Santos, Iuri Martins; Hamacher, Silvio; Oliveira, Fabricio

A data-driven optimization model for the workover rig scheduling problem: Case study in an oil company

Published in:
Computers & Chemical Engineering

DOI:
10.1016/j.compchemeng.2022.108088

Published: 01/02/2023

Document Version
Publisher's PDF, also known as Version of record

Computers and Chemical Engineering 170 (2023) 108088

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

A data-driven optimization model for the workover rig scheduling problem: Case study in an oil company

Iuri Martins Santos [a,b], Silvio Hamacher [a,b], Fabricio Oliveira [c,*]

[a] Department of Industrial Engineering, PUC-Rio, Rua Marquês de São Vicente, 225, Rio de Janeiro, 22451-900, RJ, Brazil
[b] Tecgraf Institute, Rua Marquês de São Vicente, 225, Rio de Janeiro, 22451-900, RJ, Brazil
[c] Department of Mathematics and Systems Analysis, Aalto University, Otakaari 1, PO Box 11100, Espoo, 00076, Finland

ARTICLE INFO

Keywords:
Oil and gas
Workover rig scheduling problem
Data-driven optimization
Simulation

ABSTRACT

After completion, oil wells often require intervention services to increase productivity, correct oil flow losses, and solve mechanical failures. These interventions, known as workovers, are made using oil rigs, an expensive and scarce resource. The workover rig scheduling problem (WRSP) comprises deciding which wells demanding workovers will be attended to, which rigs will serve them, and when the operations must be performed, minimizing the rig fleet costs and the oil production loss associated with the workover delay. This study presents a data-driven optimization methodology for the WRSP using text mining and regression models to predict the duration of the workover activities and a mixed-integer linear programming model to obtain the solutions for the model. A sensitivity analysis is performed using simulation to measure the impact of the regression error in the solution.

1. Introduction

Oil and gas production relies on several techniques and associated equipment that are responsible for lifting the oil to the surface of the well. Eventually, equipment failures require intervention services to restore productivity or correct oil flow losses. These interventions, known as workovers, vary from recompletion to restoration, cleaning, stimulation, and others operations that require the use of oil rigs (Chaudhuri, 2011). Oil rigs are expensive and scarce resources that cost between US$ 50,000 and US$ 700,000 per day, depending on their type, market, and operational characteristics (Kaiser and Snyder, 2013; Osmundsen et al., 2010).

An undersized fleet of rigs might lead to delays in oil production, jeopardizing the profitability of the wells. In contrast, an oversized fleet may lead to high idleness and opportunity costs. Consequently, rig fleets must be properly planned and scheduled to ensure that the rigs will be available at the right place at the right time with the lowest possible cost (Santos et al., 2021).

Each well has its characteristics and properties, which usually require a specific type of workover rig to serve it (Fernández Pérez et al., 2018). Moreover, workover operations are of varying complexity; some wells may require a single day for an intervention to be completed, while others can require months. As a result, it might not be possible to execute all workovers operations within a given planned time horizon. Therefore, companies may need to decide which wells will be attended to according to their oil production and the availability of rigs.

This decision-making process is known as the workover rig scheduling problem (WRSP). In this problem, wells require workovers (interventions with the purpose of correcting or restoring oil flow) during the scheduling horizon. Differently from traditional scheduling problems, these time horizons are typically long, in the scale of months or a few years. This is due to the nature of the activities performed, whose durations are typically of several days or months. These interventions are performed by oil rigs and can only be made on the wells after a release date related to the well's life cycle and their production schedules. Wells requiring workover have an oil production loss associated with their waiting time. As mentioned by Santos et al. (2021), oil rigs are scarce, expensive, and often custom-built resources. Consequently, the fleet of rigs that serves the wells has to be hired long before the actual need for workover. The goals of the WRSP are to determine the fleet of rigs to be hired, select the wells that will be attended to, and schedule the rigs to the wells (*i.e.*, when and by which rigs the wells will be served), aiming at minimizing the rig fleet costs and the oil production loss of the wells. As the demand for rigs is dictated by the duration and amount of workover activities, knowing the duration precisely leads to a better-sized fleet of rigs, making it necessary to use proper methods to estimate the duration of the workover activities.

* Corresponding author.
E-mail addresses: iuri.santos@tecgraf.puc-rio.br (I.M. Santos), hamacher@puc-rio.br (S. Hamacher), fabricio.oliveira@aalto.fi (F. Oliveira).

https://doi.org/10.1016/j.compchemeng.2022.108088
Received 5 May 2022; Received in revised form 22 November 2022; Accepted 26 November 2022
Available online 6 December 2022

This study addresses the workover rig scheduling problem (WRSP) and proposes a data-driven optimization model that estimates the workover duration and generates rig schedules simultaneously. The duration of the workover is predicted, taking into account the decision-dependent nature of the duration, which depends on the matching between the technical specifications of the well and the rig chosen to perform the workover. We perform such predictions by means of a combination of data science techniques, which allows us to naturally model the decision-dependent nature of the workover activity duration without compromising the linearity of the model. The prediction is made based on a combination of techniques. Specifically, text mining, clustering, and regression models were used on historical data, enabling these predictions to be utilized in a mixed-integer linear programming (MILP) model that minimizes rig fleet costs and the oil production loss of the wells.

Data-driven optimization is a recent trend in the Operation Research community that combines mathematical programming with data science and statistical algorithms. Hence, the proposed combination of mathematical programming with text mining, clustering, and regression models contributes to this trend. Furthermore, there is a lack of data-driven optimization models in the rig scheduling problem, as mentioned by Santos et al. (2021). Therefore, the main contribution of this study is the proposed data-driven methodology to improve the representation of the decision-dependent workover duration using historical data. Another contribution is the proposed mathematical model itself, which is a reformulation of Costa and Ferreira Filho (2004)'s model for WRSP with more realistic assumptions, such as a heterogeneous fleet of rigs, multi-objectives, and rig eligibility. Finally, the model is applied to realistic instances, contributing to the connection between academia and industry. These instances are generated based on historical data of the studied company and are realistic to the extent they can represent the problem's main features. Lastly, the proposed data-driven model is compared with the methodology used in practice to set the rig schedules, and this analysis demonstrated the benefits of more accurate predictions for the workover duration.

The paper is divided into six sections. Section 2 reviews the literature on the rig scheduling problem. Section 3 presents the WRSP under study and the methodology used in this research. Section 4 presents the data treatment methods utilized. This treated data is used in regression models to predict the workover duration in Section 5. Two mathematical programming formulations using the outputs from the data treatment and regression models are proposed and tested for the studied WRSP in Section 6. Section 6.3 performs a simulation of different solutions to measure their sensitivity against the prediction error associated with the regression. Lastly, Section 7 reflects on the final considerations of the research and potential future studies of the WRSP.

2. Literature review

The workover rig scheduling problem is a particular case of the rig scheduling problem (RSP), the scheduling and allocation of well activities to rigs aiming to avoid delays and optimize the use of resources (Eagle, 1996). According to Santos et al. (2021), the RSP can be divided into four major classes of problems:

- *Drilling Rig Scheduling Problem (DRSP)*: drilling and completion rig scheduling problems, where scheduling is an isolated choice from the rest of the field development decisions;
- *Workover Planning*: rig scheduling of workover activities, which is typically separated from the other rig-related decisions as they are planned in the production phase. It can be classified into two sub-groups according to the application of routing: workover rig scheduling problems (WRSP) and workover routing and scheduling problems (WRRSP);
- *Resource Planning*: rig scheduling incorporates the planning of different resources besides rigs, such as offshore support vessels (OSVs), equipment, and crews. An example is the planning of the OSVs used to lay the pipes connecting the wells and platforms; their connections can only begin after well drilling and completion (Abu-Marrul et al., 2020).
- *Field Planning*: when rig scheduling is integrated with other oilfield development decisions, such as field design, reservoir modeling, and production flow scheduling. In these cases, the RSP relies upon or affects other parts of the field development;

The first articles about RSP were from Aronofsky and Williams (1962) and Aronofsky (1962). The authors proposed two linear programming models for the planning of oil production. At that time, these mathematical models required considerable computational effort, preventing any functional application (Pittman, 1985). Consequently, most of the developments regarding the RSP were simplified, using approximation techniques (Barnes et al., 1977) or decision-making rules (Cochrane, 1989). With the improvement of computer processing capabilities and optimization techniques in the 1990s, RSP studies began to broaden themselves, as mentioned by Santos et al. (2021).

