Improving Maintenance Service Delivery Through Data and Skill-Based Task Allocation

Roberto Sala, Fabiana Pirola, Giuditta Pezzotta, Mariangela Vernieri

Improving maintenance service delivery through data and skill-based task allocation

Roberto Sala[1[0000-0001-7671-6927]], Fabiana Pirola[1[0000-0003-4633-7568]], Giuditta Pezzotta[1[0000-0001-7983-8356]], Mariangela Vernieri[2]

[1] University of Bergamo, Department of Management, Information and Production Engineering, Viale Marconi, 5, Dalmine (BG), 24044, Italy
roberto.sala@unibg.it
[2] Balance Systems, Via Ruffilli, 8/10, Pessano con Bornago (MI), 20060, Italy

Abstract. Maintenance service delivery constitutes one of the most problematic tasks for companies offering such service. Besides dealing with customers expecting to be served as soon as possible, companies must consider the penalties they are incurring if the service is delivered later than the deadline, especially if the service suppliers want to establish long and lasting relationships with customers. Despite being advisable to use appropriate tools to schedule such activity, in many companies, planners rely only on simple tools (e.g., Excel sheets) to schedule maintenance interventions. Frequently, this results in a suboptimal allocation of the interventions, which causes customer satisfaction problems. This paper, contextualised in the Balance Systems case study, proposes an optimisation model that can be used by planners to perform the intervention allocation. The optimisation model has been developed in the context of the Dual-perspective, Data-based, Decision-making process for Maintenance service delivery (D3M) framework, which aims to improve the maintenance service delivery by making a proper use of real-time and historical data related to the asset status and the service resources available. The proposed model tries to cope with the current problems present in the company's service delivery process by proposing the introduction of a mathematical instrument in support of the planner. Being strongly influenced by the contextual setting, the model discussed in this paper originates from the D3M framework logic and is adapted to the company necessities.

Keywords: Maintenance, Product-Service Systems, decision-making, task allocation.

1 Introduction

The productivity of the industrial assets depends on how production and maintenance schedules complement themselves [1]. Production constraints, established maintenance policies, and unexpected failures significantly impact the company schedule [2] and, in turn, the definition of the maintenance calendars. The definition of effective maintenance schedules, or the capacity to intervene as soon as the failure presents itself, becomes even more important in the Product-Service System (PSS) context [3], where maintenance is offered as a service by a supplier to a customer. From the maintenance supplier perspective (the one assumed in this work), knowledge and experience guide not only the definition of the maintenance policies (in agreement with the customer) but also the operational decisions (e.g., the allocation of the maintenance intervention to the technician), which are rarely executed using supporting tools (e.g., software) as guidance [4] despite the availability of data collectable from the field [5]. Even in an Industry 4.0 context, data gathering and analysis are not fully exploited if decision-makers rely on their experience instead of field data to make decisions [6]. Despite the availability of maintenance and scheduling software on the market, their adoption is obvious for various reasons. For instance, companies may be used to carry out these activities without their support or may consider too expensive the costs for the adoption of a software. On the other hand, companies may not be so aware of the potentialities and benefits that the introduction of such software in their processes may lead to and, thus, may decide to avoid such an investment.

Maintenance delivery is a complex process encompassing a series of activities and relations between decision-makers and should guarantee a satisfactory result for the customer and the supplier [7]. Especially in the PSS context, where the access to the asset operational data depends upon the PSS typology, the contractual and data sharing agreements between the stakeholders, it is necessary to be reactive and make suitable decisions to tackle in a short time the assets problems. To do so, it is necessary to make use of all the asset and service data available and evaluate the intervention alternatives matching the requests with the available solutions, figuring out the resolution scenario that maximises the utility for the customer and the provider [8]. Despite this, some authors identified a lack in terms of approaches able to manage, process and match data and information, especially as far as the service characteristics are considered [9]. A possible solution consists of adopting optimisation models that can improve the intervention allocation according to various objective functions and constraints. A literature review on the topic showed how the adoption of optimisation is becoming increasingly important [10] and requires additional research based on case studies. This paper is settled into this research stream, where an optimisation model (part of a wider framework for the improvement of the maintenance service delivery) is proposed in response to the willingness of a company to improve the maintenance delivery process in terms of management and decision-making. Considering a literature perspective, researchers proposed various resolution approaches focused on different objectives such as costs minimization [11], profit maximisation [12], optimisation of the resolution teams [13] and others. This paper proposes an approach that has as objective the minimization of the number of late (i.e., tardy) interventions justified by the idea that the maintenance supplier has to pay a penalty for every tardy intervention.

