European Journal of Operational Research 287 (2020) 820–831

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Accounting for cost heterogeneity on the demand in the context of a technician dispatching problem

Juan P. Cavada[a], Cristián E. Cortés[a,*], Marcel Goic[b], Andrés Weintraub[b], Juan I. Zambrano[b]

[a] Civil Engineering Department, Universidad de Chile, Blanco Encalada 2002, Santiago, Chile
[b] Industrial Engineering Department, Universidad de Chile, Beauchef 851, Santiago, Chile

ARTICLE INFO

Article history:
Received 30 August 2019
Accepted 28 April 2020
Available online 18 May 2020

Keywords:
Technician Dispatching
Costs Heterogeneity
Markovian Models

ABSTRACT

In the technician dispatching problem, a given number of repair teams must visit different locations to provide service support. Considering that there is a fixed vehicle capacity and variations in the demand, not all requests can be satisfied on time and therefore some of them must be delayed. Most implementations of the dispatching problem consider a penalty that might vary depending on the customers to internalize that they have heterogeneous costs for being postponed. In this research we analyze how such variations in costs affect the outcome of service planning in the context of an efficient technician dispatching problem. We focus our analysis on two objectives: first, to understand how cost heterogeneity affects the performance of optimal solutions, and second to illustrate how a firm could implement an ad-hoc methodology even in cases where only observable customers' features can be traced. Specifically, we explore how the distribution of costs affects optimal solutions of allocating teams during a daily operation of the service provider, and then we propose a Markovian model to capture cost-heterogeneity for the case where the cost of failure can be traced to observable operational characteristics. In this model we explicitly consider the cost faced by the customer by having inferior service quality. Our results indicate that when customers are sufficiently different, transportation and total penalty costs decrease gaining in operational efficiency. Moreover, results from the Markovian model indicate that firms can take advantage of these operational gains even in cases where only few customer characteristics are observed.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

Many important operational problems require the allocation of scarce resources to satisfy a myriad of customer requirements. In many of such problems, the real costs or impact on customers as a consequence of delaying service are not properly incorporated. In the present research, our focus is on the characterization of the heterogeneity in costs for different customers, taken as a case example a technician dispatch problem formulated as a variant of a Vehicle Routing Problem (VRP), and optimized in a previous work for daily operations using an efficient exact column generation approach. In the present paper the objective is to model the heterogenicity in costs across customers, to explore how the differences among them could lead to better options for optimization in technician assignments. This potential room for optimization should be reflected in a more efficient routing implying a reduction in transportation costs due to delaying service to customers with lower costs. We then consider how the company may gain knowledge of customer costs based on data that the firm usually has and how we could characterize solutions that optimize an objective function including simultaneously transportation costs and quality of service based on the assessment of the real costs of each customer. The embedded VRP is solved as efficient as in Cortés, Gendreau, Rousseau, Souyris, & Weintraub (2014), although the central focus of this paper is not the vehicle routing modelling.

Thus, in the context of a VRP, a central planner assigns vehicles and defines routes to fulfill customer demands for services. In the short run, there is usually a fixed capacity of vehicles (with a technician in each vehicle) and therefore, if variability in the demand is observed, some of the requirements cannot be satisfied and must be delayed. Literature on VRP has proposed different approaches to address this shortage of capacity, including penalizations for exceeding capacity (Cortés et al., 2014) and explicitly accounting for the costs of increasing transportation capacity (Fagerholt, 1999). In the context of VRP with time windows (VRPTW) where each costumer must be visited within a predefined time lapse (Baldacci, Bartolini, Mingozzi, & Roberti, 2010), the central planner could

* Corresponding author.
E-mail addresses: jucavada@ing.uchile.cl (J.P. Cavada), ccortes@ing.uchile.cl (C.E. Cortés), mgoic@dii.uchile.cl (M. Goic), aweintra@dii.uchile.cl (A. Weintraub), jzambran@ing.uchile.cl (J.I. Zambrano).

https://doi.org/10.1016/j.ejor.2020.04.056
0377-2217/© 2020 Elsevier B.V. All rights reserved.

consider time windows as soft constraints to avoid unfeasibility under complex configurations. Time windows can be violated if the direct costs of the violation can be compensated by a larger gain in transportation costs (Liberatore, Righini, & Salani, 2011). In these approaches, optimal solutions are obtained by carefully accounting for operational costs of the supplier. This typically includes the transportation costs and the costs for delays if time windows are considered (Kallehauge, Larsen, Madsen, & Solomon, 2005). However, the costs associated with the demand side is either neglected or considered in an ad-hoc manner. For example, the model might include linear constraints indicating that a large fraction of the demand must be satisfied (Amiri & Salari, 2018). Alternatively, the objective function could consider some penalties for not satisfying some of the customers. Moreover, customers are typically considered homogeneous in costs, and therefore the impact in the objective function for not satisfying the demand of one customer is the same regardless of the who the customer is.

To understand the potential impact of cost heterogeneity, consider for example the dispatching problem where technical teams are sent to different locations to provide service support (Hill, 1992; Weigel & Cao, 1999). If customer heterogeneity were not considered, dispatching solutions could leave unserved customers who are very cost sensitive, leading to Pareto-suboptimal solutions. Thus, homogeneous policies lead to inefficient resource allocations and they will damage customer equity in the long run (Vogel, Evanschitzky, & Ramaseshan, 2008). Moreover, the identification of customers with low costs of postponement can provide more flexibility to the problem generating better solutions. Even if the demand for service corresponds to internal customers, in order to achieve global optimality, the costs of the units that demand the service must be balanced with the costs of providing it. Thus, regardless of the nature of the relationship between customers and the service provider, ignoring opportunity costs in the demand side can lead to suboptimal decisions.

In this research, we consider that customers are heterogeneous in their costs due to delays in providing service, and therefore optimal solution should prioritize more sensitive customers. In practice, the penalty of not complying with a requirement does not distribute homogeneously. Hence, we have characterized this value considering this inherent heterogenicity in costs and we compare certain service metrics for a number of scenarios. In our analysis we are interested in two main research objectives. First, to *understand how cost heterogeneity affects the performance of the optimal solutions*. This can give general guidance to operational planners to decide when it is worth to incorporate a detailed description of the demand side in the model. The second objective is to *illustrate how a firm could implement an ad-hoc methodology if the customer costs are not known by the firm, although it can be traced to observable customers' features*.

While the general idea applies to other settings, in this research we focus on the technician dispatching problem, where customers request repair services for their equipment. In this problem, the service provider optimizes the technician routes based on a weighted-sum cost function that considers transportation costs and the customer costs for being rescheduled to the next day. Notice that customer costs can not only be associated with the reduced output in the manufacturing process, but also with longer waiting times to complete operational tasks. Thus, the difference in costs can arise not only because customers can give different use of the machines, but also because customers differ in the degree they can substitute their equipment.

To understand the impact of explicitly including cost heterogeneity in the demand side, we first provide a general characterization of how optimal solutions change depending on the nature of the heterogeneity of customer costs. In this research we analyze to which extend the optimal solution varies as a function of the variability in costs among customers and the relative importance of customer costs with respect to those experienced by the firm. For example, if there is little variation between customer costs, we expect little impact on the optimal solutions. However, as the difference in costs between customers grows, it is possible that more customers remain unattended. This is because the model would prioritize more expensive customer sacrificing efficiency in the design of traveling routes. Characterizing these solutions is important not only to understand the trade-offs between transportation costs to the firm and customer costs due to delays, but also to provide some guidance about the conditions that make more necessary to account for variations of the costs in the demand side.

After characterizing how the optimal solutions vary with different distributions of customer costs, we propose a simple Markovian model to capture cost-heterogeneity in the case where the costs of failure can be traced to observable operational characteristics. This model is well suited for our real-sized problem where machines to be repaired are homogenous in their use and the number of available equipment per customer is observable. In this case, we can derive a closed form expression for the waiting time that we can balance against the working time of the technicians.

Our results indicate that when customers are sufficiently different, the number of requirements served can be reduced. However, considering that unserved customers are precisely those with lower opportunity costs, the general impact is a sizable costs reduction derived from a gain in flexibility that positively impacts the design of optimal routes. In addition, through a Markovian model to characterize customers´costs, our results show the impact to each customer depending on the machines they have available together with their machines' utilization rates.

The remaining of the paper is organized as follows. In Section 2 we review the related literature. In Section 3 we describe our modeling approach to describe the dispatching problem. In Section 4 we describe how we propose to evaluate the impact of heterogeneity in costs of delay. At the end of the same section we present the Markovian model proposed to represent customers' costs. In section 5 we present numerical results and we conclude with a general discussion of the main findings in Section 6.

