**Accounting for cost heterogeneity on the demand**

**for the technician dispatching problem**

Juan P. Cavada (xx@xx.xx)

Cristian Cortes (xx@xx.xx)

Marcel Goic (mgoic@dii.uchile.cl)

Andrés Weintraub (aweintra@dii.uchile.cl)

Juan I. Zambrano (xx@xx.xx)

March, 2019

**Abstract**

In the technician dispatching problem, a given number of repair teams must visit different locations to provide service support. Considering there is a fixed capacity and variations in the demand, not all requests can be satisfied every on time and therefore some of them must be delayed. Most implementations of the dispatching problem consider a penalty associated to this delay that might vary depending on the customer. In this research we analyze how such variation in cost affect the outcome of service planning. More specifically, we explore how the distribution of those costs affects optimal solutions and then we propose a simple Markovian model to capture cost-heterogeneity for the case where cost of failure can be traced to observable operational characteristics. Our results indicate that when customers are different enough, transportation and total penalty costs decreases implying a sizable gain in operational efficiency. Moreover, results from the Markovian model indicate that firms can take advantage of these operational gains even with only a few customer characteristics are observed.

**Keywords:** Vehicle Routing, Costs Heterogeneity, Markovian Model.

Author affiliations…

# Introduction

Many important operational problems require the allocation of scare resources to satisfy a myriad of customer requirements. For example, in Vehicle Routing Problems (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 and therefore, if there is variability in the demand, some of the requirements cannot be satisfied and must be delayed. Literature on VRP has proposed different approaches to deal with this shortage of capacity, including penalizations for exceeding capacity (REF pending) and explicitly accounting for the cost 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 interval, central planner can consider time windows as soft constraints that can be violated if the direct cost of the violation can be compensated by a larger gain in transportation costs (Liberatore et al, 2011). Similarly, Cortés et al (2014) used fixed penalties for not responding on time to a service request.

In these approaches, optimal solutions are obtained by carefully accounting for operational costs of the supplier. This might include the cost of transportation which is usually composed of the cost of maintaining a fleet of vehicles, the cost of travel and the cost for delay in case there are time windows (Kallehauge et al, 2005). However, the costs in the demand side is either neglected or included 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 and Salari, 2018) or some penalties in the objective function for not satisfying them (Cortés et al, 2014). 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 customer. Thus, a central idea that motivates this research is that ignoring cost heterogeneity in the demand might lead to globally suboptimal solutions. Moreover, in this research we not only characterize the effect of having different cost on the demand side, but we move one step further and propose a relatively simple approach to characterize this cost.

Consider for example dispatching problem where technical teams send to different locations to provide service support (Hill, 1992; Wiegel and Cao, 1999). By not considering customer heterogeneity, dispatching solutions can leave unserved customer who are very cost sensitive, leading to Pareto-suboptimal solutions. Thus, homogeneous policies lead to inefficient resource allocation and it will damage customer equity in the long run (Vogel, 2008). Moreover, identifying customer with low cost of postponement, can give more flexibility to the problem generating better solutions. If the demand of service corresponds to internal customers, to achieve global optimality, the cost of the units demanding the service must be balanced with the cost of providing it. Thus, no matter the nature of the relationship between customers and service provider, ignoring opportunity costs in the demand side can lead to suboptimal decisions.

In this research, we consider that customers are heterogeneous in the cost they face and therefore optimal solution should prioritize more sensible customers. While the general idea applies to other settings, in this research we focus on the technician dispatching problem, where customers request equipment repair services. In this problem, customers have an operational cost for not having a machine operating. Notice the cost can not only be associated to the reduced output in the manufacturing process, but also to longer waiting times to complete operational tasks. Thus, difference in costs can arise because customer can give different use to 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 cost. Here we analyze in which extend the optimal solution varies as a function the variability of cost between customers and the relative importance of customer costs with respect to those of 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 grow, it is possible that most customers remain unattended. This is because the model prioritize more expensive customer sacrificing efficiency in the design of traveling routes. Characterizing these solutions is important not only to understand the trade-offs at play, but also provide some guidance about the conditions that make more necessary to account for variations of costs in the demand side.

After characterizing how the optimal solution varies with different distribution of customer costs, we propose a simple Markovian model to capture cost-heterogeneity for the case where cost 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 close expression for the waiting time that we can balance against the time of the technicians.

Our results indicate that when customers are different enough, the number of requirements served can decrease. However, considering that unserved customers are precisely those with lower opportunity costs, the general impact is a sizable cost reduction derived from a gain in flexibility to design optimal routes.