The authors [14] developed a Dual-perspective, Data-based, Decision-making process for Maintenance service delivery (D3M) framework to exploit at best the possibilities offered by data-based decision-making that rely on the asset and service data. The D3M framework uses the asset and service data to support the service planner in organising the service delivery process by allocating the intervention requests to service technicians. Maintenance execution decisions should be adapted to the assets requiring it as well as to other factors such as supplier constraints and customer requirements and characteristics (e.g., location, availability of maintenance workforce, contractual constraints, SLA). The development of the optimisation model discussed in the following sections of the paper should be seen as a part of the development of the D3M framework, in the scope of ameliorating the interventions allocation based on the data collected and processed during the D3M framework application. From a practical perspective, the model presented in this paper is strongly influenced by the collaboration with an Italian manufacturing company and, thus, it is also shaped based on the company necessities.

This paper deals with the proposal of an optimisation model, part of the D3M framework, able to handle asset and service data to support decision-making related to the maintenance service delivery. Section 2 describes the model formulation, while section 3 focuses on its application. Eventually, Section 4 and 5 discuss the analysis results and concludes the paper delineating future research.

2 Problem description and model formulation

2.1 Problem description

Maintenance requests are generated by customers and shared with the maintenance supplier. Each maintenance request is evaluated by the provider who selects the resolution approach. Each request can be fulfilled in one or more manners, each one requiring different skills, execution times, and costs. The maintenance supplier can satisfy the customers' requests in different ways depending on the context: remote support, send spare parts to customer premise, return the failed part to the provider premise, send technician to customer premise.

Every request *R* has its characteristics (e.g., failure typology, gravity, skills required to solve the problem), which determines if it can be solved in one or more ways. Not all the resolution modes are similar in terms of competencies required (from now on referred to as skills), execution times (varying depending on the executor's experience), and costs. Therefore, the planner must match the request with the proper resolution mode considering all these factors trying to maximise customer satisfaction and minimise the resolution's times and costs. In the model, expenditures and costs are linked to the execution of interventions happening after a specific due date, which can be defined by the forecasted failure of a component or defined by Service Level Agreements (SLA) contracts. The assumptions are: (i) the other costs related to the intervention are covered by the service contract established between the stakeholders, and (ii) the maintenance supplier wants to minimise the costs arising when the suppliers fail in executing the intervention before the due date.

The model aims to be the bridge between asset- and service-related information. Thus, the model wants to use all the available information source (e.g., maintenance requests, resolution modes, technicians' skills, and calendar) to optimise maintenance service delivery while minimising the number of tardy interventions. The originality of the model is in merging real-time (e.g., RUL, technicians' calendar, position, and availability) and historical data (e.g., customer information, skills of the technicians) coming from different sources (asset and service-related) and guided in their collection and use by the D3M framework structure, to identify the best resolution modes and executors considering the customer and suppliers constraints and interests. The inclusion of real-time data such as the technicians' and resources availability, as well as the contextual conditions (e.g., failure typology, customer information and history) would allow for an improved allocation of the interventions as well as for the selection of the solution the best fits the failure characteristics, the resolution strategies, and the companies' constraints.

The notation in the following is used to model the problem discussed hereabove:

- *R*: the set of intervention requests.
- M_r: the set of modes that can be used to fulfil the intervention request $r \in R$.
- *T*: the set of available technicians.
- *S*: the set of skills required by mode $m \in M_r$
- *W*: the set of windows available for each technician. Each window delimitates the period where the technician is available to execute the intervention.