2. Literature review

This research stands in the intersection of two main streams of literature. To be more specific, we consider the well-known marketing concept of customer heterogeneity, which is applied in the context of the technician dispatching problem where customers might have different costs for not being served on time. Thus, we first describe the literature on customer heterogeneity and its relation to operational services; finally, we discuss how customer costs structures have been approached in the dispatching and other related problems.

Customer heterogeneity has been for decades in the center of marketing plans (Wind, 1978). In essence, when firms recognize that customers have different requirements and preferences, they can design more specialized services that result in more profitable strategies (Stringfellow, Nie, & Bowen, 2004). Marketing literature has also described the benefits of evaluating and acting upon customer profitability in complex supply chain settings (Niraj, Gupta and Narasimhan, 2001). Moreover, it has been shown that prioritization of customer service can pay off (Homburg, Droll & Totzek, 2008). In the context of service planning, the concept of customer heterogeneity translates into the provision of different service levels for customers depending on their requirements. Literature on operations management is certainly aware that the nature of the customer matters when deciding how to allocate resources. For example, Güneş & Akşin (2004) propose a theoretical model where a server provider must distinguish between customers of high and

low values. Similarly, Chen (2001) considers the case where a firm faces several segments of customers with different degrees of aversion in a supply chain setting. Research on queuing systems is especially prolific on describing how different prioritization policies could affect overall performance (for instance, see Pangburn & Stavrulaki, 2008 and Afeche & Pavlin, 2016). In general, this stream of literature provides interesting insights characterizing the equilibrium outcome, although unlike our research, they propose only analytical solutions to simple problems with no explicit connection to operational programs. In the present article we consider heterogenous customers within a detailed dispatching model that can be used to support actual decision making.

Previous literature has provided operational guidelines when customers are different in a number of contexts. In the context of capacity allocation, Hu, Li, Byon & Lawrence (2015) analyze a series of properties of dynamic prioritization policies. Similarly, Zhao, Xu, Li, & Liu (2016) determine optimal assignments to maximize the long-run throughput considering prioritized customer orders. Closer to our work, Jayamohan & Rajendran (2004) and Tay & Ho (2008) study a series of rules to prioritize jobs associated with more valuable customers. While these rules can be useful in practice, they are meant to work on average, and they are not embedded in a detailed math programming model that consider capacities and sequences as we do in our analyses.

In our research we deal with the technician dispatching problem, which can be considered a variant of the capacitated vehicle routing problem (CVRP). Several aspects of this problem have been widely studied before. There is an extensive literature addressing the problem with resource constraints, such as time windows, through exact methods based on column generation and some of the most sophisticated on branch-and-cut-and-price algorithms (Baldacci et al., 2010; Baldacci, Mingozzi, & Roberti, 2012; Cordeau, Gendreau, Laporte, Potvin & Semet, 2002; Kallehauge, 2008), as well as heuristic and metaheuristic approaches constructed mostly for real size problems (Vidal, Crainic, Gendreau & Prins, 2013; Gendreau, Potvin, Bräumlaysy, Hasle & Løkketangen, 2008; Golden, Raghavan & Wasil, 2008; Gendreau, Laporte & Potvin, 2002). To avoid unfeasibility, some authors consider soft time windows as resources (Cortés et al., 2014; Liberatore et al., 2011).

A special case of CVRP is the *heterogenous* CVRP, where vehicle fleet is characterized by different capacities and costs (Baldacci, Toth & Vigo, 2009). In our research, instead of analyzing heterogeneity on the supply side we analyze how a differentiation in costs from the demand side can have relevant impact on the optimal solution. In this regard, we focus on approaches where multiple objectives are considered in a weighted-sum objective function to solve vehicle routing problems (for a review, see Jozefowiez, Semet & Talbi, 2008). Here, the basic idea consists of adding objectives to the traditional cost minimization in order to improve customer satisfaction regarding delivery dates (Sessomboon, Watanabe, Irohara & Yoshimoto, 1998). In these approaches, the goal is finding good solutions that balance operational costs with well-defined service levels. In our case, we are interested in characterizing how different levels of heterogeneity in the actual cost due to delays for the customers impact optimal solutions generated by the firm. Moreover, we move one step forward by providing a simply approach to deal with that heterogeneity for the technician dispatching problem. With regard to vehicle routing schemes for technician dispatch problems, we can mention Blakeley, Argüello, Cao, Hall & Knolmajer, (2003) and Weigel & Cao (1999) who implemented technician dispatch systems for large companies using heuristics and GIS data. Other studies of service technician routing and scheduling are Cordeau, Laporte, Pasin & Ropke (2010), Xu & Chiu (2001) and Tang, Miller-Hooks, & Tomastik (2007), who implemented a tabu search heuristic for a maintenance dispatching system, formulated as a Multiple Tour Maximum Collection Problem with Time-Dependent rewards.

The problem we consider in this research is similar to that used by Cortés et al. (2014), who developed a daily dispatch of technicians formulated as a VRP with soft time windows, where the cost had three components: travel costs, a soft time window violation and a penalty for postponing customers to the next day. The penalty considered there is fixed, regardless of the features of the customer who is postponed, and therefore, the impact in the routing due to cost heterogeneity among customers is ignored. However, in that work the actual time of attention can affect the costs associated with a delay during the same day. In our research, for practical purposes we simplify that cost structure by leaving the soft time window constraint wide open only capturing the cost that customers face for been delayed to the next day. In synthesis, the present formulation has similarities with Cortés et al. (2014) in terms of the problem to be addressed and the general conditions of the technician dispatching model. However, in addition of having completely different research objectives, in the current version we are modelling a much longer horizon than only a day of operation, considering fixed capacity and variable penalties for being postponed, which are based on the heterogeneity cost models presented in section 4. In addition, Cortés et al. (2014) used a branch and price scheme to solve the problem. They show that for this particular application, the branch and price approach did not lead to significantly better solutions that just generating columns at the root node, and that is why in the model presented in section 3 we only generated solutions at the root node. This simplification notoriously reduces the computational work, which is fundamental for testing many scenarios and replications to discover the behavior of the optimization when customers are different in terms of cost heterogeneity.

To summarize, the present paper contributes to the literature by exploring how firms can benefit from this idea in the context of the dispatching problem, and then we provide a simple approach to endogenize customer cost heterogeneity using commonly available information such as the utilization rates and number of machines.

3. The technician dispatching model

In this section we describe the formal model we use for the technician dispatching problem. Let us assume a set of requests asking for repair services of office machines in a geographically dispersed area. The service is provided by a set of technicians working on the area, under a 24-hours-in advance service protocol.

The daily dispatching model we consider in this research is a Vehicle Routing Problem with Soft Time Windows (VRPSTW), in which the cost in the objective function has three components: travel costs, a wide-open soft time window violation and a customer-dependent penalty for postponing the attention to the next day. Thus, we consider a simple costs structure associated with the whole day of delay and we only study heterogeneity in the costs that customers experience due to postponing their attention to the next day. Although this is a simplifying assumption, it has no major impact on the main tradeoffs of the problem. The model can be extended to incorporate a cost structure that increases continuously with the delays, but it would make the decomposition computationally much more challenging. Moreover, in the model there is a hard constraint setting the maximum length of the working day of each technician, which means that not all the demand is necessarily attended during the requested day.

The context behind the modelling ideas requires to run simulations for multiple consecutive days. We consider optimizing the schedule for a full week where the capacity does not change from

one day to another. At the beginning of each day, the scheduler of the firm knows the demands left from the previous day plus the demands that need to be covered during the current day. The model is inspired in a real problem based on data from a major company offering repair services of office machines in Santiago, Chile. For the daily dispatching model, the set of service requests assigned during a given day come from the previous days, usually the day before, since the company attempts to enforce a 24-hour service policy. In the present model, the dispatcher selects which requests should be handled during the day, regardless of the time of attention during that day. Thus, those customers who are not served in a given day must be visited the next day.

Considering that most VRPSTW are quite difficult to solve using traditional network flow models (Taillard, Badeau, Gendreau, Guertin & Potvin, 1997), in this research, we formulate the problem through a column generation (CG) approach. In this work we decided to add columns only at the root node to make the model computationally easier to solve and the conclusions of this article are consistent and valid even tough in theory we do not reach optimality in some of the instances.

In general, column generation approaches have been very successful in solving various types of VRPs. CG approaches allow splitting the problem into two parts: the master problem to select the routes with the minimum total cost from a pool of feasible routes; and the sub-problems, which generate feasible routes that could potentially reduce the total costs. To solve the sub-problems, in this work we use Constraint Programming (CP), which has already been used successfully for a similar formulation of this problem we studied in Cortés et al. (2014).