# Literature Review

This research stands in the intersection of two main streams of literature. More specifically, we consider the well-known marketing concept of customer heterogeneity and we apply it in the context of the technician dispatching problem where customers might have different costs by not being served on time. Thus, we first describe the literature on customer heterogeneity and its relation to operational services and then we discuss how customer costs structures has 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 and Bowen, 2004). 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 matter when deciding how to allocate resources. For example, Günes and Aksin (2004) propose a theoretical model where a server provider must distinguish between high and low value generator type of customers. 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 in describing how different prioritization policies affects overall performance (for instance, see Pangburn and Travulaki, 2008 and Afeche and Pavlin, 2016). In general, this stream of literature provide interesting insights characterizing the equilibrium outcome, but unlike this research, they propose very stylized models that are not connected to operational programs. In this 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 et al (2015) analyze a series of properties of dynamic prioritization policies. Similarly, Zhao et al (2016) determine optimal assignments to maximize the long-run throughput considering prioritized customer orders. Closer to our application, Jayamohan and Rajendran (2004) and Tay and Ho (2008) study a series of rules to prioritize jobs associated to 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 this research.

In this research we deal with the technician dispatching problem, which can be considered a variant of the capacited vehicle routing problem (CVRP). Several facets of this problem have been widely studied in the literature, the development of efficient methods to solve real sized instances one of the most prominent directions of research (see for example Gendrau et al. 2002 and Cortes et al, 2014). A special case of CVRP is the *heterogenous* CVRP, where vehicle fleet is characterized by different capacities and costs (Baldacci, Toth and Vigo, 2009). In our investigation, instead of analyzing heterogeneity on the supply side we analyze how different cost in the demand side can have relevant impact on the optimal solution. In this regard, our investigation is related to multi-objective approaches to solve vehicle routing problems (for a review, see Jozefowiez, Semet and Talbi, 2008). Here, the basic idea consist in adding objectives to the traditional cost minimization in order to improve customer satisfaction regarding delivery dates (Sessonboon et al, 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 aversion to delays impact optimal solutions. Moreover, we move one step forward by providing a simply approach to deal with that heterogeneity for the technician dispatching problem.

Marketing literature has described the benefits of evaluating and acting upon customer profitability even 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 and Totzek, 2008). In summary, this 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 endogeneize customer cost heterogeneity using commonly available information such as the utilization rates and number of machines.

# Modeling Approach

Let us assume a set of *n* requests distributed over a modelling area, asking repair service for office machines provided by a set of technicians working on the area, under a 24-in advance service protocol. In Cortés et al, (2014), the daily dispatch of technicians was formulated as a VRP with soft time windows, where the cost had three components: travel cost, a soft time window violation and a penalty for sending customers to the next day. 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 in the requested day. In this research, we have added the feature of heterogeneity in the cost that customers face, which is considered in the computation of the penalty of sending customers to the next day prioritizing those customers that are more sensible to being unattended at the required day. For practical purposes, in the proposed model, the soft time window constraint is left wide open, considering that each customer has no penalty if he(she) is at any time but during the requested day, regardless of the specific time of attention. In this paper, this constraint only set the final time of the labor day for technicians’ shifts. This apparent simplification allows us to clearly identify the impact of different penalties for postponing the attention of customers to the next day. Also, in the original GC model the number of technicians is not limited, instead every new route has a base cost that has to be paid. In our formulation, the firm must find the best way to use her available resources, therefore we added a hard constraint on the number of technicians.

In the next subsection we describe the technician dispatch model and the way it is modelled. Then, we describe how the vehicle routing model is combined with different ways to model cost heterogeneity by means of the penalty of sending requests to the next day, which in this research is differentiated depending on the cost faced by each customer. The context behind the modelling ideas require to run simulations (next section) over a longer period than a single day.