Each intervention request $r \in R$ defines a set of modes M_r that can be used to execute the intervention and fulfil the request. Each technician $t \in T$ has a set of intervention request $r \in R$ already assigned before the next planning. Thus, before the planning, technicians already have some free and busy spots in their calendar. The availability windows can be described through a start date s_w, and an end date e_w, which can be used to define the length $\theta_w = e_w - s_w$ of the window $w \in W$. There is an infinity window available for each technician, which is the one just after the last busy block in the calendar.

The problem that wants to be modelled consists in assigning an intervention request $r \in R$ to be executed in a specific mode $m \in M_r$ to a specific window $w \in W$, which is associated with a single technician $t \in T$ simultaneously minimising the number of tardy jobs executed. Other assumptions characterising the model are:

- Each technician $t \in T$ owns a set of skills that define its competencies and the intervention modes they can execute. Skillsets influence the technician's ability to execute deal with certain requests, resolution typologies (e.g., on-field vs remote) and execution length.
- At the moment of the intervention allocation, the schedule of the technicians is not blank. There are windows of availability and unavailability for all the technicians.
- Each time an on-field intervention is performed, the technician leaves from (and returns to) the headquarter before executing the following one.

The parameters used in the model are the following:

- δ_{rms}: binary, representing the requirements of each mode in terms of skills. In particular:

$$\delta_{rms} = \begin{cases} 1 \; if \; the \; mode \; m \in M_r \; requires \; the \; skill \; s \in S \\ 0 \; otherwise \end{cases}$$

- ω_{mts}: binary, associating the skills of each technician with the mode of resolution. In particular:

$$\omega_{mts} = \begin{cases} 1 \; if \; the \; technician \; t \in T \; has \; the \; skill \; s \in S \\ required \; for \; the \; mode \; m \in M_r \\ 0 \; otherwise \end{cases}$$

- M_{rm}: number of skills required to perform the intervention that satisfied the request $r \in R$ with the mode $m \in M_r$ so that $M_{rm} = \sum_{s \in S} \delta_{rms}$.
- DD_r: due date of intervention request r, defined as min $\{SLA_r; RUL_r\}$, where SLA_r is the date before which the request r has to be executed according to the SLA stipulated between the stakeholders and RUL_r is the residual life before the breakdown of the component associated with the request $r \in R$.
- t_{rm}^{TOP}: travelling time for the technician to reach (and come back from) the location of the intervention request r addressed with the mode $m \in M_r$. This value is not dependent on the single technician $t \in T$.
- t_r^{SS}: time to get the spare parts in place for the execution of the intervention in mode $m \in M_r$ fulfilling the request $r \in R$. This time is dependent upon the intervention request r, which determines the necessity of spare parts.
- t_{rmt}^{INT}: the time required to perform the intervention in mode $m \in M_r$ by the technician $t \in T$;
- M: a constant, large number for modelling purpose.

Finally, the following variables:

- C_r= completion time of intervention $r \in R$;
- $U_r = \begin{cases} 1 \; if \; the \; request \; r \in R \; is \; satisfied \; after \; the \; due \; data \; DD_R \\ 0 \; otherwise \end{cases}$ = tardiness
- $x_{rmtw} = \begin{cases} 1 \; if \; request \; r \in R \; is \; satisfied \; in \; mode \; m \in M_r \\ by \; operator \; t \in T \; in \; windows \; w \in W \\ 0 \; otherwise \end{cases}$

Each request must be allocated to a single technician, and all the tasks have to be allocated.