There are several literature reviews considering the RSP. Bassi et al. (2012) studied the workover rig routing and scheduling problem and presented a literature review about its setting. Bissoli et al. (2016) also performed an extensive review on the workover routing and scheduling problems, focusing on its drivers. According to the authors, the RSP trends were to approximate the problem with real-life scenarios through new objective functions, mathematical formulations, solution methods, and dynamic or stochastic approaches. Santos et al. (2021) expanded on Bissoli et al. (2016)'s study with a systematic literature review covering most variants of the rig scheduling problem. The authors proposed a unique taxonomy for the RSP addressing its key features and reviewed 130 studies, detecting several gaps and trends in the literature, such as a trend for optimization under uncertainty and a lack of data-driven optimization models, which this paper intends to fulfill.

Others authors have provided a general analysis that relates to the RSP. Tavallali and Karimi (2014) and Tavallali et al. (2016) discussed the planning and development of oilfield decisions and associated perspectives, reviewing several studies, including some on rig scheduling. According to Tavallali and Karimi (2014), rig scheduling is an open research topic that needs more attention. Tavallali et al. (2016) focused on reservoir models and their optimization approaches but proposed a general classification for field development problems, in which the rig scheduling is an oilfield operation decision. The authors highlighted the lack of scheduling studies for drilling new wells and suggested that it should be an integral part of well placement models and oilfield development planning. Khor et al. (2017) also performs a review of field development problems but focuses on the optimization methods used rather than the problems.

This study focuses on the workover rig scheduling problem. Therefore, the literature review presented in this section will be limited to workover planning problems and separated according to the use or not of routing: workover rig scheduling problems (WRSP), Section 2.1, and workover rig scheduling and routing problems (WRRSP), Section 2.2.

2.1. Workover rig scheduling problem

The workover rig scheduling problem was first addressed by Barnes et al. (1977), proposing two approximation techniques to minimize the loss of oil production and testing them on a small and short-term instance. Pioneering advances in the WRSP were made by Costa and Ferreira Filho (2004, 2005). The authors proposed a linear integer programming model and 300 real-life instances for the problem that was used in many other studies later. Thus, different heuristics were

Table 1
Summary of the studies approaching the workover rig scheduling problem (WRSP).

Authors (Year)	Field	Instances	Jobs	Fleet	Approach	Objectives
Barnes et al. (1977)	–	Real data	Single	Homogeneous	Heuristic	Single
Costa and Ferreira Filho (2004, 2005)	Onshore	Real data	Single	Homogeneous	Heuristic	Single
Lasrado (2008)	Offshore	Theoretical data	Single	Homogeneous	Simulation	Multi-Objective
Ribeiro et al. (2011)	Onshore	Public data	Multiple	Homogeneous	Heuristic	Single
Marques et al. (2014)	Offshore	Real data	Single	Homogeneous	Exact	Single
Monemi et al. (2015)	Onshore	Real data	Single	Heterogeneous	Heuristic; Matheuristic	Single
Danach (2016)	Onshore	Real data	Single	Heterogeneous	Heuristic	Single
Kromodihardjo and Kromodihardjo (2016)	–	Real data	Single	Homogeneous	Heuristic	Single
Pérez et al. (2016)	Onshore	Public data	Single	Heterogeneous	Exact	Single
Vasconcelos et al. (2017)	Offshore	Real data	Single	Heterogeneous	Heuristic	Single
Fernández Pérez et al. (2018)	Onshore	Public data	Single	Heterogeneous	Simu-Optimization	Single

tested or created for the problem, such as a maximum priority three-criteria heuristic, MPTH (Costa and Ferreira Filho, 2004); a dynamical assemble heuristic, DAH (Costa and Ferreira Filho, 2005).

Aiming to address large instances, Ribeiro et al. (2011) proposed a simulated annealing (SA)-based heuristic that uses SA to create a preliminary solution and iteratively enhance it with SA, which allowed it to surpass other methods in the instances of Costa and Ferreira Filho (2004), such as GRASP, GRASP-PR, DAH, BS, SS, MA, and GA-2opt.

A few other variations of the WRSP can be found in the literature. For instance, Lasrado (2008) developed a software application using manual procedures combined with reservoir simulation (de Andrade Filho, 1994) to create schedules minimizing the number of rigs and the traveling distances, which reduces contract and transportation costs. Marques et al. (2014) proposed a decision support system that schedules a homogeneous fleet of offshore rigs aiming to minimize its size and utilization through MILP.

Monemi et al. (2015) considered a heterogeneous fleet of rigs, presenting a new MILP model with arc-time-indexed formulations and two techniques: branch-price-and-cut (BPC) and hyper-heuristic (HH) that obtained near-optimal results in a remarkably short time. This same problem was addressed by Danach (2016) with a binary linear programming model and a HH, which was examined in a real case, and presented problems solving the large instances. The researchers suggested future improvements in the efficiency of the mathematical formulation.

Pérez et al. (2016) adapted the binary linear model from Costa and Ferreira Filho (2004) to the case of heterogeneous onshore rigs, proposing a decomposed reformulation with fewer variables and constraints, obtaining new exact solutions for Costa and Ferreira Filho (2004)'s large instances and surpassing the heuristic methods. This mathematical model was later reformulated by Fernández Pérez et al. (2018) to take into account uncertainty in the duration of tasks through a stochastic programming model that minimizes the loss of oil production and the costs of the drilling fleet. The model was tested in instances adapted from Paiva et al. (2000), Costa and Ferreira Filho (2004) and Ribeiro et al. (2012a) in terms of the problem's features, using different scenario generation methods, such as Monte Carlo simulation and Quasi-Monte Carlo. Next, Table 1 summarizes the WRSP studies presented in this section.

2.2. Workover rig routing and scheduling problem

When the wells demanding workovers are not concentrated near to each other and the traveling time between the wells is not negligible, routing techniques are required, which leads to the workover rig routing and scheduling problem (WRRSP) (Bissoli et al., 2016). The WRRSP discussion began with a SA proposed by Paiva et al. (2000) aiming to minimize the oil production losses and costs of a homogeneous fleet of workover rigs.

After that, several heuristics were proposed to solve the homogeneous WRRSP, such as: ILS, clustering search, and an adaptive large neighborhood search (ALNS) (Ribeiro et al., 2012b); ALNS with aggregated rank removal heuristic (ARRH), GA, and GA with VNS (GA + VNS) (Shaji et al., 2019). Of these different heuristics, the best results were obtained with ALNS from Ribeiro et al. (2012b) and ARRH-based ALNS (Shaji et al., 2019).

Meanwhile, other researchers concentrated on new modeling approaches for the WRRSP with a homogeneous fleet. Duhamel et al. (2012) proposed a MILP model based on Aloise et al. (2006), another method based on the open vehicle routing problem, and a set-covering model using Dantzig–Wolfe decomposition and an alternative column generation method with variable neighborhood descent and GRASP. Finally, Kromodihardjo and Kromodihardjo (2016), in a combinatorial optimization approach, employed discrete simulation to perform an exhaustive search in the problem, which also led to reasonable solutions in small real-life instances.