3.1. Master problem

The master problem of the proposed VRPTW can be formulated as a set partitioning model assuming that it is possible to choose among different service routes for each technician available in an existing set of routes R. Let $\mathcal{K} = \{1, \ldots, K\}$ be the set of available technicians. Technicians must start the day at a location of a high priority request (in the set of those customers postponed from the day before) or at the depot. Let $I_1 = \{1, \ldots, K\}$ be the set of these locations. Then, each route $r \in R$ is characterized by a technician who starts a path at a specific location $i_1 \in I_1$, and then continues to visit a sequence of locations $\{i_2, .., i_e\} \in I_2$. Here I_2 is the set of all customers to be served during a specific day, not included in I_1. Each route, therefore, is described by the set r $\in$ {2, ..., e} where e is the last service request of a path. Therefore, the mathematical statement of the master problem is formulated as:

(M.P.)

$$\min \sum_{r \in R} c_r \theta_r + \sum_{i \in I_1 U I_2} p_i v_i \tag{1}$$

s.t.

$$\sum_{r \in R} a_{ir} \theta_r + v_i = 1 \qquad i \in I_1 U I_2 \tag{2}$$

$$\sum_{r \in R} \theta_r \leq V \tag{3}$$

$$\theta_r \in \{0, 1\} \qquad r \in R \tag{4}$$

$$v_i \in \{0, 1\} \qquad i \in I_1 U I_2 \tag{5}$$

This mathematical formulation considers two set of binary variables: θ_r that indicates whether the route $r \in R$ should be chosen or not, and variables v_i that are equal to one if customer i is not included in any of the chosen routes. The objective function is the sum of the costs of selecting a route c_r and the costs for postponing a customer to the following day p_i.In constraint (2) the binary parameter a_{ir} indicates if customer i belongs to route r, so this ensures that all customers are either in a selected route or moved to the next day. In constraint (3) the total number of routes that can be selected is limited by V the number of technicians. This formulation guarantees that there is always a feasible solution to the problem regardless of the definition of the set R.

3.2. Sub-Problems

The objective of the CG sub-problem is to produce new columns (routes) for the master problem, considering that if a column r not previously included in R, has a negative reduced cost, it can potentially improve the solution of the master problem. The reduced costs of a column is defined as the costs of the route c_r, minus the sum of the master problem dual variables $\pi_{s[l]}$ associated with constraint (1) and ϕ associated with constraint (2). The cost of the new route is the sum of the travel time cost and the violation of the time windows.

Let L be the maximum possible length of a route and $s[l]$, $l = 1 \ldots L$ an array of variables that represents a route, where the l^{th} element of $s[l]$ is the client in the position l. The first positions of all new routes must be a customer from set I_1 (below we explain implementation details and starting routes), from which the technician continues the route to either another customer from *set* I_2 or a fictional customer from the set I_3. These fictitious customers are added for CP modelling purposes and ensures that all routes are of length L. The maximum number of fictitious nodes that can be included in a route is $L-2$, therefore the set of fictitious nodes is defined as $I_3 = \{C+1, ..., C+L-2\}$, where C is the cardinality of set $I_1 U I_2$. Additionally, we define the variables w, d and t for the service start time at the l^{th} client, the violation of the time window and the travel time between two clients respectively. The sub-problem used for the generation of feasible routes is presented below.

(S.P.)

$$\min \beta \sum_{l=1}^{L} d[l] + (1-\beta) \sum_{l=1}^{L} t_{s[l-1],s[l]} - \sum_{l=1}^{L} \pi_{s[l]} - \phi \tag{6}$$

s.t.

$$w[1] = 0 \tag{7}$$

$$w[l] = w[l-1] + u_{s[t-1]} + t_{s[l-1],s[l]} \qquad l = 2, \ldots, L \tag{8}$$

$$d[l] = \max(0, w[l] - b) \qquad l = 1, \ldots, L \tag{9}$$

$$\text{Alldifferent(s)} \tag{10}$$

$$s[l] \in I_1 \qquad l = 1 \tag{11}$$

$$s[l] \in I_2 \quad l = 2 \tag{12}$$

$$s[l] \in I_2 \cup I_3 \quad l = 3, \ldots, L \tag{13}$$

$$s[l] = i, i \in I_3 \Rightarrow s[l+1] = i+1 \quad l = 3, \ldots, L \tag{14}$$

$$s[l] > first(I_3) \Rightarrow s[l-1] = s[l] - 1 \quad l = 3, \ldots, L \tag{15}$$

$$s[l] \leq first(I_3) \Rightarrow s[l-1] \leq first(I_3) \quad l = 3, \ldots, L \tag{16}$$

The objective function in (6) searches for the column (route) with minimum reduced costs, with four terms; the first two account for the real costs of the route (delay and travel time) while

the last two terms are the dual variables, associated with constraints (2) and (3) of the master problem, respectively. Constraints (7) and (8) compute the starting time of the service at the l^{th} customer in the route, where $u_{s[t-1]}$ is the required service time in the previous node. In (9) we define the violation of the time windows, which in this case is the same for all customers. Constraints (10) to (14) build the routes, forcing the first two nodes to be customers, and once the route reaches a fictitious node it can only go to another fictitious node until the end of the route. Constraints (15) and (16) are redundant constraints that greatly improve the CP resolution.

3.3. Implementation

In order to solve this particular set of experiments, As mentioned above, for the sake of simplicity, we do not consider time windows within the day, but only postponements to the next day. Also, we added a hard constraint in the number of technicians in the MP and the corresponding dual variable in the SP. As we pointed out before, to solve this problem we only generated columns at the root node, avoiding the use of a branch-and-price algorithm.

The master problem considers a penalty p_i that represent the cost that customer i experiences for not been served during the day. A key idea of this research is that an estimation of this penalty based on customers' internal costs can lead to different solutions. This is not only because it would allow prioritizing the provision of service for customers with higher costs, but also because it could provide better operational flexibility. We expect the impact in the solution to be dependent on several key parameters on the customer costs, and therefore we analyze several scenarios in which we vary the relative importance of customers' costs and degree of heterogeneity among them.

4. Treatment of cost heterogeneity in optimal dispatching

In this section we describe how the vehicle routing model described in Section 3 is combined with different ways to model cost heterogeneity.

As we highlighted in the introduction, we consider that the penalty of not complying with a requirement does not distribute homogeneously among customers. Our premise is that costs are different across customers, focusing our analysis on comparing service metrics for various scenarios. As introduced before, our analysis will pursue two main research objectives. First, to *understand how cost heterogeneity affects the performance of the optimal solutions*. To do so, we assume that customer costs p_i is known by the firm, and we run different scenarios considering the distribution of those costs among customers. We use those scenarios to evaluate variations in the total costs and other relevant operational performance metrics. The second objective is to *illustrate how a firm could implement our proposal if the customer costs are not known by the firm, but can be traced to observable characteristics of the customers*. More specifically, we describe the case of a service repairing homogeneous machines where both, the total number of available equipment and the utilization rate are observable at the customer level. Here, the total costs of customers waiting for service can be approximated through a closed form equation derived from an Erlang-C type of models. Next, we revise how we address these two research objectives.

4.1. Modeling the cost heterogeneity through penalties in the objective function

We start by assuming that the firm knows the magnitude of the penalties associated with not satisfying the requests during the day and we analyze how different distributions of those penalties affect the performance of the solutions. For example, the firm might have negotiated individually with each customer and determined a penalty that closely matches customer costs. In this research, we are interested in evaluating whether the mean magnitude and the variability of those costs affect optimal solutions. We build our scenarios based on a log-normal distribution for the penalties p_i, assuming that $\log(p_i)$ is normally distributed. The Normal distribution is a frequent choice to characterize randomness (Steward & Golden, 1982; Abdelaziz, Aouni & El Fayedh, 2007) as well as a distribution easy to be interpreted. The logarithm is justified because, by definition, delay penalties are positive. Thus, a given scenario is determined by a pair (m,s) corresponding to the mean and variance of the underlying normal distribution. To build the associated instance of the dispatching problem, we sample the values of p_i according to that normal distribution. To guarantee that changes in the optimal solutions are only explained by variations in mean and variance of the postponement costs, in our sampling procedure we keep the ordering in costs. This is to say, if the cost of a customer a are larger than the costs of a customer b in one instance, these relationships will hold for all scenarios. In the numerical evaluations we consider a total of twenty-one scenarios generated from three values for the mean m and seven values for the variance s for each value of m. The values of the means were selected to represent a wide range of reasonable scenarios of customer's costs. More specifically we consider three values of m: m_{Low}, m_{Medium} and m_{High}. The value of m_{Medium} is calibrated to be comparable with the current costs of a one day of delay. The values of m_{Low} and m_{High} are derived by simply increasing and decreasing the values of m_{Medium} by 50%.