2.2 Model formulation

The model formulated is described by equations from (1) to (11). In particular, the objective function (1) minimises the number of tardy interventions. Such an objective is relevant in the considered case because it minimises the number of requests that should be re-negotiated with the customers (i.e., if the planning returns a tardy job, it may be possible to renegotiate the terms of that job with the customer). Constraint set (2) stipulates that each intervention request is allocated precisely once, whereas constraint set (3) guarantees that each window contains at most one task. Constraint sets (4), (5), and (6) define the completion time of the intervention executed through the mode $m \in Mr$, assuring that the intervention starts after the beginning of the availability window and concludes before its end, leaving the technician the time to travel back to the headquarter. Constraint set (7) defines the match between technician and intervention mode based on the skills required (by the mode) and owned (by the technician). Constraint set (8) introduces the decision variable U_r that assumes the value $Ur=1$ only if the intervention is tardy, which means that completion time $Cr \geq DDr$, otherwise $Ur=0$, which means that the intervention satisfied the condition of being completed before the due date ($Cr \leq DDr$). Constraint sets from (9) and (11) define the domains of the variables.

$$\min Z = \sum_{r \in R} U_r \tag{1}$$

$$\sum_{m \in M_r, t \in T,, w \in W : \theta_w \geq \min(t_{rmt}^{INT})} x_{rmtw} = 1 \ \forall r \ in \ R \tag{2}$$

$$\sum_{r \in R, m \in M_r, t \in T} x_{rmtw} \leq 1 \ \forall w \in W \tag{3}$$

$$C_r \geq \sum_{m \in M_r, t \in T, w \in W} (s_w + t_{rmt}^{INT} + \max(t_{rm}^{TOP}; t^{SS})) \ \ \forall r \in R \tag{4}$$

$$\theta_w \geq (t_{rmt}^{INT} + t_{rm}^{TOP} * 2) * x_{rmtw} \ \forall r \in R, m \in M_r, t \in T, w \in W \tag{5}$$

$$C_r \leq \sum_{m \in M_R, t \in T, w \in W} (e_w - t_{rm}^{TOP}) \ \ \forall r \in R \tag{6}$$

$$x_{rmtw} * M_{rm} \leq \sum_{s \in S} \delta_{rms} * \omega_{mts} \ \forall r \in R, m \in M_r, t \in T, w \in W \tag{7}$$

$$C_r \leq DD_r + M * U_r \ \forall r \in R \tag{8}$$

$$U_r \in \{0,1\} \ \forall r \in R \tag{9}$$

$$x_{rmtw} \in \{0,1\} \ \forall r \in R, m \in M_r, t \in T, w \in W \tag{10}$$

$$C_r \geq 0 \forall r \in R \tag{11}$$

3 Model application

3.1 Balance Systems

Due to space constraints, a complete description of the model application cannot be reported. For this reason, in the following, a summary of three applications run using historical data of Balance Systems (BS) is reported. BS is a manufacturing company headquartered in Italy that sells balancing machines and offers maintenance services to customers who buy the company products. The company manages the scheduling of the maintenance requests through a Microsoft Excel® sheet that is filled by the planner with the support of service technicians. Due to how this process is managed, the scheduling activity is time-consuming. The company has five service technicians, two of them execute only remote support, while the remaining carry out on-field activities.

An analysis of more than 150 maintenance reports allowed to summarise the information useful to run the model (e.g., request, technician, customer, travel time, execution time, spare parts used). Data has been used to create scenarios that differentiate themselves for the number of requests. The data used in the three scenarios have been randomly generated based on the historical data available from BS. Each scenario aimed to verify that the model could guarantee the minimisation of the number of tardy interventions while demonstrating that it would be possible to introduce such an instrument in the planning process to support the planner in his activity.

This information has been used to feed the optimisation model that has been modelled in Cplex12 and solved using an Intel® Core™ i5-7200 CPU @ 2.50 GHz, 2 cores. Table 1 reports the results of the application in three different scenarios. The number of requests in each scenario has been established following a discussion with the company, interested in verifying the applicability of the model under different circumstances. The number of requests in each scenario has been calibrated based on the average number of requests that the company is usually responding to in a normal situation. Variations to the number of requests allowed to test the behaviour of the model in various scenarios. Following, the applicability of the model in different (more complex) scenario is foreseen in the future also to investigate the model limitations in the of variables' handling capabilities.