Similarly to the WRSP, some authors address the WRRSP with heterogeneous rigs. Aloise et al. (2006) designed a VNS heuristic mixing swap (changing the wells allocated to a rig) and insert move (inserting wells to a rig itinerary) and implemented it in a Brazilian company, which led to savings of approximately 2.5 million dollars per year. Using column generation, ng-path relaxation, subset-row inequalities, and TS, Ribeiro et al. (2012a) proposed a BPC algorithm to optimally solve real-life examples with as many as ten rigs and two hundred wells. Ribeiro et al. (2014) compared this BPC from Ribeiro et al. (2012a), the ALNS made by Ribeiro et al. (2012b), and the VNS from Aloise et al. (2006) with a hybrid-GA (HGA) that outperformed the other methods.

Focusing on the data exploration to enhance the solution quality, Vasconcelos et al. (2017) combined a GA and operational historical data to minimize the non-productive time of wells, testing it on a petroleum company and improving 20 to 40% of the operational and navigation time. Another GA was proposed by Tozzo et al. (2020) to minimize multiple objectives (rig fleet costs and oil production loss).

As the business environment has become more dynamic nowadays and many decisions are made without knowing the full picture, there is a trend in the Operations Research community to optimize under uncertainty, which can be observed for the WRRSP in the studies of Bassi et al. (2012), and Silva and Silva (2018). Bassi et al. (2012) developed a method to simulate the duration of the workovers and optimize the schedule with GRASP. Last, Silva and Silva (2018) introduced a WRRSP in which the decision maker does not know beforehand where the workovers will be required (which wells will need maintenance), naming it Dynamic WRRSP (D-WRRSP). The proposed formulation was based on Ribeiro et al. (2012a)'s formulation and tested in short-term instances modified from Costa and Ferreira Filho (2004). Next, Table 2 summarizes the WRRSP discussed in this section.

2.3. Review outline and insights

The first RSP studies focused on the DRSP. Research considering workover planning only began to grow in the 2000s, with studies addressing the WRSP, most of them proposing heuristics for the problem. Sometime later, with the advances in techniques for VRP, the WRRSP started to gain attention. Nowadays, several model formulations and

Table 2
Summary of the studies approaching the workover rig routing and scheduling problem (WRRSP).

Authors (Year)	Field	Instances	Jobs	Fleet	Approach	Objectives
Paiva et al. (2000)	Onshore	Real data	Single	Homogeneous	Heuristic	Multi-Objective
Aloise et al. (2006)	Onshore	Real data	Multiple	Heterogeneous	Heuristic	Single
Bassi et al. (2012)	Offshore	Theoretical data	Single	Heterogeneous	Simu-Optimization	Single
Duhamel et al. (2012)	Onshore	Real data	Single	Homogeneous	Heuristic; Matheuristic	Single
Ribeiro et al. (2012a)	Onshore	Public data	Single	Heterogeneous	Matheuristic	Single
Ribeiro et al. (2012b)	Onshore	Public data	Single	Homogeneous	Heuristic	Single
Ribeiro et al. (2014)	Onshore	Public data	Multiple	Heterogeneous	Heuristic; Matheuristic	Single
Kromodihardjo and Kromodihardjo (2016)	–	Real data	Single	Homogeneous	Heuristic	Single
Silva and Silva (2018)	Onshore	Theoretical data	Single	Heterogeneous	Exact	Single
Shaji et al. (2019)	Onshore	Theoretical data	Single	Heterogeneous	Heuristic	Multi-Objective
Tozzo et al. (2020)	Onshore	Public data	Single	Heterogeneous	Heuristic	Multi-Objective

heuristic methods have already been proposed, both for the WRSP and WRRSP. According to Santos et al. (2021), workover planning is now the most popular subject concerning rig scheduling problems.

Currently, the approaches tend to combine mathematical programming, heuristics, and simulation and take into account more realistic assumptions and objective functions, such as fleet availability and eligibility considerations (heterogeneous rigs), multiple objectives (rigs fleet costs and oil production loss), net present value, and costs varying over the scheduling horizon.

Furthermore, the complex and risky workover environment requires techniques that reduce uncertainty and can cope with errors in the data, such as stochastic/robust optimization, simulation optimization, dynamic programming, or data-driven optimization. Most of these techniques have been applied in some way in the WRRSP (Bassi et al., 2012; Silva and Silva, 2018; Vasconcelos et al., 2017). However, the WRSP has received less attention in these types of approaches. Some stochastic and robust models were proposed by Fernández Pérez et al. (2018), but there is no data-driven optimization study for the WRSP.

Another literature gap detected by Santos et al. (2021) is that more studies need to be applied in real instances and validated with the decision-makers, strengthening the integration between the academic and industry perspectives.

Aiming to fulfill these gaps, this study proposes a data-driven optimization framework for the workover rig scheduling problem for a heterogeneous fleet of offshore rigs. This data-driven approach first uses text mining and clustering algorithms to extract information from historical data from a Brazilian oil company. Then, this information is used in regression models to predict the duration of the workover activities according to the rig. Finally, an optimized workover rig schedule is obtained with an MILP model that aims to minimize oil production losses and rig fleet costs. Further details on the problem at hand and the methodology used are given in the next section.

3. Materials and methods

This section defines the workover rig scheduling problem, proposes a data-driven optimization methodology that tackles some of the literature gaps detected in the last section, and clarifies some key elements of the techniques used in the methodology.

3.1. Problem definition

This article considers a Brazilian oil company that operates a large number of oil fields and needs to plan a fleet of rigs to operate its offshore wells. As a result, this case study has some particularities. This large set of wells requires workover activities, and a fleet of rigs must be hired to serve them. The goal is to decide which wells will be served by which rig in the scheduling horizon, minimizing the costs associated with hiring the rigs and the oil production loss of the wells waiting for workover service. The offshore wells are relatively close to each other, and their processing times are much longer than the traveling times between them, making thus traveling times negligible. Therefore, routing considerations can be disregarded, and their scheduling sequence naturally yields a route for the rig. As a result, we can classify this problem as a workover rig scheduling problem (WRSP), which is a particular case of the rig scheduling problem for workover operations.

Workover planning is performed separately from the other operations on a stand-alone planning level. In that, a fleet of heterogeneous rigs is hired to execute them. Each rig has a particular maximum water depth and a drilling depth. Moreover, each well has a water depth and a drilling depth that cannot exceed the rig limits. Rigs have a fixed cost when hired. Others resources, in addition to rigs, are not considered in this case study.

Each well has an oil production associated with it, regardless of whether it is an injector or producer well. Further details on the oil production of the wells are provided later when we describe the instance generation (Section 6.2). Every well requires only one maintenance (or rework) operation (job or task). Basically, it is a single job scheduling problem for which we use the terms *well*, *workover*, *operation*, *task*, and *job* interchangeably. Furthermore, every well has a release date related to the date it starts needing workover, and there is a cost associated with the oil production loss of the wells waiting to be served, which extends until the end of the scheduling horizon if the well is not served.

Lastly, the processing time for each workover operation varies for each class of rig. However, these processing times are not known before scheduling a well to a rig. Currently, the company studied uses the average duration for the type of workover. However, historical data from the workover operations is available and can be used to predict the processing time of a particular rig in a well. Details on the historical data will be presented in Section 4.

3.2. Methodology

This section proposes a data-driven methodology for the workover rig scheduling problem, which is separated into three major phases: data treatment (in which the workover historical data is cleaned, shortened, and labeled using data science techniques, including text mining and clustering); predictive models (when the treated data is applied into predictive models to estimate the workover duration according to a well and a rig); optimization (a mixed-integer linear programming model is used to determine an optimal workover rig schedule). Fig. 1 summarizes these three phases presented in Sections 4, 5, and 6, respectively.