The values of the variances were also selected to represent a wide spectrum of values ranging from scenarios with almost no variation to scenarios where the maximum penalty is twenty-five times the minimum value. In these twenty-one scenarios we consider that the penalties are independent of the complexity of the repair tasks. For a detailed description of the numerical values we considered for means and variances, see Appendix. Fig. 1 shows examples of the distribution of penalties we used to represent customer costs for scenarios with relatively small and relatively large values of the variance s.

4.2. A Markovian model to characterize customer costs

Oftentimes, firms do not have precise information of customer costs and they must approximate such costs based on observable characteristics. This is for example the case of internal customers where no explicit contracts are available to characterize the terms of the service. In this section we propose a simple model to estimate the costs for not providing the service on the requested day. We believe that this approach is suitable for the case of a service repairing homogeneous machines where both, the total number of available equipment and the utilization rate, are observable by the firm at the customer level. Here, the total costs of client waiting for the service can be approximated through a closed form equation, where we rely on queuing theory that has been extensively used in other domains such as health care (Lakshmi and Iyer, 2013) and telecommunications (Giambene, 2005).

Our main objective of this exercise is to use specific customer information, which the firm should have, to approximate a penalty representing customer costs of being delayed. For that purpose, we assume that each customer i has n_i machines and each one of them processes its jobs in an exponentially distributed service time of mean $1/\mu$. The processing time of each server is an operational feature of a printer machine. In our empirical application this time exhibits little variation between printer models, and therefore we assume that this time is homogeneous. We also

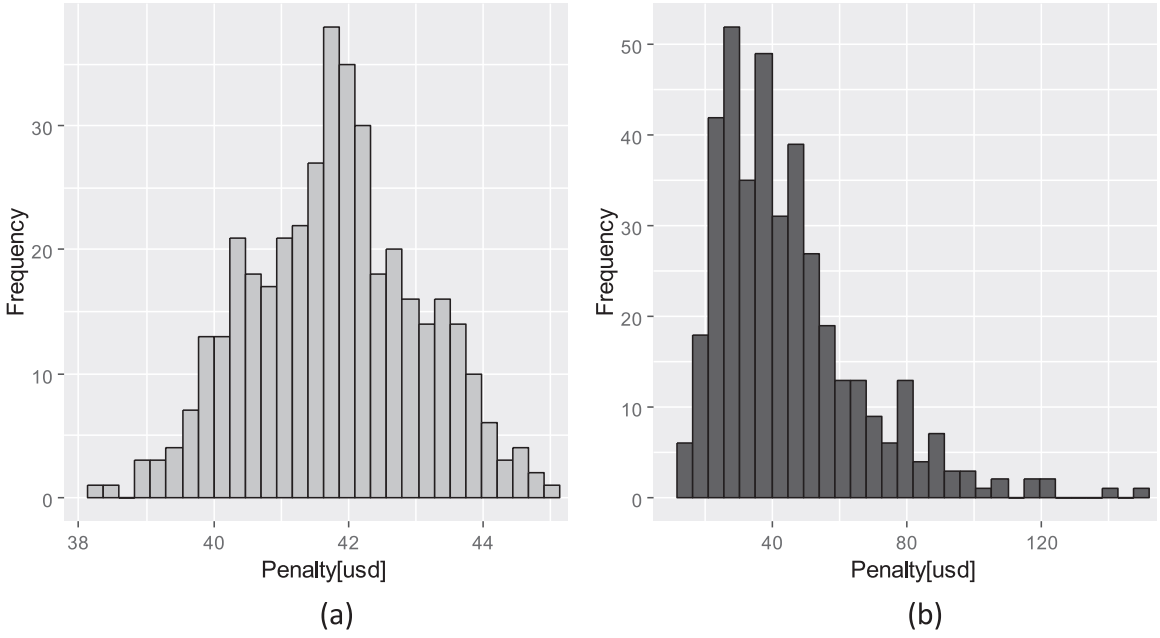

Fig. 1. Distribution of Penalties among customers. Left panel (a) displays the distribution with small variance. Right panel (b) displays the distribution with large variance.

assume the workload is given by a Poisson arrival process with rate λ_i. Under these assumptions, the internal use of the machines can be described as bird-death process characterized by a utilization rate $\rho_i = \lambda_i/(n_i \cdot \mu)$ that represents the average fraction of time that each machine is being used.

When a machine is broken and requires to be repaired, the system reduces its capacity from n_i to $n_i - 1$ servers. These reductions translate into longer waiting times that can be translated into a monetary cost. Formally speaking, under the assumptions we describe above, the mean time in the system of customer i (W_{s_i}) has a closed form expression given by equation (17)

$$W_{s_i} = \frac{C(n_i, \rho_i)}{n_i\mu - \lambda_i} + \frac{1}{\mu} \tag{17}$$

In this expression, $C(n_i, \rho_i)$ corresponds to the probability that an incoming job must wait because the servers are busy. This probability also has a closed form solution as shown in Equation (18) (Tackacs, 1969; Harel, 1988).

$$C(n_i, \rho_i) = \frac{1}{1 + (1-\rho_i)\left(\frac{n_i!}{(n_i\rho_i)^{n_i}}\right)\sum_{k=0}^{n-1}\frac{(n_i\rho_i)^k}{k!}} \tag{18}$$

To derive the monetary value of not being served, we assume that such costs are proportional to the additional time in the system given by the difference between $W_{s_i}(n_i - 1) - W_{s_i}(n_i)$. Notice that this value depends on the number of machines and the utilization rate, information that is readily available in many real cases. In our application, the number of machines is directly observable because firm's maintenance databases register not only the number of machines, but also the specific model and tenure of each machine. While these results are fairly standard from queuing theory, we consider them useful to provide some intuition about how they translate to our specific setting. Fig. 2 illustrates how the impact of total time in the system is affected by the two observable variables (utilization and number of servers).

In both panels of Fig. 2, we show how the waiting times are affected when one machine is not available. In the left panel we display the variation in waiting times as a function of number of servers, while in the right panel we display such a variation as a function of utilization rates. When the firm has a large number of servers, the waiting times are not very sensitive to the failure of a single machine, but the impact is large when the firm only has few machines to compensate for the unavailability of one of them. Similarly, the waiting times are fairly robust if the utilization rates are mild, but they become very sensitive to the availability of machines when the utilization rate is high. Thus, using our proposed Markovian model, the dispatching problem should prioritize customer with either fewer machines or larger utilization rates.

5. Results

For the numerical analyses, in all instances we use the same set of customers demanding repair services over one week of operation. To reduce the computational time in the GC search, we defined a large set of starting columns. To characterize the effect of including cost heterogeneity in the dispatching problem, we performed a collection of computational exercises where we characterize how the optimal solutions change when customer costs are incorporated. In these exercises, we use data of real requirements for a company serving customers located in the city of Santiago, Chile. The total number of requests received everyday shows important variation and weekly seasonality as shown in Fig. 3(a). On average, the firm receives 194.63 repair request/day. Monday is the day of the week with highest demand, while Friday has the lowest demand. While demand includes several geographical zones, in our empirical application we only consider a subset covering nearly a half of those requests. For the whole dataset, repair tasks also differ in their complexity and some of them take longer than others, ranging from 7 to 435 minutes to be completed as is illustrated in the histogram of Fig. 3(b). In our empirical setting, there is a fleet of 20 vehicles (technicians) that should be used to satisfy all the weekly demand. We note that this is the situation where technicians are regular employees of the firm. We evaluated other capacity profiles and the results remain qualitatively unaltered. For example, if we move 5 capacity units from Friday to Monday, so that we have 25 vehicles available on Monday and only 15 on Friday, the fraction of postponed customers drops from 8 to 5%. In this scenario the firm still has to decide which clients to postpone and therefore the methodology we propose here is still useful.