## The Technician Dispatching Model

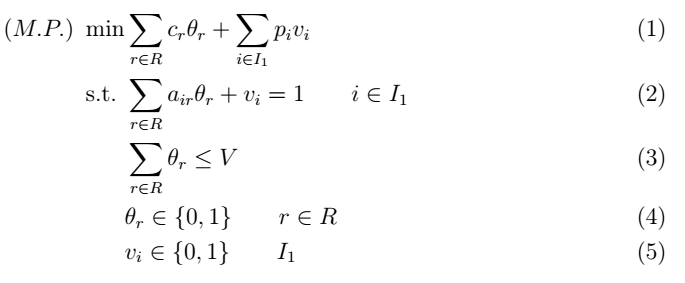
The technician dispatch model we consider in this research is based on the branch and price scheme proposed in Cortés et al. (2014). We use the same data from a major company offering repair services of office machines in Santiago, Chile. For the daily dispatching model, the set of service requests assigned during a given day come from the previous days, usually the day before, since the company attempts to enforce a 24-hour service policy. The dispatcher selects which requests should be handled during the coming day, regardless of the time of attention during the day, unlike Cortés et al. (2014), in which depending on the time of attention there might be cost associated to a delay. Technicians will start his working day at one of a set of high-priority customer locations, which correspond to those customers pending from the day before. In some cases, a technician can start the working day at the depot if high priority clients are either zero or assigned to other technicians. In addition, all the tasks assigned to each technician must be completed during the day.

To ensure feasibility, Cortés et al. (2014) modified the classical VRPSTW formulation to be able to decide which service requests can be handled during the day by the available technicians; any remaining requests are postponed to the next day and given a very high priority. Considering that most VRPSTW are quite difficult to resolve using traditional network flow models (Taillard, Badeau, Gendreau, Guertin, and Potvin, 1997), in this research, we formulate the problem in the same way Cortés et al (2014) did, through a column generation approach,.

In general, column generations (CG) approaches have been very successful in solving various types of VRPs. CG approaches allow splitting the problem into two parts: a main model, which chooses the routes with the minimum total cost from a pool of feasible routes; and a secondary model, the subproblem, which generates feasible routes that could potentially reduce the total cost. In the literature, the subproblem is usually solved via dynamic programming (DP). Cortés et al., (2014) decided to use Constraint Programming (CP) instead, for several reasons. First, it seems that CP works well when the length of each route is relatively small, which is the case for this application. Second, the CP model we implement is simple and easy to code and many additional constraints can be incorporated directly into the code without a special modeling technique.

**Master Problem**

The master problem of the VRPTW in Cortés et al (2014) 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 . Each route 𝑟 ∈ 𝑅 is characterized by a technician who starts a path at a specific location , and then continues to visit a sequence of locations , where is the set of all customers to be served during a specific day. Each route, therefore, is described by the set r ∈ {2, ..., 𝑒} where 𝑒 is the last service request of a path. Therefore, the mathematical statement of the master problem was formulated as:

 \begin{align}

\text(M.P.) \ \min & \sum\_{r \in R} c\_r \theta\_r + \sum\_{i \in I\_1} p\_i v\_i \\

\text{s.t.} \ & \sum\_{r \in R} a\_{ir}\theta\_r + v\_i = 1 \qquad i \in I\_1 \\

& \sum\_{r \in R} \theta\_r \leq V \\

& \theta\_r \in \{0,1\} \qquad r \in R \\

& v\_i \in \{0,1\} \qquad I\_1

\end{align}

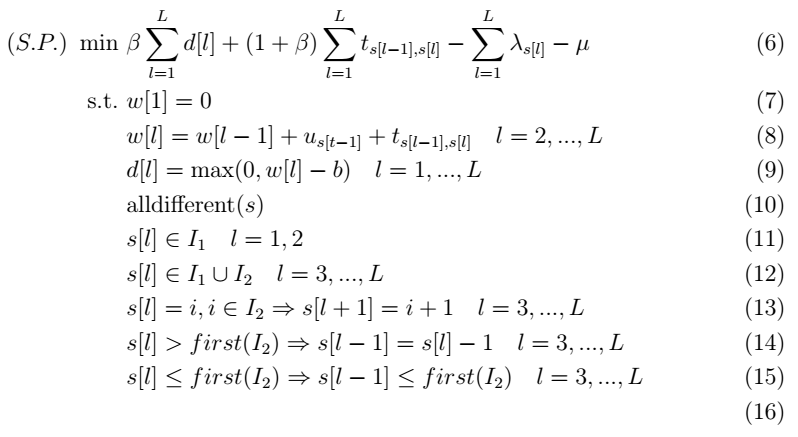
This mathematical formulation considers two binary variables that indicates whether the route should be chosen or not, and variable that is equal to one if client is not included in any of the chosen routes, The objective function is the sum of the cost of selecting a route c\_r and the penalty for postponing a client to the following day p\_i. In the first constraint the binary parameter indicates if customer belongs to route , so this ensures that all clients are either in a selected route or moved to the next day. In the constraint (2) the total number of routes that can be selected is limited by v\_i the number of technicians. It should be noted that this formulation guarantees that there is always a feasible solution to the problem, even if the set R is empty the solution where no client is served is possible .