Data treatment is based on the data science framework from Shcherbakov et al. (2014) and separates data into two types, qualitative and quantitative data, applying text mining, clustering, and statistical techniques. As explained by Srnka and Koeszegi (2007), quantitative data refers to numerical variables, such as duration, costs, and other measures of value. On the other hand, qualitative data are categorical variables, usually represented with text, symbols, codes, and other nominal categories. The quantitative data is cleaned by removing errors, duplicated rows, and empty fields. With the assistance of plots, such as boxplots (with a multiplier of $1.5 \times IQR$, where IQR is the interquartile range) and histograms, outliers are eliminated, generating numerical variables for the predictive models.

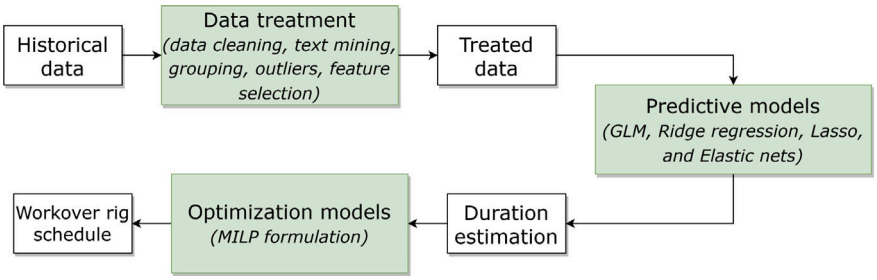


Fig. 1. Data-driven optimization methodology.

The qualitative data is treated with text mining techniques (responsible for cleaning the data) and clustering models (which propose better groups for the treated data) to generate dummy variables.

The text mining procedures were generated using the *R* public packages "tau", "tm", "SnowballC", "wordcloud", and "stringdist" and include:

- *Data cleaning:* which is the removal of symbols (such as: "/,@,',",|, -,_"), the converting of the text to lower case only, and the removal of numbers, accent marks, dots, and extra spaces.
- *Data simplification:* removal of stopwords and use of the stemming technique (adapted for the Portuguese language) (Lang, 2004). Stopwords are uninformative words often common in a text, such as: articles, pronouns, and conjunctions (Sarica and Luo, 2021). The complete list of the Portuguese stopwords used is shown in Appendix A. Meanwhile, the stemming technique reduces inflected or derived words to their respective word stems, simplifying the text and making it easier to identify fields with the same meaning (Jivani et al., 2011). For instance, words such as "removal", "removing", "removed", and "removes" are replaced by their word stem "remov". Basically, the stemming technique and the data cleaning simplify the data. However, these techniques would still not recognize texts with the same meaning as similar. For instance, the terms "Removing of equipment" and "Equipment removal". The stopword removal would remove the "of" from the first text, and the stemming would transform each one of them into "Remov equip" and "Equip remov", respectively. A clustering model is used to detect these similar text fragments and group them.

The grouping of the text data was made using the *R* public packages "pheatmap", "dendextend", "ggdendro", and "cluster" and include the following procedures:

- *Distance measure:* which uses string similarity and distance tools to measure how close the sentences of the qualitative data are to each other. After several tests, a custom string similarity measure was created using the Levenshtein (LV) (Yujian and Bo, 2007) and the Longest Common Substring (LCS) (Sun et al., 2015) distances. This custom string similarity measure for two strings is the mean between both these measures:

$$\text{String Similarity}\,(s1, s2) = \frac{LV\,(s1, s2) + LCS\,(s1, s2)}{2}, \tag{1}$$

  where $s1$ and $s2$ in Eq. (1) refer to "String1" and "String2", respectively. The LV distance is an edit-based string similarity, whereas the LCS similarity is a sequence-based measure. Both similarity measures are efficient for short strings like the task description, and the combination of the two resulted in suitable matches.
- *Clustering methods:* which uses the k-means algorithm (Likas et al., 2003), a partition method that separates the data into a predefined number of mutually exclusive clusters (k). It is a point-based clustering method that starts with the cluster centers initially placed in arbitrary positions and proceeds by moving the cluster centers at each step to minimize the clustering error (Likas et al., 2003). A crucial part of the k-means algorithm is the definition of the number of clusters (k), which is usually defined using the average silhouette analysis. The silhouette score measures how similar objects are to their assigned clusters compared to other clusters. The score varies between −1 and +1, and a higher score indicates that the object is well-matched to its own cluster and poorly matched to other neighboring clusters (Rousseeuw, 1987).

The string similarity measure in Eq. (1) was used as the distance for clustering algorithms that aim to group textual descriptions according to their similarities.

As illustrated in Fig. 2, linear regression models are applied in the treated data aiming to predict the duration of the workovers. Linear regression models are statistical models used to determine the relationship between a response variable (Y) and its explanatory variables (X), which can then be used to predict response values for newly observed explanatory variable values. Two types of regression are tested and evaluated:

- *Generalized linear models (GLMs):* it is a generalization of ordinary linear regression models that accepts response variables with errors following an exponential family distribution, not necessarily a normal distribution as the ordinary models (Nelder and Wedderburn, 1972). The value predicted by the GLM for the observation Y_n is a linear sum of the effects of one or more explanatory variables X_{nm}, as shown in Eq. (2):

$$Y_n = \beta_0 + \beta_1 X_{n1} + \cdots + \beta_m X_{nm} + \cdots + \beta_M X_{nM} + \epsilon_n,\ \forall n \in N, \tag{2}$$

  where $n = \{1, \ldots, N\}$ represents the set for all observations, $m = \{1, \ldots, M\}$ denotes the number of explanatory variables (or features) (X_{nm}) used, and β_m represents their effect on the response variable Y_n (McCullagh and Nelder, 2019).
- *Ridge regression (RR) models:* RR is a multiple regression technique adapted for data with multicollinearity (when the least-squares estimates are unbiased, but their variances are significant, causing them to be far away from the actual value). Ridge regression adds a degree of bias to the regression estimates by adding a penalty in the sum of the squares (L2 normalization), reducing standard errors. This technique is recommended for regression models with near-linear relationships among independent variables or many independent dummy variables (Hoerl and Kennard, 1970).
- *Lasso regression models:* Lasso or least absolute shrinkage and selection operator is another type of multiple regression technique with regularization that adds bias by penalizing the sum of the absolute values (L1 normalization). This technique is also recommended for regression models with a near-linear relationship among independent variables or a large number of dummy variables (Tibshirani, 1996). As mentioned by James et al. (2013), the Lasso regression can sometimes be used for feature selection as it can completely reset the coefficients.