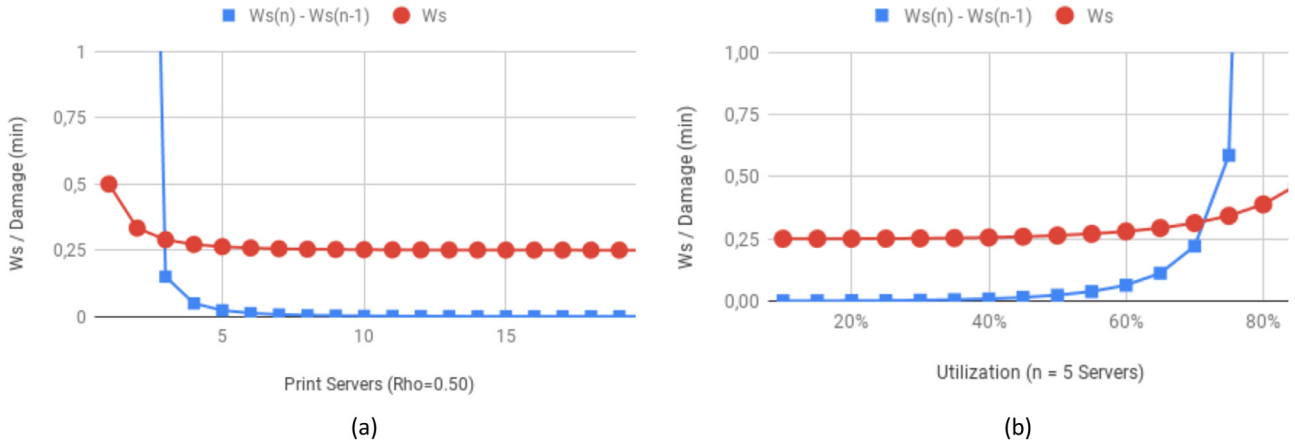

Fig. 2. Illustration of time in the system for representative values of number of servers and utilization rates (a) W_s and difference in W_s for different values of number of servers (the utilization rate is kept fixed at ρ=0.5). (b) W_s and difference in W_s for different values of the utilization rate (the number of servers is kept fixed at n=5).

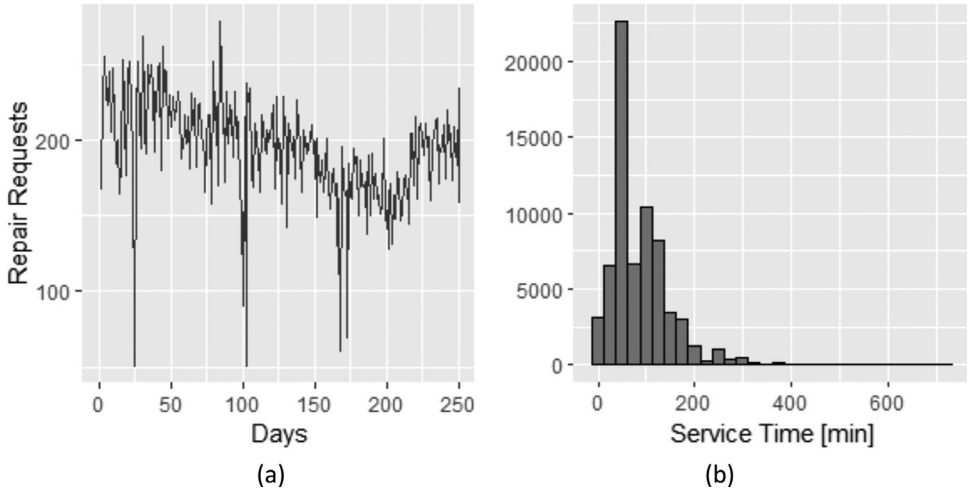

Fig. 3. Daily Demand (a) and Distribution of service times (b).

Table 1
Instance Demand by Weekday and Sector.

Location	Monday	Tuesday	Wednesday	Thursday	Friday
Sector 1	5	7	2	2	1
Sector 2	14	14	8	10	11
Sector 3	20	19	22	19	19
Sector 4	57	48	46	36	28
Sector 5	4	2	2	3	1

Table 2
Instance Average Service Time [min] by Weekday and Sector.

Location	Monday	Tuesday	Wednesday	Thursday	Friday
Sector 1	188.00	76.43	45.00	60.00	150.00
Sector 2	98.21	78.64	108.75	85.50	92.18
Sector 3	90.50	106.58	66.36	81.63	66.63
Sector 4	99.16	96.17	86.09	116.14	85.36
Sector 5	53.75	95.00	57.50	150.00	10.00

The instances we use in our numerical evaluations correspond to a full labor week (Monday to Friday). Within each instance, we determine which customer is going to be served everyday by solving the vehicle routing model previously described in section 3. If the request of a given customer is not satisfied during the day, the corresponding demand is delayed to the next day when it must be served with the highest priority. All these instances are built using the same customer demand derived from a representative week. Tables 1 and 2 display the number of requests per day and the corresponding average service times for that representative week. In this table, we decompose the demand by sector representing geographically clustered customers.

The main operating costs of the dispatching firm are associated with the remuneration of personnel and the maintenance of the necessary assets to provide technical assistance. For the computational exercise we have considered two operational costs: the travel costs equal to USD 5.07 per hour and the overtime costs equal to USD 7.38 per hour.

5.1. Evaluation of heterogeneity of penalty costs

In our first experiment we assume that there is a probability distribution that characterizes the penalty associated with delaying a request that is known by the service provider. We evaluate the impact of cost heterogeneity using some key performance indices. We start by looking at the fraction of customers who are delayed to the next day as shown in Fig. 4. In this figure, the scenarios denoted by p05, p10, p15, p20, p50 and p100 are differentiated by the distribution of the costs of delay; the larger the number tagged in this notation, the higher the variability assumed for the costs of delay in the population of customers. We observe that regardless of how large the penalties are, more variation in their values

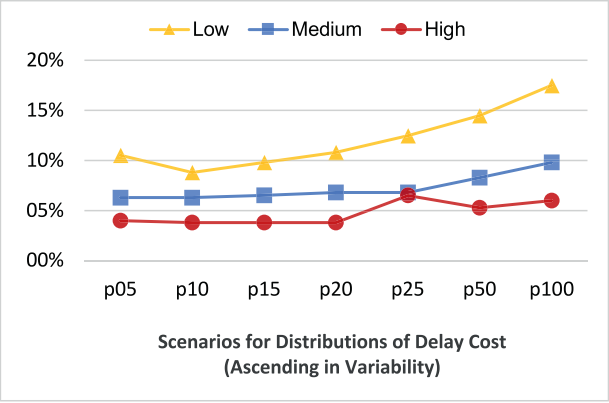

Fig. 4. Percentage of customers who are not served in the requested day (1-week). Each line corresponds to a given value for the mean of the distribution.

is associated with a larger proportion of delayed customers. The intuition is straightforward: if customers are more diverse in the costs they assign for not being served during the same day, then the firm is going to find more customers who are not sensitive to these delays, and therefore they are going to be postponed to favor transportation efficiency. From the figure, we also observe that the effect appears to be more pronounced for relatively low values of the penalty. The intuition behind these results is clear: when the costs are high, even for customers at the bottom of the distribution, the cost of delay is large enough to discourage the firm to postpone their service and pay the corresponding costs.

Knowing that more variation in costs is associated with more customers being delayed, we complement our previous analysis by analyzing the delay costs of those customers who were delayed compared against those who did not. Boxplots of Fig. 5 compares the delay costs of satisfied (S) against non-satisfied customers (N). The boxplots show that the routing algorithm systematically chooses to serve customers with higher penalties and delays those customers with lower penalties. While this difference is marginal for the scenarios of low penalty costs, the difference gets more pronounced when the costs becomes larger. In fact, when the penalties are large on average, it is more expensive for the firm to delay customers, and therefore, they are postponed only in those few cases where the costs are affordable. In the figure we only display the case for a relatively high variability scenario, but qualitatively speaking, these patterns are consistent regardless of the degree of variability associated with the penalties.

We now turn our attention to the components of the objective function to evaluate whether these qualitative changes in the solution translate into operational efficiencies. Recall that in the objective function of the master problem we have two main costs components: the transportation costs (driven by c_i parameters) and the penalty costs associated with order delays (driven by p_i parameters). We first analyze the transportation costs as illustrated in Fig. 6(a). As we already explained, more variability is associated with larger flexibility to postpone some of the requests, which has a direct impact on transportation costs. If some customers can be delayed at a low cost, the firm can benefit by designing more efficient routes. Thus, as the variation increases, transportation costs are reduced. This pattern is consistent for all mean values of the distribution.

Fig. 6(b) shows how the other component of the costs is affected by larger variation in the penalties. Considering the previous results showing that more variation implies a larger proportion of customers being delayed, we expected that total costs associated with the delays are going to be increasing with the variability. Surprisingly, this is not the case, and the costs that the firm must incur to compensate these delays tends to decrease with more variation in the values of the penalties. The intuition behind this result is as follows: the instances we considered exhibit important daily variation in the demand and there are some days where the entire set of requests cannot be satisfied with the installed capacity. This is especially true on Mondays where the additional demand generated over the weekend must be satisfied. In these days, the firm is forced to delay some customers. Recognizing the variation in the penalties allows the firm to select those customers with lower costs.

Overall, these results provide clear evidence that a detailed assessment of customer costs can have a relevant impact in how the firm allocates resources. Furthermore, those changes directly translate into a meaningful reduction of total costs.