**Sub-Problems**

The objective of the CG subproblem 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 cost of a column is defined as the cost of the route , minus the sum of the master problems dual variables: associated with constraint (1) and associated with constraint (2). The cost of 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 … L an array of variables that represents a route, where the lth element of s[l] is the client in the position l of it. The first to positions of all new routes must be a client from set I\_1 (see forward for the implementation details and starting routes), from which the technician continues the route towards either another client from set I\_1 or a fictional client from set I\_2. These fictional clients are added for CP modelling purpose and ensures that all routes are of length L. (Cortés et al, 2014). The maximum number of fictitious nodes that can be included in a route is , therefore the set of fictitious nodes is defined as , where is the cardinality of set . Additionally, we have defined variables , and for the service start time at the lth client, the violation of the time window and the travel time between two clients respectively. A small difference from the original model is that we are starting all routes from clients and not from depots.

The sub-problem used for the generation of feasible routes is presented below.



\begin{align}

\text(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 \lambda\_{s[l]} - \mu \\

\text{s.t.} \ & w[1] = 0 \\

& w[l] = w[l-1] + u\_{s[t-1]} + t\_{s[l-1],s[l]} \quad l = 2,...,L \\

& d[l] = \max(0,w[l] - b) \quad l = 1,...,L \\

& \text{alldifferent}(s) \\

& s[l] \in I\_1 \quad l = 1,2 \\

& s[l] \in I\_1 \cup I\_2 \quad l =3,...,L \\

& s[l]=i, i \in I\_2 \Rightarrow s[l+1] = i+1 \quad l=3,...,L \\

& s[l] > first(I\_2) \Rightarrow s[l-1] = s[l]-1 \quad l=3,...,L\\

& s[l] \leq first(I\_2) \Rightarrow s[l-1] \leq first(I\_2) \quad l=3,...,L\\

\end{align}

Constraints (7) and (8) calculates the starting time of the service at the lth client in the route, where u[l] is the required service time in the node. In (9) we define the violation of the time windows, which in this case is equal to for all customers. Constraints (10) to (13) builds the routes, forcing that the first two nodes must be clients, and once the route reaches a fictional node it can only go to another fictional node until the end of the route. Constraints (14) and (15) are redundant constraints that greatly improve the CP resolution as shown in Cortes et. al. (2014)

**Implementation**

In order to solve this particular set of experiments our implementation differs from the one in Cortes et al. (2014) in several ways. As mentioned, we consider much wider time windows and equal to all clients, the reason for this is that we want to capture the effect of postponing a client’s attention and therefore we don’t want to include a second cost specific for each client. Also, we added a hard constraint in the number of technicians in the MP and the corresponding dual variable in the SP. For solving the problem, we opted to only generate columns in the root mode and not use a branch-and-price algorithm. The reason is that as the found in Cortes et al. (2014) the branch-and-price algorithm only slightly improved the branch-and-bound solution for this problem.

Furthermore, as we decided to use a single dataset of clients for one week of operation for testing this approach (see computational experiments) we opted to not implement all the time improving methods proposed. Instead, we a) manually added all the single client routes as starting routes for R and b) we computed a large set of starting columns to further reduce the GC search.

The master problem of formulation proposed above considers a penalty *pi* for not serving a request during the working day. A key idea of this investigation 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 services of customers with higher costs of not being served, but also because it could provide better operational flexibility. We expect the impact of the solution to be dependent on several key parameters on the customer costs and therefore we analyze a series of scenarios in which we vary the relative importance of customers’ costs and degree of heterogeneity among them.

In practice, the penalty of not complying with a requirement does not distribute homogeneously. Hence, we have characterized this value considering that the cost is different for each costumer 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. To do so, we assume that customer costs *pi* is known by the firm, and we run different scenarios about 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 cost is not known by then firm, but they 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 cost of customers waiting for the service can be approximated through a closed form equation derived from an Erlang-C type of models (or MMC, REF). Now, we revise how we address these two research objectives.