Each of the requests has been associated with one or more resolution modes, depending on its characteristics. The resolution mode "Send spare parts" has not been modelled, since it does not occupy the technicians' time. Once defined all the inputs, these have been used to run the model and verify its performance. The model has been evaluated according to running time and the capability of minimising the number of tardy interventions.

Table 1. Summary of model application in three scenarios

Scenario	Request	Available Technicians	Resolution Time	Tardy Interventions
1	7	5	5-10 s	0
2	10	5	20-30 s	1
3	12	5	25-35 s	1

4 Discussion

The optimisation model was tested in three scenarios with an increasing number of intervention requests. The results show that the model can optimise the intervention allocation and select the approach to maintenance delivery to minimise the number of tardy interventions. Scenario 2 and Scenario 3 see the presence of one tardy intervention each caused the impossibility, for the model in its current version, to allocate more than one request in the same window. Thus, this forces the model to search and allocate new requests in the successive windows, limiting the possibility to execute some interventions in the shortest possible time.

The model is not free form limitations and opportunities for improvement:

- *The model allocates only one intervention request per window*. Currently, even though a customer sends multiple requests, each one is treated singularly due to modelling purposes. Relaxing this constraint would favour a faster resolution of the intervention requests, especially when multiple intervention requests are received from the same customer. In this case, the travel time would reduce (the technician would need to travel only once to solve multiple problems). In other words, this would disclose the opportunity to execute opportunistic maintenance, anticipating future interventions.
- *The model allocates an intervention to a single technician*. Based on the context in which the model has been developed (i.e., the Balance Systems context), the model currently does not include the possibility to create teams of technicians for the resolution of interventions requests. At the moment, this limitation may prevent the application of this model to other, more complex, contexts.
- *The model cannot postpone, re-schedule or re-allocate interventions already confirmed*. Currently, the model is not able to change the size of the availability windows by moving the interventions or re-allocating them to technicians that may be free and that can execute them before the moment initially established. Of course, anticipating the intervention also requires confirmation from the customer.
- *The model does not consider costs.* This assumption is part of the model boundaries, which assume the execution costs as part of the contract signed with the customer and already covered. Something that could help in optimising the intervention allocation is considering the penalty cost associated with each customer in case of tardy intervention.
- *The model does not consider weight for the requests.* All the requests are characterised by the same weight, which means that simple and complex problems have the same importance for the model and are differentiated only by the execution and travel times. Something that could help in prioritising the intervention requests is considering weights for the interventions that would allow distinguishing between them facilitating their prioritisation.
- *The model works in a deterministic fashion.* The model considers data and inputs as deterministic, neglecting the possibility that there could be variations in the time required for reaching the customer or the time required to perform the maintenance (e.g., because new problems emerge). Using a stochastics approach would increase the realism and the efficiency of the model in terms of allocation.

5 Conclusions

This paper presented an optimisation model developed in the context of the D3M framework [14]. The model aims to optimise the maintenance delivery by improving the intervention allocation and identifying the optimal resolution strategy considering the objective function (i.e., minimisation of the number of tardy interventions).

The model, adapted to fit Balance Systems necessities, was applied in three different scenarios created using the data collected from the company. The three scenarios were dimensioned based on the average workload of the company in terms of service requests. The application of the model allowed to show the benefits and limitations connected to the model development status. The model presented in this paper should be considered as a first version of a model able to overcome the limitations listed in section 4 and, thus, able to be applied in different a various scenario. In addition, it should be clarified that the baseline used for the model development was based on to the D3M framework logical structure. Being still in a development phase, the model can, at the moment, handle simple situations and requires to be integrated with the human knowledge for the consideration of external factors (e.g., costs).

Following the model application and the company's discussion, it has been possible to identify a set of limitations (Section 4) that will guide future development and research.

Acknowledgements

This research is supported MADE Competence Center in the project “PRocessi, strumEnti e dAti a supporto delle deciSiOni di MaNutenzione 4.0 (REASON4.0)”.