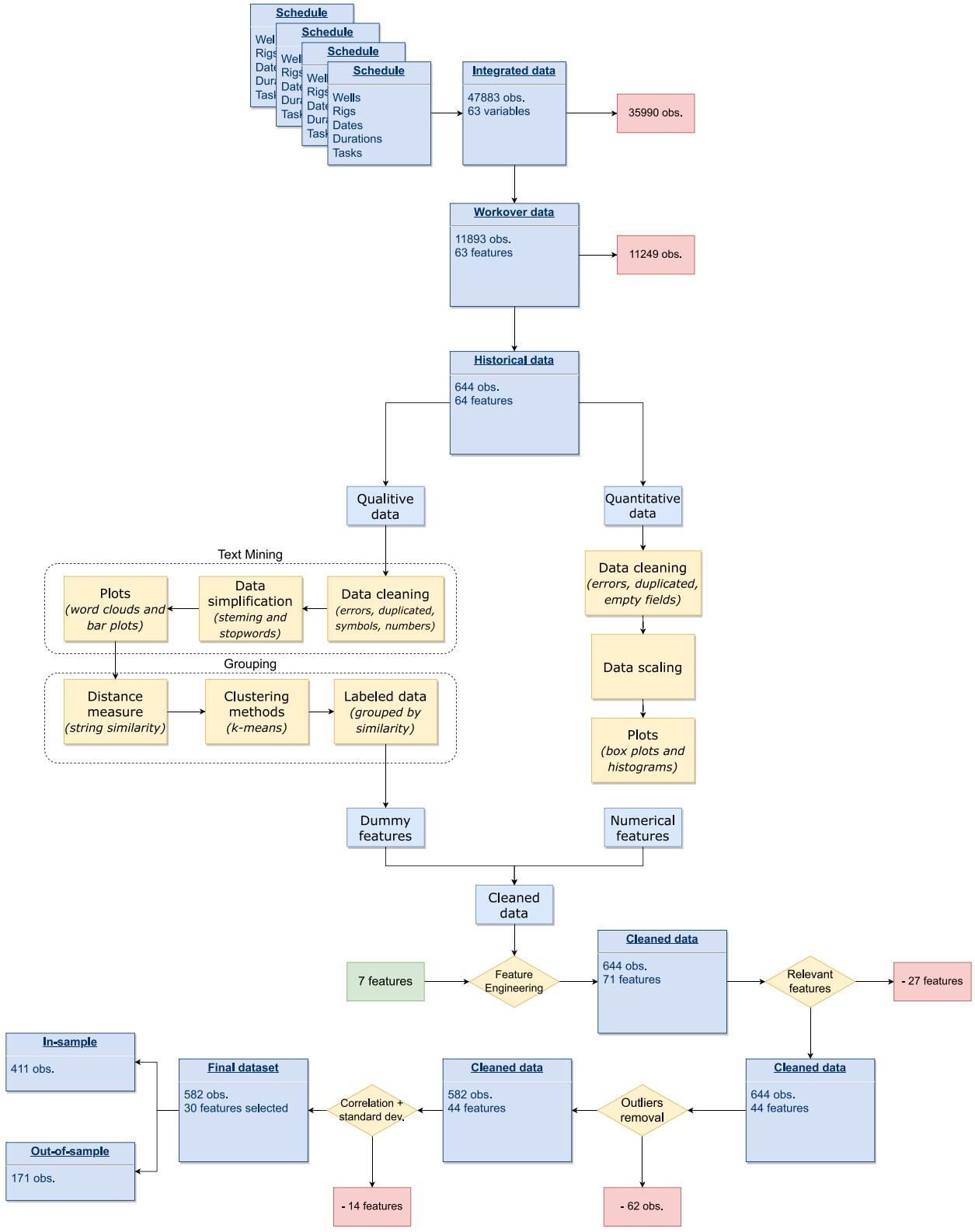
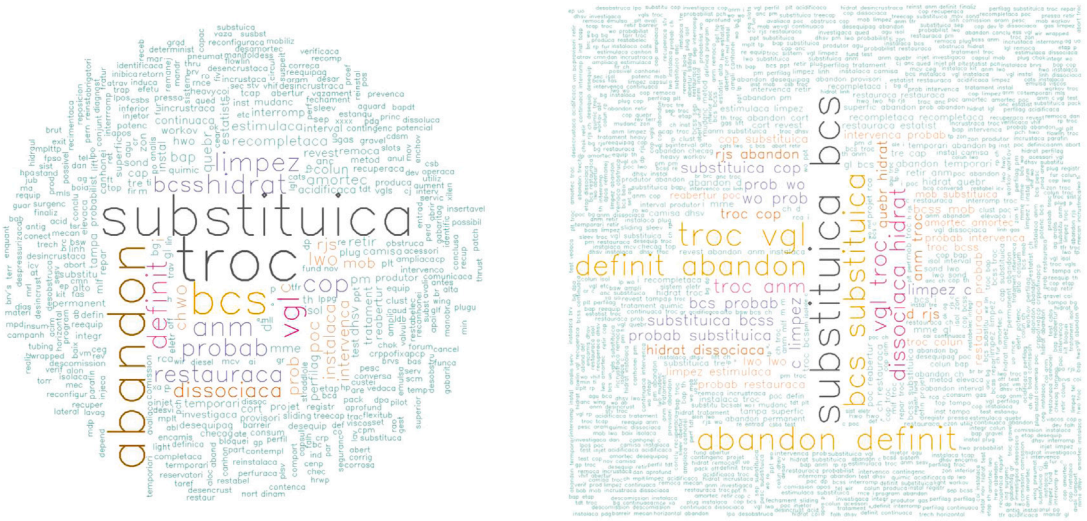


Fig. 2. Data treatment methodology.

- *Elastic net regression models:* Elastic nets are another type of regularized linear regression that combines the L1 and L2 normalizations, *i.e.,* the ridge and lasso regression models, resulting in a more stable feature selection from the L1-normalization and grouping correlated variables using the L2-normalization (Zou and Zhang, 2009).



(a) One word (b) Two word

Fig. 3. Word clouds for one word (a) and two words (b) using the simplified task description.

In the GLM, the error variable ϵ follows a distribution of the exponential family, which includes the Normal, Poisson, Binomial, and Gamma distributions. Linear coefficients are estimated using the maximum likelihood estimation (MLE) method if the residuals are non-Normal or ordinary least squares (OLS) otherwise (Yuan and Yang, 2005; Yan and Su, 2009; Mahmoud, 2019). Several packages are available in the R programming language to estimate generalized linear models. In this study, we used the native library *Stats* (R Core Team, 2013) and the package *olsrr* (Hebbali and Hebbali, 2017). These packages allow one to estimate the coefficients of the model that minimize the loss function.

However, if there are many dummy variables (as a result, a large number of coefficients), the model can overfit the training data and might not perform properly on an out-of-sample data set. Aiming to assist in those cases, regularization techniques can be used to reduce the number of features and prevent overfitting results, such as ridge regression (McDonald, 2009). As this study proposes using qualitative data as an input to predict the unknown workover duration, a large number of independent dummy variables may be generated. Therefore, the ridge model has been chosen as an alternative testing method. The ridge, lasso, and elastic net regression models were estimated using the *glmnet* (Engebretsen and Bohlin, 2019), *stats* (R Core Team, 2013), and *Caret* (Kuhn et al., 2020) libraries for the R programming language.

Using the previous libraries for GLMs and ridge regression, a procedure was created, exhaustively testing all possible combinations of response variables to predict each of the regressions mentioned above. Based on the hold-out validation, the procedure separates 80% of the data as an in-sample and the others 20% left as out-of-sample data. In-sample data is used to train the regression model, and the out-of-sample data is used to predict and evaluate the trained models. The GLMs are fitted using the iteratively reweighted least squares (IWLS) (Street et al., 1988). Meanwhile, the ridge regression models are trained using a 10-fold cross-validation (Bengio and Grandvalet, 2004) within the in-sample data. The trained models are then evaluated for their prediction capabilities using the out-of-sample data with the following metrics: root-mean-square error (RMSE), R-squared (R^2), and p-value fit for residuals normally distributed. The goal is to choose a model with a high R-squared, low error, and possibly low complexity and having residuals normally distributed. The Caret package (Kuhn et al., 2020) was used to train and select the regression models as it automatically selects the optimal features and parameters, allowing to decide the algorithm to choose between ridge, lasso, and elastic nets. Last, the selected model is used to predict the duration. In what follows, we apply the methods described in Section 3.2 and present the results.

With the duration predictions, a MILP model is optimized using the Gurobi solver v. 9.1.2 (Gurobi Optimization, 2018), generating a workover rig schedule. Next, we apply this proposed data-driven optimization methodology to the workover rig scheduling problem. Section 4 presents workover data treatment results. Section 5 tests and selects the regression models for the workover duration. Finally, Section 6 compares different mathematical programming formulations for the WRSP.