## The impact of cost heterogeneity in optimal dispatching

We start by assuming that the firm knows the magnitude of the penalties associated to not satisfying the request on time 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 determine a penalty that closely matches customer costs. Here 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 on the penalties *pi*, assuming that log(*pi*) is normally distributed. The Normal distribution is not only a frequent choice to characterize randomness (Steward and Goldenm 1982, Abdelaziz, Aouni, and El Fayedh, 2007), but also is easy to interpret. 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 *pi* according to that normal distribution. To guarantee that changes of the optimal solutions are only explained by variations in mean and variance of the penalty cost, in our sampling procedure we keep the ordering in costs. This is, if the penalty of a customer *a* is larger than the penalty 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* of each value of *m*. The values of the means were selected in order to represent the average penalty moves around the cost of receiving the service approximately ten hours outside the time window associated with each requirement. In this way μ -value was determined using the delay cost per hour contemplated in the routing model. That base scenario is designed to capture the average penalty cost currently faced by the company. To evaluate how different values of *m* affect the performance of the system, we also considered lower and higher values for *m.*. The values of the variances were also selected to represent a wide spectrum of values ranging from almost no variation to scenarios where the maximum penalty is twenty five times the minimum. In these twenty-one scenarios we consider that the penalties are independent of the complexity of the repair tasks. In reality, we consider plausible they might be positively correlated and therefore we also explore another set of twenty one scenarios where longer task are associated to larger penalties. While different degrees of correlation can be explores, for simplicity we only consider the case with a perfect ordering. In this case, penalties are ordered according to the length of the repair task. Figure 1 shows example of the distribution of penalties for relatively small and relatively large values of the variance *s*. Figure 2 illustrate the correlations between service time and delay penalties for the cases of no correlation and perfect ordering.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

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

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

**Figure 2:** Correlation between Expected Requests Service Time and Penalty. Left panel (a) displays the case of no correlation between them. Right panel (b) displays the case of a perfect ordering.

## A Markovian model to characterize costs on the consumer side

Oftentimes, firms do not have precise assessment of customer costs and they must approximate the cost of the delay for their customers 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 cost of not being served. We believe 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 at the customer level. Here, the total cost of customers waiting for the service can be approximated through a closed form equation derived from an Erlang-C type of models (or MMC, REF pending).

Our main objective of this exercise is to use specific customer information to approximate a penalty representing customer cost of being delayed. For that purpose, we assume that each customer has *i* has *ni* machines and each one of them process their assignment jobs in an exponentially distributed service time of mean 1/*μ*. 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 is homogeneous. We also assume the workload is given by a Poisson arrival process with rate λ. Under these assumptions, the internal use of the machines can be described as bird-death process characterized by a utilization rate ρ*i=λ/*(ci μ), that represents the average fraction of time that each machine is being used.

When a machine is broken and it requires to be repaired, the system reduces its capacity from *ni* to *ni* -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 (*Ws*) has close form expression given by equation (X)

|  |  |  |
| --- | --- | --- |
|  |  | (X) |

Here, correspond to the probability that an incoming job must wait because the servers are busy. This probability also has a close form solution as is shown in Equation (Y).

|  |  |  |
| --- | --- | --- |
|  |  | (Y) |

To derive a monetary cost of not being served, we assume that such a cost is proportional to the additional time in the system given by the difference between *Ws*(*n-1*)-*Ws*(*n*). Notice that this cost depends on the number of machines and the utilization rate, which is information readily available for 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. To understand the intuition of our model, Figure 3 displays XXX

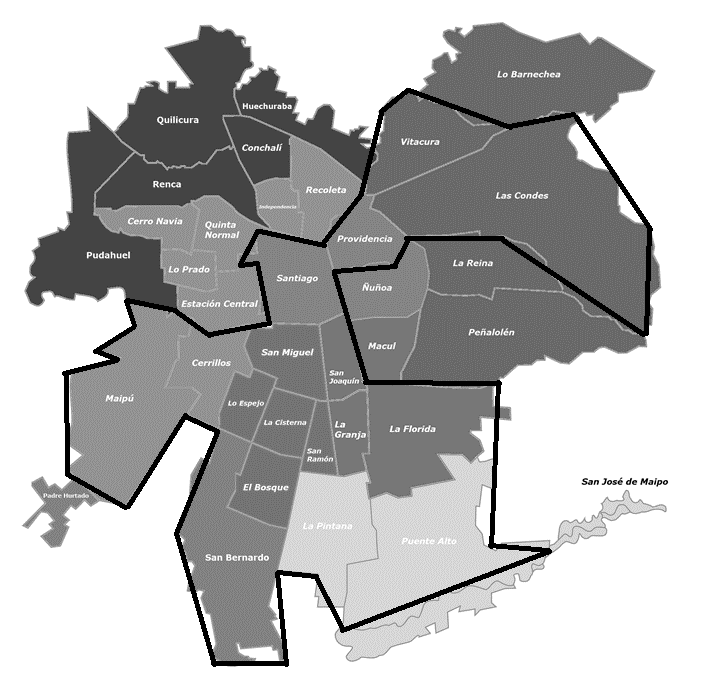
|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

Figure X. (a) Ws and Damage / Number of Servers (Constant Utilization Rate) (b) Ws and Damage / Utilization Level (Constant Server Number)

SHORT EXPLANATION OF FIGURE.

