result of timeline is not provided here due to lake of space). In practice, waiting time depends closely on the quantity of customers to deliver in a tour.

<table>
<thead>
<tr>
<th>Sc.</th>
<th>Time Windows</th>
<th>Saturday Delivery</th>
<th>Waiting Time ≤60 mn</th>
<th>Distance of First Delivery (KM)</th>
<th>Total Distance (KM)</th>
<th>ΔKM vs S0</th>
<th>Average Probability of Successful First Delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>111</td>
<td>367</td>
<td>-</td>
<td>37%</td>
</tr>
<tr>
<td>S1</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>130</td>
<td>320</td>
<td>-13%</td>
<td>55%</td>
</tr>
<tr>
<td>S2</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>166</td>
<td>356</td>
<td>-3%</td>
<td>55%</td>
</tr>
<tr>
<td>S3</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>143</td>
<td>295</td>
<td>-20%</td>
<td>63%</td>
</tr>
<tr>
<td>S4</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>176</td>
<td>327</td>
<td>-11%</td>
<td>63%</td>
</tr>
</tbody>
</table>

Table 3. Optimal route planning of scenarios

<table>
<thead>
<tr>
<th>Routes</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0 R1=(0-10-6-1-13-8-0), R2=(0-7-12-5-4-2-0), R3=(0-11-15-14-9-3-0)</td>
</tr>
<tr>
<td>S1 R1=(0-4-2-5-12-7-0), R2=(0-8-10-1-3-6-0), R3=(0-11-9-14-15-3-0)</td>
</tr>
<tr>
<td>S2 R1=(0-7-0); R2=(0-4-0), R3=(0-6-12-5-0), R4=(0-8-13-15-3-2-0), R5=(0-9-14-11-1-10-0)</td>
</tr>
<tr>
<td>S3 WD: R1=(0-4-8-13-12-7-0); Saturday: R2=(0-1-11-14-15-6-0), R3=(0-10-9-3-5-2-0)</td>
</tr>
<tr>
<td>S4 WD: R1=(0-4-8-13-0), R2=(0-12-7-0); Saturday: R3=(0-5-2-0), R4=(0-11-11-10-0), R5=(0-9-14-3-15-6-0)</td>
</tr>
</tbody>
</table>

6 Conclusion

This paper presents an experimental study to investigate how data-related techniques or data-based solutions can help improving AHDP. A two-stages methodological approach is proposed to experiment
on customer’s historical electricity consumption data. Thanks to data mining techniques the electricity consumption data may help us to estimate inoccupation (absence) probability of each customer. Then we developed a VRPTW model considering time windows with inoccupation probability to improve delivery success rate. A numerical study has been conducted to demonstrate the effectiveness of the proposed approach. The paper is among the first studies integrating data mining techniques in urban freight transport. Some prospects can be thus identified to this line of research. For example, one may test the methodology with some other datasets, e.g., water consumption, Internet debit etc., in order to compare the accuracy and performance. Moreover, this study may also contribute to the slot management problem in home delivery. The proposed approach may help retailers to better understand the optimal delivery time of their customers, then to propose a more adaptive price and slot management mechanism. Further, customers’ satisfaction may also be improved thanks to the proposed approach, e.g., customers may reduce their unnecessary moving to pick up parcels and the risk of product damage, the ship time could also be shorten.

Several limits exist in this work. First relevant legal issues are not discussed in this paper, for example the issues of data privacy or security etc. These kinds of issues should be carefully studied before field-testing and application. Second, the number of customers is limited by computational capacity. The original data set contains more than 4000 consumer profiles, but only 15 of them are included in the study. Knowing that such number of nods is already a challenge to the current VRP models, an adapted heuristic approach is necessary if we want to investigate the approach with customers of large scale.

References