4. Workover data treatment

As mentioned in Section 3.1, the workover duration is unknown before scheduling the workover rigs. Currently, the studied company uses an average duration according to the type of workover. However, there are historical data that can be used to estimate the workovers following the methodology proposed in Section 3.2. Table 3 summarizes this historical data according to the data group (well or rig attributes) and type (qualitative or quantitative data):

Most of this information is qualitative data, i.e., non-numerical. Only a few fields are quantitative data (numerical), such as those related to depth and water depth. Furthermore, there are several issues with the qualitative data that require corrections. For instance, the workover groups and workover types are poorly grouped, making it hard to obtain any distribution for the duration using only this information. Aiming to enhance the task grouping, a data treatment methodology based on the data science framework by Shcherbakov et al. (2014) is used to obtain representative task groups and to improve the qualitative data in the case study. The proposed method uses the well data with the task description, which is unstructured, with unnecessary words and letters, and prone to errors. Fig. 2 illustrates the proposed methodology.

An example of the cleaned and simplified data is shown in Appendix B. Word cloud plots were made to check for any patterns in the data. Fig. 3 contains two word-cloud plots, (a) for one word alone (1-g) and (b) for two words together (2-g). We can observe that some words are more common in the task description, such as “abandon” (when a well needs to be abandoned), “troc” and “substitu” (related to the replacement of equipment in the well), and “bcs” (which is a Portuguese acronym for *Bombeio Centrifugo Submerso*, in English: Electrical Submersible Pump, ESP). However, many sentences still have similar meanings and could technically be considered the same sentence. For

Table 3
Description of the historical data gathered.

Data	Group	Type	Description
Workover group	Well	Qualitative	The workover operations are grouped according to the complexity: workover, light workover, and heavy workover.
Workover type	Well	Qualitative	Specifies the type of workover made, such as drilling, completion, appraisal, or abandonment.
Task description	Well	Qualitative	Describes all the essential information about the workover and the well.
Well's project	Well	Qualitative	Specifies the company's project of which the well is part. A project represents a set of wells that share budgets, resources, and performance expectations.
Well's basin	Well	Qualitative	Related to the basin in which the reserve is located.
Well's subpool	Well	Qualitative	Specifies the company's department responsible for the well operation and planning.
Well's water depth	Well	Quantitative	Stores the distance between the sea level and bottom in which the well is located.
Well's depth	Well	Quantitative	Stores the distance between the sea bottom in which the well is located and the oil reserve.
Rig's type	Rig	Qualitative	Specifies if the offshore rig is a fixed rig, a semi-submersible, a jack-up rig, or a drill-ship.
Rig's maximum water depth	Rig	Quantitative	Defines the rig's maximum water depth that it can operate.
Rig's maximum depth	Rig	Quantitative	Defines the rig's maximum depth that it can operate.

instance, "substitu bcs" (replacement of ESP) and "bcs substitu" (ESP replacement) share the same meaning. This issue also occurs with "abandon definit" (abandon definitively), "definit abandon" (definitive abandonment), and other sentences. String similarities combined with clustering algorithms can be used as a grouping model to detect text with similar meaning.

The string similarity measure in Eq. (1) was used as the distance measure of a k-means algorithm (Likas et al., 2003) to group cleaned textual descriptions according to their similarities. With the silhouette analysis, two strategies were selected to cluster and classify the workover tasks. The first clustering strategy separates the task description into major groups of tasks ($k = 7$, fewer clusters). Meanwhile, the second clustering strategy selects smaller groups of tasks description, but not too small ($k = 45$, more clusters). We have chosen to use the second with k (clusters) equal to 45 as they contained more information that was hidden in the historical data, providing 45 new groupings for the workover operations based on the string similarity of the task descriptions.

Overall, the text mining procedures were able to clean the qualitative data, which had several errors, and to extract only the critical information. Furthermore, the clustering algorithms are powerful tools to group the essential knowledge and obtain new data classifying the workovers. Finally, this data with the new grouping is analyzed in a feature engineering perspective, using correlation, standard deviation, and pair plots to carefully select the features that are associated with workover duration and are more likely to improve the regression models. Fig. 4 presents the correlation or strength-of-association of the features in the data set with the workover duration, using Pearson's R for continuous–continuous cases, correlation ratio for categorical–continuous cases, and Cramer's V for categorical–categorical cases.

The first features in Fig. 4 are over-correlated with the workover duration as there are not enough observations for its several categories and, therefore, were removed as a possible feature to be used. Nonetheless, many other significant features were detected, such as 'Bloc', 'Rig type', 'Clusters45', and 'Rig Water Depth' have a significant association. With the support of a standard deviation analysis and a complete correlation matrix of the features (presented in the Appendix), 30 features were selected to be used as an input to the duration prediction in the following section, which presents the regression models used to model the workover durations after the data treatment.

5. Regression models for the workover duration

Statistical techniques play an essential role in the oil and gas upstream. There have been several successful cases using statistics to predict operation times and to support their planning. Desai et al. (2020) reviewed some of these studies and mentioned techniques such as regression models, neural networks, machine learning, and support vector machine models. Motivated by Desai et al. (2020), this study uses the treated workover data (Section 4) to obtain parametric regression models to predict the workover duration, as explained earlier in Section 3.2. Two types of regression are tested and evaluated: GLMs and ridge regression models.

To test for a setting with the better fitting of the regression models, some transformations of the well i workover duration, when served by rig k (d_i^k), were considered. Specifically, a logarithmic scale ($log(d_i^k)$) and a normalization ($\frac{d_i^k - \min(d_i^k)}{\max(d_i^k) - min(d_i^k)}$) were applied to the data. Finally, alternative settings for the regression modes were considered. For example, GLMs were tested using Gaussian and Gamma distributions, and ridge regression (RR) models were tested using Gaussian and Poisson distributions.

Using the testing procedure described in Section 5, all combinations of response variables to predict the workover duration were exhaustively tested for each of these regression settings. The best results for each regression model and setting are presented in Table 4. The labels generated with the data treatment and clustering are represented by the field $Clusters^{45}$, where each task description is associated with one of these 45 clusters. The other independent variables are the data fields described in Table 3. The column "R^2" is the adjusted R-squared for the regression; "RMSE" refers to the root-mean-squared deviation; "MAE" refers to the mean absolute error. The subscripts *in* and *out* refer to in-sample and out-of-sample, respectively. Last, the column "p-value" refers to the hypothesis that the errors of the regression estimation for the duration are normally distributed.

Analyzing Table 4, we can observe that all the best-performing regressions use data related to the well (i) with some data from the rig. Attributes such as *Basin* (the basin in which the well is associated) and *RigType* (the type of rig used) are important dependent variables selected in all the best regressions. The smaller clusters ($Clusters^{45}$) resulting from the text mining and grouping (Section 4) were also a common attribute in most of the regression models, which indicates that the techniques were successful in revealing the underlying task description. As expected, the number of independent variables is smaller in the ridge regression as this technique penalizes the models for an excess of size and dummy variables. The best-fitted model was the ridge regression using a logarithmic duration for the workover ($log(d_i^k)$). The Gaussian distribution has a good adjusted R^2 (slightly lower than using the Poisson distribution) and a better p-value for a normal distribution for the errors, suggesting that it would be easier to fit distributions for them. Therefore, we have chosen to work with the duration log as