1. **Results**

To characterize the effect of including cost heterogeneity in the dispatching problem, we have made a series of computational exercises where we characterize how the optimal solutions change when customer costs are incorporated. In this exercises, we use real requirement data from Santiago, Chile. Each daily instance consists of a set of service calls that where solved using the routing model described above. To meet daily demand, the firm has subdivided the city into two main service areas from where the vehicles are dispatched with technical personnel. Both zones are highlighted within the polygonal line in the following figure.



**East**

**Zone (A)**

**South**

**Zone (B)**

(Taken from Leng’s)

Daily demand characterization is important since its variability affects the response time the firm is able to provide to its customers. During high demand days, we observe that the percentage of customers who are not satisfied on time increases. Along with this, service requests are not concentrated homogeneously throughout the week. For the most part, the strongest demand tends to concentrate at the beginning of each week and decreases throughout each day.

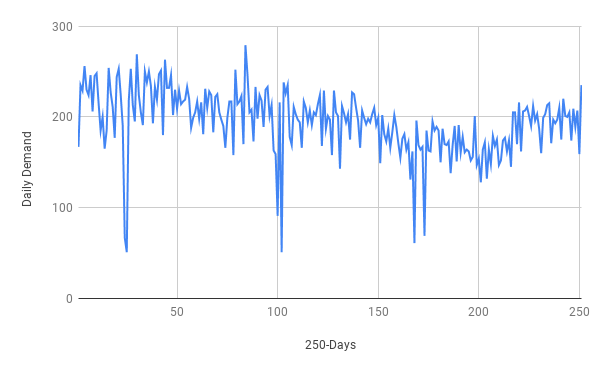
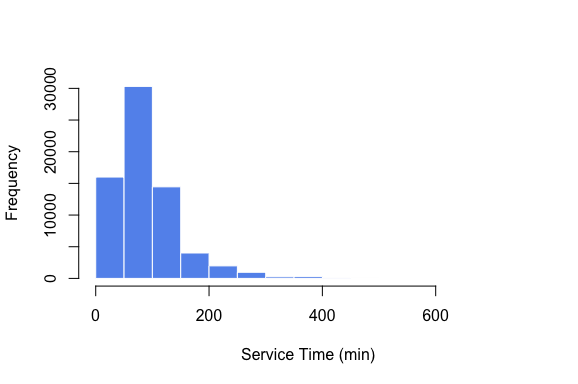
 

Figure X. Daily Demand Variability and Service Time

On average the firm receive X repair request a day being Monday the day of the week with highest demand. Repair task differ in their complexity and some of them take longer than other ranging from X to Y minutes to be completed. To server daily requirements we considered a fixed float of X vehicles. ADD QUE EXISTEN SECTORES

All instances we use in our numerical evaluations correspond to a full labor week and for every penalty scenario, we solved each weekday using the vehicle routing model previously described. During the resolution of each day an identification is used for each customer that requests the service during a weekday, the sector of the city to which it belongs, the expected time in hours that it will take the resolution the technical incident, and a cost associated with the non-resolution of the requirement during the corresponding day. For the resolution of each instance we have assumed that the number of resources of the firm to satisfy the demand has been previously calculated and therefore it is operating with a balanced and fixed number of vehicles and technicians. The following tables show the instance demand and the average service time recorded for each zone A subsector.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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 |
| **Total** | **100** | **90** | **80** | **70** | **60** |

Table 1. Instance Demand by Weekday and Sector (Zone A)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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 |
| **Average** | **105,92** | **90,56** | **72,74** | **98,65** | **80,83** |

Table 1. Instance Average Service Time by Weekday and Sector (Zone A)

The main operating costs of the firm studied 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 cost per distance traveled equal to CLP $ 3,300 per hour and a Delay Cost equal to CLP $ 4,800 per hour. These costs are referential and do not represent real values provided by the service firm.