5.2. A markovian model for determining customer's costs

In this subsection we show the results of estimating customer's costs using the Markovian model described in Section 4.2. This model uses specific customers' information, such as the number of print terminals and their utilization rates, to approximate a penalty value associated with each service request. Using observable values of these variables, we computed a penalty for each customer

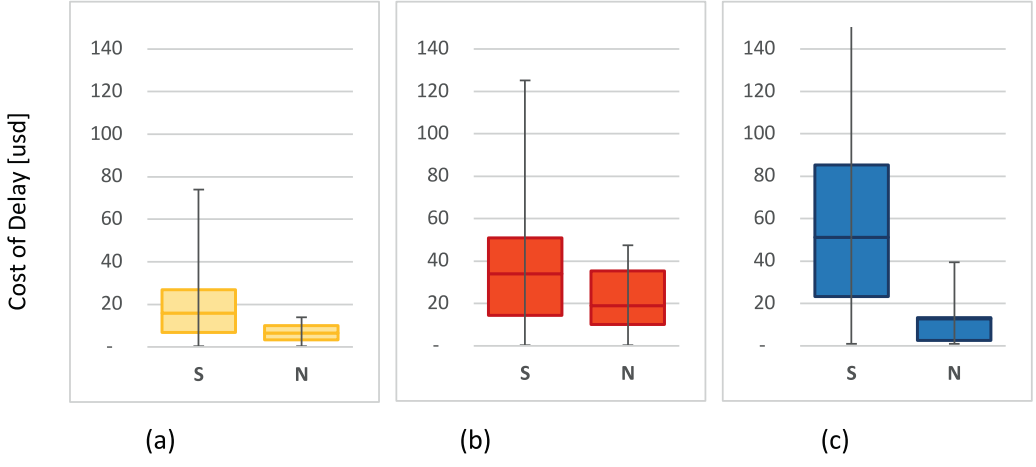

Fig. 5. Boxplots comparing the difference in the penalty cost [USD] between satisfied (S) and unsatisfied costumers (N). For simplicity we only display scenarios with relatively high variance. (a) low penalty mean, (b) medium penalty mean and (c) large penalty mean.

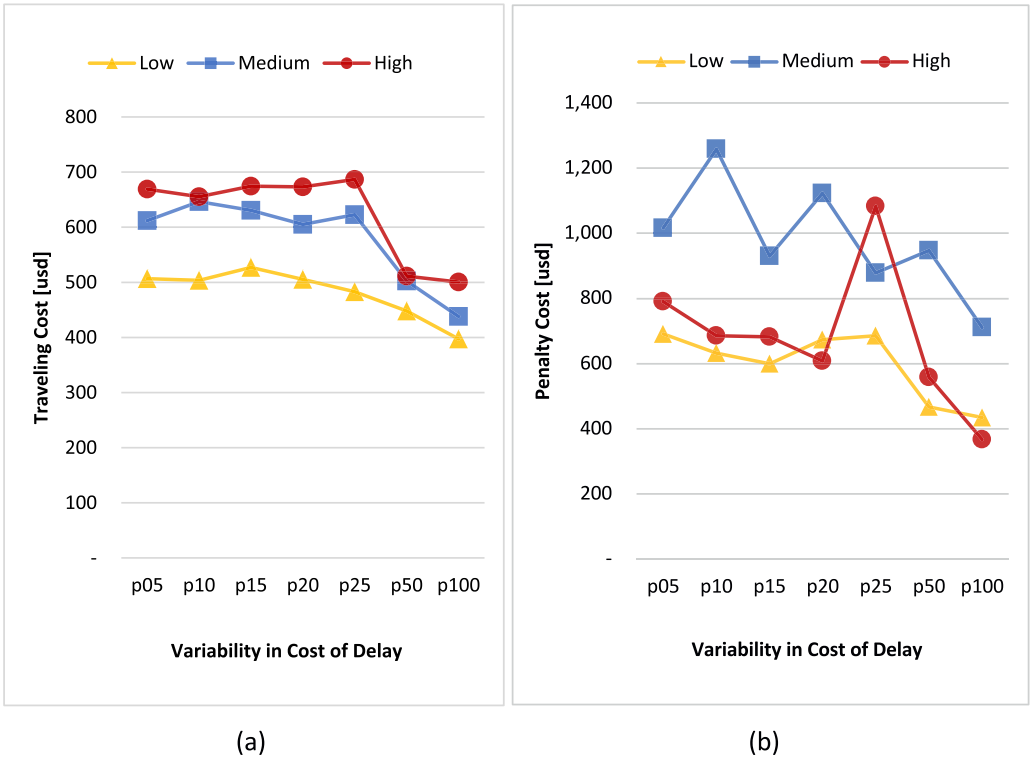

Fig. 6. Components of cost function. (a) Traveling cost (1-week) (b) penalty cost.

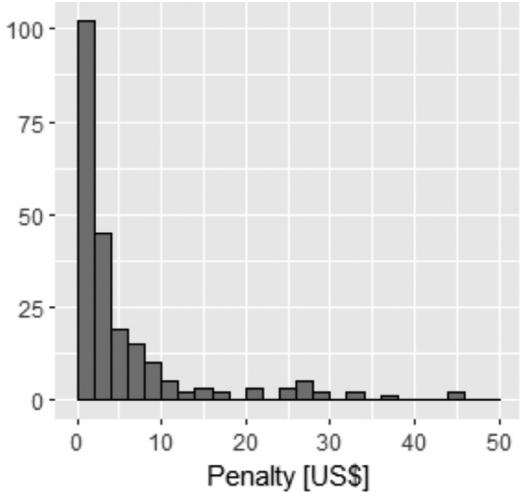

Fig. 7. Distribution of penalty costs inferred from the Markovian Model.

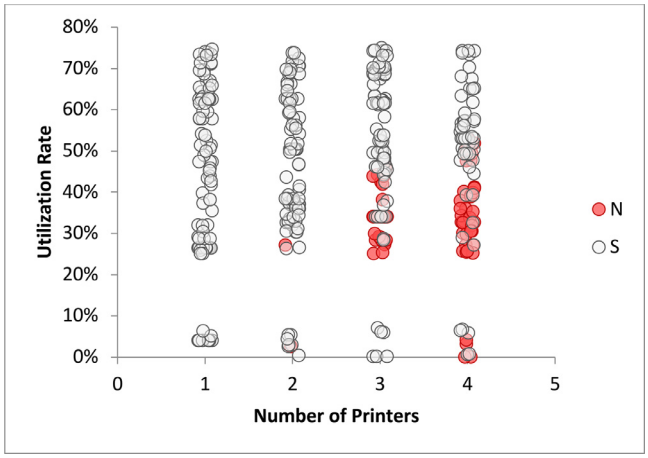

Fig. 8. Comparison between satisfied and unsatisfied customers with respect their number of servers (c_i) and Utilization rates (ρ_i).

according to closed form expressions of Equations (17) and (18). The resulting distribution of penalties is displayed in Fig. 7.

Unlike previous parametric distribution we used, this distribution is the result of using individual level data to provide an estimate to the costs of each customer. Similar to other scenarios we already analyzed, the Markovian model captures important variations in the penalty costs, but compared to the log-normal densities we used before, this empirical distribution is more skewed with a long positive tail. Our Markovian model indicates that a large fraction of customers have enough flexibility in their printing systems to tolerate a delay for a day with minimal impact in their functioning. However, the other group of customers show a much more saturated behavior and therefore require urgent satisfaction of their repair requests. In terms of the optimization problem, the Markovian model is effectively discriminating which customers must be served in the requested day and which customers can be delayed if there are enough reductions in transportation costs. To understand how the optimization model decided which customers to serve each day, in Fig. 8, we compare satisfied (S) and non-satisfied (N) customers in terms of their number of machines and utilization rates.

In the Figure, customers in the left and upper zone, with low number of printers and high utilization will not be delayed, while customers in the lower, right hand side, with a higher number of printers and low utilization will tend to be delayed for the next day.