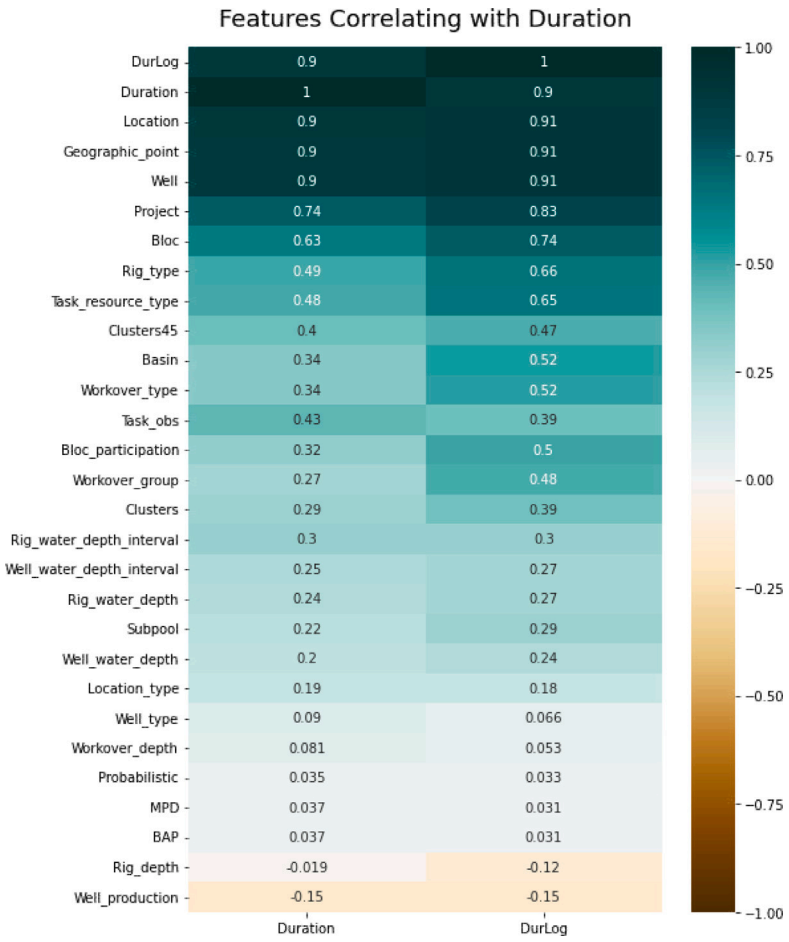


Fig. 4. Associations between features and the workover duration and its logarithmic scale.

Table 4
Best results for the regressions models using Caret package.

#	Method	Dist.	Variable	R^2_{in}	R^2_{out}	$RMSE_{in}$	$RMSE_{out}$	MAE_{in}	MAE_{out}	*p*-value
1	GLM	Gaussian	*Duration*	0.47	0.25	6.7	8.0	5.38	6.38	0.00
2	**GLM**	**Gaussian**	***DurLog***	**0.59**	**0.47**	**0.4**	**0.5**	**0.32**	**0.40**	**0.15**
3	GLM	Gaussian	*DurScale*	0.47	0.25	0.7	0.9	0.59	0.69	0.00
4	GLM	Gaussian	*DurSqrt*	0.52	0.35	0.8	1.0	0.63	0.75	0.74
5	GLM	Poisson	*Duration*	0.47	0.21	6.7	8.2	5.41	6.49	0.30
6	GLM	Poisson	*DurLog*	0.57	0.44	0.4	0.6	0.33	0.41	0.59
7	GLM	Poisson	*DurSqrt*	0.52	0.33	0.8	1.0	0.63	0.77	0.03
8	GLMNET	Gaussian	*Duration*	0.32	0.30	7.6	7.7	6.37	6.33	0.03
9	GLMNET	Gaussian	*DurLog*	0.46	0.46	0.5	0.5	0.38	0.41	0.15
10	GLMNET	Gaussian	*DurScale*	0.32	0.30	0.8	0.8	0.69	0.69	0.30
11	GLMNET	Gaussian	*DurSqrt*	0.38	0.38	0.9	0.9	0.75	0.76	0.00
12	GLMNET	Poisson	*Duration*	0.33	0.30	7.5	7.7	6.36	6.34	0.00
13	GLMNET	Poisson	*DurLog*	0.46	0.46	0.5	0.5	0.38	0.41	0.00
14	GLMNET	Poisson	*DurSqrt*	0.38	0.38	0.9	0.9	0.75	0.76	0.09

a dependent variable (the 2th row from Table 4) that has the largest adjusted R^2, lowest RMSE, and significant *p*-value (greater than 0.05). This results in the following Eq. (3) obtained via the generalized linear regression model:

$$\begin{aligned} log\left(d_i^k\right) \sim & (Intercept) + \beta_1 WellDepth_i + \beta_2 Subpool_i \\ & + \beta_3 Basin_i + \beta_4 Cluster_i^{45} + \\ & \beta_5 LocationType_i + \beta_6 Prob_i + \beta_7 BAP_i + \beta_8 clusters_i \\ & + \beta_9 WorkoverGroup_i + \\ & \beta_{10} WorkoverType_i + \beta_{11} WellWaterDepth_i \\ & + \beta_{12} WorkoverRigType_i + \\ & \beta_{13} BlocSharehold_i + \beta_{14} WorkoverRigType_i \\ & + \beta_{15} BlocSharehold_i + \\ & \beta_{16} RigDepth^k + \beta_{17} RigWaterDepth^k + \beta_{18} RigType^k + \varepsilon, \end{aligned} \quad (3)$$

where d_i^k is the duration of the well i workover performed by rig k, $WellDepth_i$ is the depth of the well i, $Subpool_i$ represents the subpool responsible for the well i, $Basin_i$ refers to the exploratory basin where

the well i is located, $Cluster_i^{45}$ is the cluster for the descriptions of the operation executed in the well i (obtained using k-means for $k = 45$), $RigType^k$ indicates the rig k type, and ε is the residual or error of the regression.

Using this regression, Eq. (3) can be rewritten and simplified to the following linear regression.

$$d_i^k \sim e^{Intercept + WellEffect_i + RigEffect^k} + \varepsilon = \hat{d}_i^k + \varepsilon = \tilde{d}_i^k, \quad (4)$$

where d_i^k is the actual duration of workover i in rig k, $WellEffect_i = \beta_1 WellDepth_i + \beta_2 Subpool_i + \beta_3 Basin_i + \beta_4 Cluster_i^{45} + \beta_5 LocationType_i + \beta_6 Prob_i + \beta_7 BAP_i + \beta_8 clusters_i + \beta_9 WorkoverGroup_i + \beta_{10} WorkoverType_i + \beta_{11} WellWaterDepth_i + \beta_{12} WorkoverRigType_i + \beta_{13} BlocShareholid_i + \beta_{14} WorkoverRigType_i$ and $RigEffect^k = \beta_{16} RigDepth^k + \beta_{17} RigWaterDepth^k + \beta_{18} RigType^k$. Finally, $\tilde{d}_i^k$ is its approximation, $\hat{d}_i^k$ is its prediction from the regression, i.e.,

$$\hat{d}_i^k = e^{Intercept + \alpha WellData_i + \beta RigData^k},$$

and the distribution of ε can be estimated using the regression residuals.

The following section describes the use of the workover data treated in Section 4 and the workover duration estimated in this section to optimize the workover rig schedule.

6. Optimization models

As mentioned in the literature review in Section 2, several formulations have been proposed for the rig scheduling problem. Costa and Ferreira Filho (2004, 2005) proposed models using a time-indexed formulation for the WRSP, consisting of the first formulations for the WRSP. The authors used routing elements to define the sequence in which the rigs serve the wells and scheduling rules to determine when each workover is performed. Although it was a time-index formulation, the model proposed in Costa and Ferreira Filho (2004, 2005) had several routing elements, such as flow balance constraints to ensure the correct sequencing of workover activities in each rig. Their objective function aimed to minimize oil production. As a result, this formulation was easily adapted for this WRSP study, removing the time-index elements and modifying it to a routing formulation with release dates for the operations, rig hiring costs, and the selection of which wells to serve as part of the WRSP.

Costa and Ferreira Filho (2004, 2005) did not consider any release date for the workover activities, so a new constraint for the release date was created. Their objective function was to minimize oil production loss only, and all wells were required to be served. We modified the objective function to consider the rig hiring costs and a penalty for not performing a workover in a well. Furthermore, we added a fictional depot node 0, in which all hired rigs must start their "routes" and return to it at the end of the scheduling horizon. Despite being a routing model, the travel times between the wells were considered to be negligible. However, the formulation can be easily adapted to a workover rig routing and scheduling problem (WRRSP) if the context requires it. This new model, its objective function, and its constraints are presented below. In addition, its sets, parameters, and variables are detailed in Appendix D.