## Evaluation of Heterogeneity of Penalty Costs

Distributional model used to characterize customer’s cost assume that there is a probability distribution that characterizes penalty associated with the set of requirements and said distribution is known by the service provider. We simulate the resolution of multiple service scenarios conditional to which the planner knows the value of penalty associated with each customer. This exercise allows obtaining average global service indicators that help us to evaluate whether the firm should worry about customer heterogeneity or not. As explained before, the system composed by the firm and its customers has been characterized assuming a probability distribution and establishing assumptions about its parameters allow us to describe how different customers are from each other. As was explain above, we supposed that customer’s penalization follows a normal distribution and it was our special interest to observe the behavior of the system when service operation is exposed to requirements with different degrees of dispersion among customer’s penalty. For this we set different values of σ on each scenario. The referred service simulation was carried out assigning all customers a different parameter Pi.

The increase in distribution dispersion does not affects customer’s penalty average. By positively varying σ, the average penalty only increases in high variance scenarios This means that the difference between each scenario is given only by the variation of the sigma-parameter. The resolution of the generated instances shows that the assignment of a heterogeneous value to the penalty, allows to separate customers into two groups: One composed of those priority customers that receives service immediately and a second group composed of those that must be assisted with delay. The following boxplots show that the routing algorithm chooses to systematically go for customers with high penalty values to include them in service routes and choose to make the service more flexible to customers with low penalty.

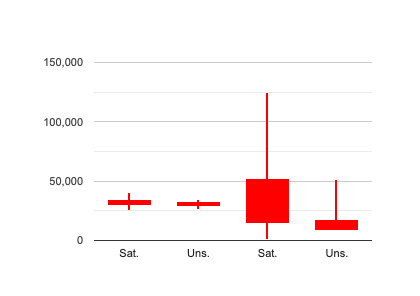
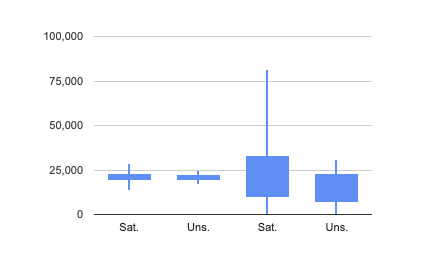
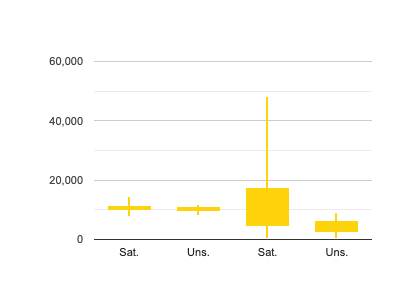


Figure X. Penalty Boxplots Satisfied / Unsatisfied Costumers

(Low / High Variance)

This figure shows that by introducing a penalty value that distributes heterogeneously, the firm can generate a separation between two groups of costumers, being possible to find a significant difference among them at the end of the week. In this case, the model consistently chooses to serve customers with higher costs of not receiving care, above those of lower cost. As the dispersion of the studied penalty increases, the separation between groups also increases and costumers that are not served tends to be more homogeneous. It is also possible to observe that the percentage of costumers unsatisfied during the week tends to increase as the difference between customers increases. Our service scenarios show how the percentage of requirements changes as the dispersion of the distribution increases. At the same time, we consider an instance where the expected service time of the clients is perfectly correlated with the penalty of the customers. In this way, the most difficult locations are those with the highest penalty.

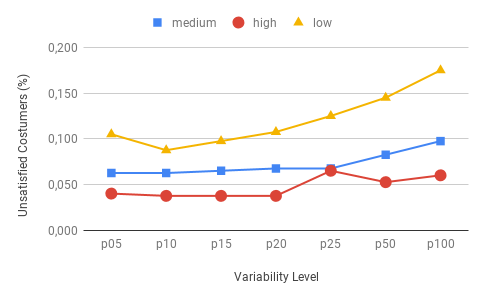


Figure X. % Unsatisfied costumers (1-week)

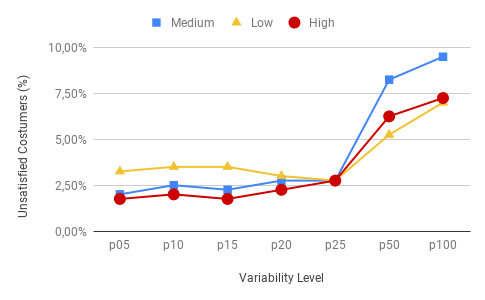


Figure X. % Unsatisfied costumers (1-week) (Service Time and Penalty Correlated)

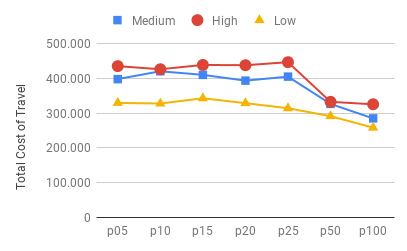


Figure X. Traveling Cost (1-week) (Uncorrelated)