The results of the optimization problem show that the model is indeed guided by the costs estimated from the Markovian model and the majority of delayed customers have low utilization rates

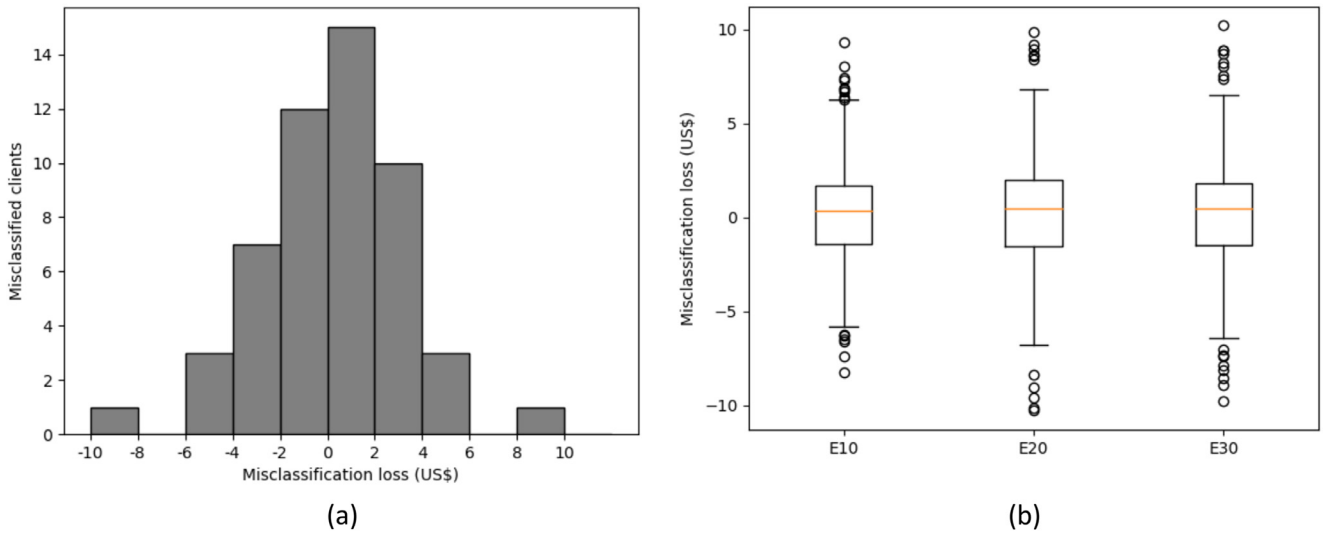

Fig. 9. Misclassification Losses. (a) Distribution for a representative instance (b) Summary of all instances grouped according to the magnitude of the forecasting error.

and large number of machines. For example, customers with only one printer were never delayed regardless of their utilization rate. This is reasonable as customers with a single machine have no means to compensate the failure and therefore their waiting times explode. On the other hand, customers with 4 machines are much more likely to be delayed, but among those, customers with lower utilization rates are more likely to be postponed. This is again reasonable as the impact on their functioning of having one printer not working is small.

In the previous exercise, we implicitly assume that Markovian model perfectly captures the customer costs of delay. Certainly, by adding more observable variables we can refine the model to have more precise estimate of the costs. However, it is plausible to think that the Markovian model is only going to be an approximation to the true costs of the customers. To understand the role of estimation errors in the dispatching problem we run a series of complementary scenarios where we progressively increase the mean error associated to the Markovian model. In these models we assume the real costs of customer i ($R(n_i, \rho_i)$) is equal to the cost derived from the Markovian model plus a zero-mean error term ε_i, as indicated by Equation (18).

$$R(n_i, \rho_i) = W_{s_i}(n_i - 1) - W_{s_i}(n_i) + \varepsilon_i \tag{18}$$

In these scenarios we consider that ε_i is normally distributed with variances that we vary to represent between 10 and 30% variation with respect to the real cost. To assess the impact in the solution we compare the solution of the Markovian model against a hypothetical oracle solution where we observe true customer costs with no uncertainty. In this comparison we first compute the number of customers that, due to the estimation errors, were postponed. Here we find that a relevant fraction of the customers is affected by the forecasting error and they must wait an additional day that they would not need to wait if the central planer would have perfect costs information. In fact, on the 30 instances we run, we find that on average 6.9% of customers are incorrectly delayed. Notice however that, as these customers are postponed, it leads to a higher number of vehicles being available that can be used to serve other customers that would have been delayed with perfect information. On average, we find that 5.3% of customers benefit from this miscalculation. To understand if these two subsets of misclassified customers are balanced or not, we compute for each customer, the *misclassification loss* (ML), corresponding to the effective postponement costs incurred by the dispatching solution. If $ML_i > 0$ then customer i is worse off with the solution implied by the imprecise estimation of the postponement costs. Similarly, $ML_i < 0$ indicates that customer is better off. These results are illustrated in Fig. 9.

Fig. 9a, displays the distribution of ML for a representative scenario (cases with $ML_i = 0$ are not relevant and therefore not considered in the figure). The distribution is fairly symmetric around zero, indicating that the additional penalty cost associated to customers who were incorrectly delayed is mostly compensated by the gain of those who were not postponed. For this scenario, the mean misclassification cost is only 0.28USD which is negligible compare to postponement costs we use in our instances. Fig. 9b shows the values of *ML* for all misclassified customers in all 30 scenarios and confirms the pattern observed in the example of Fig. 9a. More precisely, Fig. 9b confirms that (i) in general the misclassification loss of incorrectly delayed customers is quite compensated by the capacity gain that is used to promptly serve other customers and (ii) this pattern is not very sensitive to the magnitude of the forecasting error (at least in the 10-30% range we explored). One explanation to these results is that the empirical distribution of costs we used have enough customers with small costs of being delayed (see Fig. 7) and therefore any forecasting error only implies the delay of relatively low cost customers. We test this in our dataset and we did find that all customers what were incorrectly delayed have costs of postponement that are below-the-median.

6. Synthesis and conclusions

In this research we analyzed the impact in the optimal daily dispatching of repair services during a week of operation, in the case where customers have different costs of delaying the service. We used an optimal routing model for dispatching the technicians, considering penalties in the objective function associated with the delay of the service from one day to the next. The vehicle routing model follows a typical column generation approach, in this case solved at the root with high efficiency. In general, previous literature has assumed as given the aforementioned penalizations, either based on contractual rules, or as perceived loss of goodwill in customers. In our case we tried to characterize the real costs incurred by the customer due to delay in service. In our analysis we compared key performance metrics of optimal dispatching solutions for a number of scenarios, which lead us to several conclusions. As a first order effect, we found that the way in which delay costs are distributed in the population matter and that they

can lead to sizable cost reductions. We built our setting based on the data and operation of a real company that provides office equipment repair service and therefore we worked with the original fleet as well as the strategy used by the company, which consist of a fixed capacity. In our analysis, we considered the case where the firm has the flexibility of allocating more capacity for days with larger demand which leads to mild cost reductions. In this regard, we consider interesting to analyze in future research how the benefits of learning about customers' costs of delays can be combined with a more flexible contract for allocating capacity over time.

Our analysis also indicates that, keeping everything else constant, firms are better off with more variation in customer costs. The underlying intuition behind this result is that more variability implies a larger fraction of customer not sensitive to delays. These customers can be postponed to give more degree of freedom in vehicle routing leading to more cost-effective solutions. As a consequence of this gain in flexibility we did expect that transportation costs go down and this is indeed the case. However, considering that the number of delayed customers increases, we expected that total costs associated with delays compensation would also increase. Our results indicate that this is not the case.

The intuition behind this result is that when, on some days, the system receives more requests that its capacity can handle, and therefore some customers must be delayed, more variability in costs provides the flexibility to delay only cheap customers.

While the degree of variation has an important effect on optimal solutions, the mean values of the penalties also matters. In fact, the penalties moderate the operational gains of exploiting costs heterogeneity. When the mean of customers' costs is high, delaying any customer is expensive, and therefore there is little room to improve vehicle routing efficiency. Acknowledging that elicitation of customer costs might be difficult in some cases, in this research we also propose a simple Markovian model to show that costs of the delay can be approximated using only the number of servers and the utilization rate, which are known. We applied this approach and solved the corresponding dispatching problem, showing that the solution gives higher priority to customer who are more sensitive to machine malfunctioning, due to having fewer machines or high use of them. These results hold even if the Markovian model provides a noisy signal of the true costs of delaying the service.

In this research we investigated the technician dispatching problem and we demonstrate that heterogeneity in the costs of the delay can have a relevant impact on operational efficiency. Our choice of focusing on this specific problem is to illustrate the relevance of accounting for cost heterogeneity in a practical problem in which most operational considerations are considered in the model. Nevertheless, we expect the general idea can be extended to other classes of vehicle routing problems. We believe that firms can benefit from our approach, as far there is given capacity and resources can be reallocated depending on customers costs.

There is a number of interesting extensions that can be implemented to gain further insight on the role of variability in dispatching solutions. When solving this problem, a central planner must balance the relative costs of serving customers and their waiting costs. Our methodology certainly allows for an arbitrary weight to balance those costs; however, we have assumed a simple cost structure. In some cases, the dispatching firm can have a well identified group of costumers of higher strategic value with larger costs of being delayed. These would translate into our model in a bimodal distribution for the penalty costs with one mode for regular customers and another for strategic ones. To be more general, while we restricted our attention to normal case, our analysis can be easily extended to working with other distributions.

To conclude, through this article, we evaluated different scenarios considering the information set as given. However, the results of this research could motivate practitioners in the domain of transportation planning to actively learn from their customers. While the recent advances in data analytics suggest that learning from customers should be readily accessible to firms, the results derived from this investigation indicate that such learning can have a direct impact on operational planning policies.