References

1. Rahman, A.R.A., Husen, C.V., Pallot, M., Richir, S.: Innovation by Service Prototyping Design Dimensions & Attributes, Key Design Aspects, & Toolbox Abdul. In: 23rd ICE/IEEE International Technology Management Conference. pp. 587–592 (2017).
2. Potes Ruiz, P.A., Kamsu-Foguem, B., Noyes, D.: Knowledge reuse integrating the collaboration from experts in industrial maintenance management. Knowledge-Based Systems. 50, 171–186 (2013). https://doi.org/10.1016/j.knosys.2013.06.005.
3. Ardolino, M., Rapaccini, M., Saccani, N., Gaiardelli, P., Crespi, G., Ruggeri, C.: The role of digital technologies for the service transformation of industrial companies. International Journal of Production Research. 1–17 (2017). https://doi.org/10.1080/00207543.2017.1324224.
4. Gopalakrishnan, M., Bokrantz, J., Ylipää, T., Skoogh, A.: Planning of maintenance activities - A current state mapping in industry. Procedia CIRP. 30, 480–485 (2015). https://doi.org/10.1016/j.procir.2015.02.093.

5. Bumblauskas, D., Gemmill, D., Igou, A., Anzengruber, J.: Smart Maintenance Decision Support Systems (SMDSS) based on corporate big data analytics. Expert Systems with Applications. 90, 303–317 (2017). https://doi.org/10.1016/j.eswa.2017.08.025.
6. Karim, R., Westerberg, J., Galar, D., Kumar, U., Karim, R.: Maintenance Analytics – The New Know in Maintenance. IFAC-PapersOnLine. 49, 214–219 (2016). https://doi.org/10.1016/j.ifacol.2016.11.037.
7. Mathieu, V.: Service strategies within the manufacturing sector: Benefits, costs and partnership. International Journal of Service Industry Management. 12, 451–475 (2001). https://doi.org/10.1108/EUM0000000006093.
8. Kuo, T.C., Wang, M.L.: The optimisation of maintenance service levels to support the product service system. International Journal of Production Research. 50, 6691–6708 (2012). https://doi.org/10.1080/00207543.2011.616916.
9. Rondini, A., Tornese, F., Gnoni, M.G., Pezzotta, G., Pinto, R.: Hybrid simulation modelling as a supporting tool for sustainable product service systems: a critical analysis. International Journal of Production Research. 55, 6932–6945 (2017). https://doi.org/10.1080/00207543.2017.1330569.
10. Afshar-Nadjafi, B.: Multi-skilling in Scheduling Problems: A Review on Models, Methods and Applications. Computers & Industrial Engineering. 107004 (2020).
11. Agnihothri, S.R., Mishra, A.K.: Cross-training Decisions in Field Services with Three Job Types and Server-Job Mismatch*. Decision Sciences. 35, 239–257 (2004). https://doi.org/10.1111/j.00117315.2004.02642.x.
12. Pal, D., Vain, J., Srinivasan, S., Ramaswamy, S.: Model-based maintenance scheduling in flexible modular automation systems. In: IEEE International Conference on Emerging Technologies and Factory Automation, ETFA. pp. 1–6. Institute of Electrical and Electronics Engineers Inc. (2017). https://doi.org/10.1109/ETFA.2017.8247738.
13. Xu, Z., Ming, X.G., Zheng, M., Li, M., He, L., Song, W.: Cross-trained workers scheduling for field service using improved NSGA-II Cross-trained workers scheduling for field service using improved NSGA-II. International Journal of Production Research. 53, 1255–1272 (2014). https://doi.org/10.1080/00207543.2014.955923.
14. Sala, R., Bertoni, M., Pirola, F., Pezzotta, G.: Data-based decision-making in maintenance service delivery: the D3M framework. Journal of Manufacturing Technology Management. 32, 122–141 (2021). https://doi.org/10.1108/JMTM-08-2020-0301.