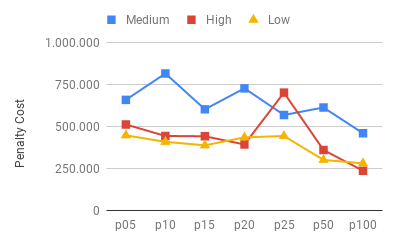


Figure X. Total Cost (1-week) (Uncorrelated)

(Comments on this graphics)

* 1. **Markovian model**

Markovian Model introduces specific customers information such as prints terminal and server utilization to approximate a penalty value associated with each service request. The following figures seek to represent the same service metrics characterized in the previous section, but this time we seek to evaluate the performance of the system in relation to variables that the service can easily get. Following the same scheme use in previous section, we first described the system under a random combination of print terminals and utilization level. Then we added an additional scenarios where longer average service time are associated with a higher number of servers and utilization levels. The following graphics shows how the routing algorithm tends to serve first customers with less replacement capacity among from those costumers that are less sensitive to the failure of their equipment.

Figure X. Relation between (N and Rho) Color: Att vs NAT (Uncorrelated / Correlated)

(comments on this graphics)

# Conclusion

In this research we analyze the impact in optimal dispatching of having customers with different cost of delay in the service repair. Previous literature in general has assumed these costs to be given and it has remained mostly silent in providing a systematic evaluation of these impacts. In our analysis we compare key performance metrics of optimal dispatching solutions for a number of scenarios, which lead us to several conclusions.

1. It matters and it can lead to sizable cost reductions.
2. Ceteris paribus, more variation is better because implies a larger fraction of customer that are not sensible to delays. That gives more degrees of freedom to find more cost-effective solutions. Explain mechanism some complicated customers can be delayed
3. As a consequence of larger degrees of freedom we did expect that transportation costs go down. Considering that number of delays increases we expected that total cost of compensating for delays would also increase, but that is not the case. Intuition: first couple of days the system has more request than capacity and therefore they optimal solutions would require to delay always. Variability provides the flexibility of delaying only those with smaller costs.
4. Mean values of the penalties do matter and can moderate the operational gains of exploiting cost heterogeneity. If they are very large they leave less room to improve

More specifically, we explore how the distribution of those costs affects optimal solutions and then we propose a simple Markovian model to capture cost-heterogeneity for the case where cost of failure can be traced to observable operational characteristics. Our results indicate that when customers are different enough, transportation and total penalty costs decreases implying a sizable gain in operational efficiency. Moreover, results from the Markovian model indicate that firms can take advantage of these operational gains even with only a few customer characteristics are observed.

Dos ideas para concluir:

* A aumentar la heterogeneidad de los clientes, la planificación sea más flexible y por ende podamos mostrar que al incluir el costo-heterogeneidad la firma experimente un incremento operacional.
* El modelo markoviano sirve para darle prioridad a los clientes con menor capacidad de reemplazo, en el caso aplicado son aquellos que son más sensibles a la falla en una maquina.

In this research we focused on the technician dispatching problem and we demonstrate that heterogeneity in the cost of the delay can have a relevant impact on operational efficiency. We believe the general idea can be extended to other classes of vehicle routing problems.

Used for a given information set. Furthermore, the results of this research should motivate practitioners in the domain of transportation planning to better learn from their customers. While the recent advances in data analytics suggest that learning from customer should be accessible to firms, results derived from this investigation indicate that such learning can have a direct impact on operational planning.

# 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 (2018). OR Spectrum

Baldacci, R., Toth, P. & Vigo, D. (2010) Exact algorithms for routing problems under vehicle capacity constraints. Ann Oper Res, 175-213

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, 1–6.

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

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. In The vehicle routing problem (pp. 129-154). Society for Industrial and Applied Mathematics.

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.

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.

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. In: Desaulniers G., Desrosiers J., Solomon M.M. (eds) Column Generation. Springer, Boston, MA

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.

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.

Stewart Jr, W. R., & Golden, B. L. (1983). Stochastic vehicle routing: A comprehensive approach. European Journal of Operational Research, 14(4), 371-385.

Stringfellow, A., Nie, W., & Bowen, D. E. (2004). CRM: Profiting from understanding customer needs. Business Horizons, 47(5), 45-52.

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.

Wind, Y. (1978). Issues and advances in segmentation research. Journal of marketing research, 15(3), 317-337.

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.