Funding

This work was supported by projects ANID/FONDECYT/REGULAR # 1191531; ANID/FONDECYT/REGULAR # 1191200; and the Complex Engineering Systems Institute ANID PIA/APOYO AFB180003.

Declaration of competing interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix mean and variances of each scenario

We build our scenarios assuming $log(p_i)$ is normally distributed with mean m and variance s. The value of $m_{Medium} = 33.231\,USD$ is calibrated to be comparable with the current cost of a one day of delay. The values of $m_{Low} = 16.615$ and $m_{High} = 49.846$ are computed by increasing and decreasing the values of m_{Medium} by 50%. In terms of the variance s,

The values of s are chosen to represent a wide range of variability. The variance of each scenario p0x for mean m is computed as s=m·x/100. For example, the scenario p10 for medium variance is computed as s=33.231·0.1=3.323. The whole set of mean and variances used in the analysis are reported in Table A1.

Table A1
Numerical values for mean and variances for the scenarios included in the analysis.

	m	s						
		P05	P10	P15	P20	P25	P50	P100
Low	16.615	0.831	1.662	2.492	3.323	4.154	8.308	16.615
Medium	33.231	1.662	3.323	4.985	6.646	8.308	16.615	33.231
High	49.846	2.492	4.985	7.477	9.969	12.462	24.923	49.846

References

Abdelaziz, F. B., Aouni, B., & El Fayedh, R. (2007). Multi-objective stochastic programming for portfolio selection. *European Journal of Operational Research, 177*(3), 1811–1823.

Afeche, P., & Pavlin, J. M. (2016). Optimal price/lead-time menus for queues with customer choice: Segmentation, pooling, and strategic delay. *Management Science, 62*(8), 2412–2436.

Amiri, A., & Salari, M. (2018). Time-constrained maximal covering routing problem. *OR Spectrum, 41*(2), 415–468.

Baldacci, R., Toth, P., & Vigo, D. (2009). Exact algorithms for routing problems under vehicle capacity constraints. *Annals of Operational Research, 175*(1), 213–245.

Baldacci, R., Bartolini, E., Mingozzi, A., & Roberti, R. (2010). An exact solution framework for a broad class of vehicle routing problems. *Computational Management Science, 7*(3), 229–268.

Baldacci, R., Mingozzi, A., & Roberti, R. (2012). Recent exact algorithms for solving the vehicle routing problem under capacity and time window constraints. *European Journal of Operational Research, 218*(1), 1–6.

Blakeley, F., Argüello, B., Cao, B., Hall, W., & Knolmajer, J. (2003). Optimizing periodic maintenance operations for Schindler Elevator Corporation. *Interfaces, 33*(1), 67–79.

Chen, F. (2001). Market segmentation, advanced demand information, and supply chain performance. *Manufacturing & Service Operations Management, 3*(1), 53–67.

Cordeau, J. F., Gendreau, M., Laporte, G., Potvin, J. Y., & Semet, F. (2002). A guide to vehicle routing heuristics. *Journal of the Operational Research society, 53*(5), 512–522.
Cordeau, J. F., Laporte, G., Pasin, F., & Ropke, S. (2010). Scheduling technicians and tasks in a telecommunications company. *Journal of Scheduling, 13*(4), 393–409.
Cortés, C. E., Gendreau, M., Rousseau, L. M., Souyris, S., & Weintraub, A. (2014). Branch-and-price and constraint programming for solving a real-life technician dispatching problem. *European Journal of Operational Research, 238*(1), 300–312.
Fagerholt, K. (1999). Optimal fleet design in a ship routing problem. *International transactions in operational research, 6*(5), 453–464.
Gendreau, M., Laporte, G., & Potvin, J. Y. (2002). Metaheuristics for the capacitated VRP. *The vehicle routing problem* (pp. 129–154). Society for Industrial and Applied Mathematics.
Gendreau, M., Potvin, J. Y., Bräumlaysy, O., Hasle, G., & Løkketangen, A. (2008). Metaheuristics for the vehicle routing problem and its extensions: A categorized bibliography. *The vehicle routing problem: latest advances and new challenges* (pp. 143–169). Boston, MA: Springer.
Giambene, G. (2005). Queuing theory and telecommunications. Springer US.
Golden, B. L., Raghavan, S., & Wasil, E. A. (Eds.) (2008). *The vehicle routing problem: latest advances and new challenges* (43). Springer Science & Business Media.
Güneş, E. D., & Akşin, O. Z. (2004). Value creation in service delivery: Relating market segmentation, incentives, and operational performance. *Manufacturing & Service Operations Management, 6*(4), 338–357.
Harel, A. (1988). Sharp bounds and simple approximations for the Erlang delay and loss formulas. *Management Science, 34*(8), 959–972.
Homburg, C., Droll, M., & Totzek, D. (2008). Customer prioritization: does it pay off, and how should it be implemented? *Journal of Marketing, 72*(5), 110–130.
Hill, A. V. (1992). An experimental comparison of dispatching rules for field service support. *Decision Sciences, 23*(1), 235–249.
Hu, X., Li, Y., Byon, E., & Lawrence, F. B. (2015). Prioritizing regular demand while reserving capacity for emergency demand. *European Journal of Operational Research, 247*(2), 472–487.
Jayamohan, M. S., & Rajendran, C. (2004). Development and analysis of cost-based dispatching rules for job shop scheduling. *European journal of operational research, 157*(2), 307–321.
Jozefowiez, N., Semet, F., & Talbi, E. G. (2008). Multi-objective vehicle routing problems. *European journal of operational research, 189*(2), 293–309.
Kallehauge, B., Larsen, J., Madsen, O. B., & Solomon, M. M. (2005). Vehicle routing problem with time windows. *Column generation* (pp. 67–98). Boston, MA: Springer.
Kallehauge, B. (2008). Formulations and exact algorithms for the vehicle routing problem with time windows. *Computers & Operations Research, 35*(7), 2307–2330.
Lakshmi, C., & Iyer, S. A. (2013). Application of queueing theory in health care: A literature review. *Operations research for health care, 2*(1-2), 25–39.
Liberatore, F., Righini, G., & Salani, M. (2011). A column generation algorithm for the vehicle routing problem with soft time windows. *4OR, 9*(1), 49–82.
Niraj, R., Gupta, M., & Narasimhan, C. (2001). Customer profitability in a supply chain. *Journal of marketing, 65*(3), 1–16.
Pangburn, M. S., & Stavrulaki, E. (2008). Capacity and price setting for dispersed, time-sensitive customer segments. *European Journal of Operational Research, 184*(3), 1100–1121.
Sessomboon, W., Watanabe, K., Irohara, T., & Yoshimoto, K. (1998). A study on multi-objective vehicle routing problem considering customer satisfaction with due-time (the creation of Pareto optimal solutions by hybrid genetic algorithm). *Transaction of the Japan Society of Mechanical Engineering*.
Stringfellow, A., Nie, W., & Bowen, D. E. (2004). CRM: Profiting from understanding customer needs. *Business Horizons, 47*(5), 45–52.
Takacs, L. (1969). On Erlang's formula. *The annals of mathematical statistics, 40*(1), 71–78.
Taillard, É., Badeau, P., Gendreau, M., Guertin, F., & Potvin, J. Y. (1997). A tabu search heuristic for the vehicle routing problem with soft time windows. *Transportation science, 31*(2), 170–186.
Tang, H., Miller-Hooks, E., & Tomastik, R. (2007). Scheduling technicians for planned maintenance of geographically distributed equipment. *Transportation Research Part E: Logistics and Transportation Review, 43*(5), 591–609.
Tay, J. C., & Ho, N. B. (2008). Evolving dispatching rules using genetic programming for solving multi-objective flexible job-shop problems. *Computers & Industrial Engineering, 54*(3), 453–473.
Vidal, T., Crainic, T. G., Gendreau, M., & Prins, C. (2013). Heuristics for multi-attribute vehicle routing problems: A survey and synthesis. *European Journal of Operational Research, 231*(1), 1–21.
Vogel, V., Evanschitzky, H., & Ramaseshan, B. (2008). Customer equity drivers and future sales. *Journal of marketing, 72*(6), 98–108.
Weigel, D., & Cao, B. (1999). Applying GIS and OR techniques to solve Sears technician-dispatching and home delivery problems. *Interfaces, 29*(1), 112–130.
Xu, J., & Chiu, S. Y. (2001). Effective heuristic procedures for a field technician scheduling problem. *Journal of Heuristics, 7*(5), 495–509.
Zhao, Y., Xu, X., Li, H., & Liu, Y. (2016). Prioritized customer order scheduling to maximize throughput. *European Journal of Operational Research, 255*(2), 345–356.