The objective function (5) minimizes the total cost. The first two terms represent the oil production loss, which can be associated with the time until the execution of the task after it is released (first term) or the production loss from the entire time horizon (since the well is released) when the well is not served (second term). The last term of the objective function is related to the fleet size cost.

$$\text{Min} \quad \sum_{i \in J | i \neq 0} l_i \left[S_j + \sum_{j \in J} \sum_{k \in K} (\hat{d}_i^k - a_i) X_{ij}^k + (H - a_i)\left(1 - \sum_{j \in J} \sum_{k \in K} X_{ij}^k\right) \right] + \sum_{k \in K} c^k Z^k \quad (5)$$

$$\text{Subject to} \quad \sum_{j \in J} X_{ji}^k = \sum_{j \in J} X_{ij}^k \qquad \forall i \in J, k \in K \quad (6)$$

$$\sum_{k \in K} \sum_{i \in J} X_{ij}^k \leq 1 \qquad \forall j \in J | j \neq 0 \quad (7)$$

$$\sum_{k \in K} \sum_{j \in J} X_{ij}^k \leq 1 \qquad \forall i \in J | i \neq 0 \quad (8)$$

$$S_j - \hat{d}_i^k \geq S_i - M(1 - X_{ij}^k) \qquad \forall i \in J, j \in J, k \in K | i \neq 0 \quad (9)$$

$$S_i \geq a_i \sum_{k \in K} \sum_{j \in J} X_{ij}^k \qquad \forall i \in J | i \neq 0 \quad (10)$$

$$\sum_{j \in J} X_{ij}^k \leq Z^k \qquad \forall i \in J, k \in K \quad (11)$$

$$X_{ij}^k \in \{1, 0\} \qquad \forall i \in J, j \in J, k \in K | i \neq j \quad (12)$$

$$S_i \in \mathbb{Z}^+ \qquad \forall i \in J | i \neq 0 \quad (13)$$

$$Z^k \in \{1, 0\} \qquad \forall k \in K. \quad (14)$$

Constraints (6), (7), and (8) are flow balance rules from the vehicle routing formulation, where the last two constraint guarantees that a well i or j can only be served once. Constraints (9) calculate each task j starting time (S_j) according to the previous service of the rig ($S_i + d_i^k$). Notice that the dependence between workover duration and the allocated rig is represented by the index k in the parameter $\hat{d}_i^k$. The actual duration of workover activity i is then given by $\sum_{j \in J} \hat{d}_i^k X_{ij}^k$. However, in constraint (9), we can remove X_{ij}^k, which we noticed to make the linear formulation stronger. Constraints (10) guarantee that the task i starting time (S_i) respects its release date (a_i). Constraints (11) connect variables Z^k and X_{ij}^k, forcing the model to hire a rig (Z^k) to execute a task i with this rig k. The other constraints (12), (13), and (14) are related to the variables' domains. Note that this model could be easily adapted to a WRRSP by simply adding the duration of the travels between well i and j using rig k with the duration of the intervention in well j ($d_{ij}^{k'} = d_{ij}^k + d_j^k$) and replacing it in the model, more specifically in Eqs. (5) and (9). Next, we show how we have reformulated the model (5)–(14) to achieve better computational performance.

6.1. Reformulated workover rig scheduling problem model

Aiming to improve the performance of the WRSP model, we propose a reformulation adding new auxiliary variables hoping to help the branching process of the MILP solver employed. The additional auxiliary variables required are detailed in Appendix D. Their use aims to avoid summations inside the constraints, which can then improve the linear programming relaxation of the problem. The objective function terms were equivalently reformulated with the auxiliary variables. As shown in Eq. (15), it minimizes the total costs associated with the oil production losses and the fleet size cost.

$$\text{Min} \quad \sum_{i \in J | i \neq 0} l_i \left[S_i + \sum_{k \in K} (d_i^k - a_i) X1_i^k + (H - a_i)(1 - W_i) \right] + \sum_{k \in K} c^k Z^k \quad (15)$$

$$\text{Subject to:} \quad X1_i^k = X2_i^k \qquad \forall i \in J, k \in K \quad (16)$$

$$X1_i^k = \sum_{j \in J} X_{ji}^k \qquad \forall i \in J, k \in K \quad (17)$$

$$X2_i^k = \sum_{j \in J} X_{ij}^k \qquad \forall i \in J, k \in K \quad (18)$$

$$W_i = \sum_{k \in K} X1_i^k \qquad \forall i \in J | i \neq 0 \quad (19)$$

$$W_i = \sum_{k \in K} X2_i^k \qquad \forall i \in J | i \neq 0 \quad (20)$$

$$S_i - d_j^k \geq S_j - M(1 - X_{ij}^k) \qquad \forall i \in J, j \in J, k \in K | i \neq j \quad (21)$$

$$S_i \geq a_i W_i \qquad \forall i \in J | i \neq 0 \quad (22)$$

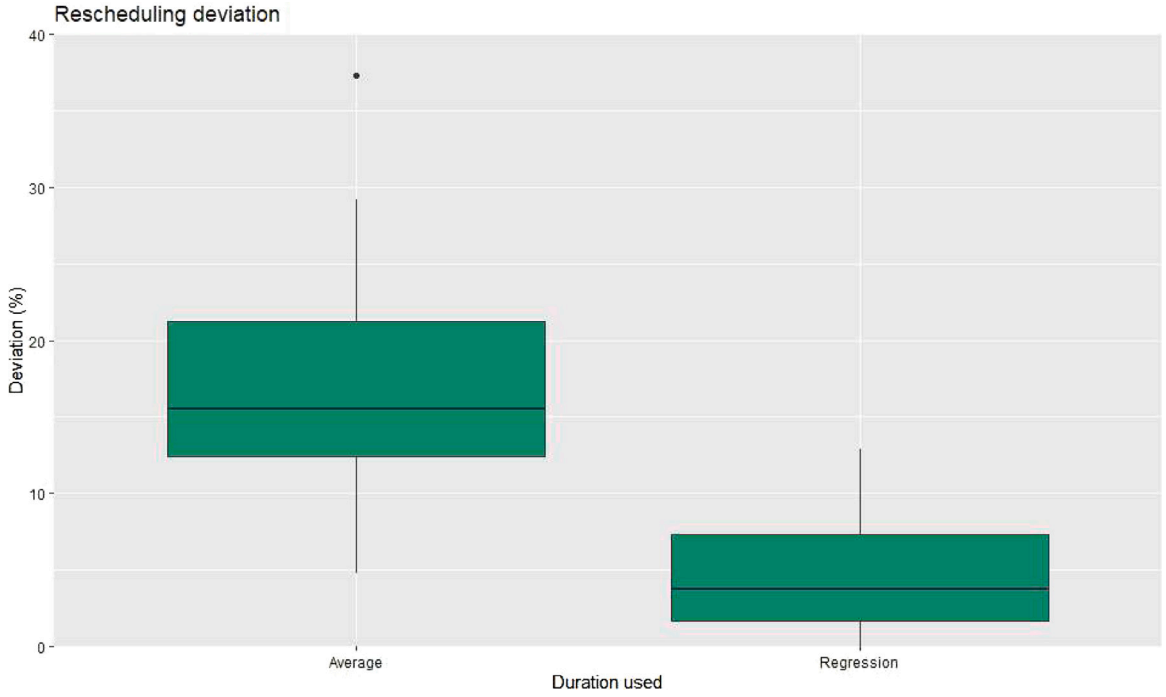


Fig. 5. Box-plot comparing the objective function difference to the schedule with perfect information according to the duration estimation.

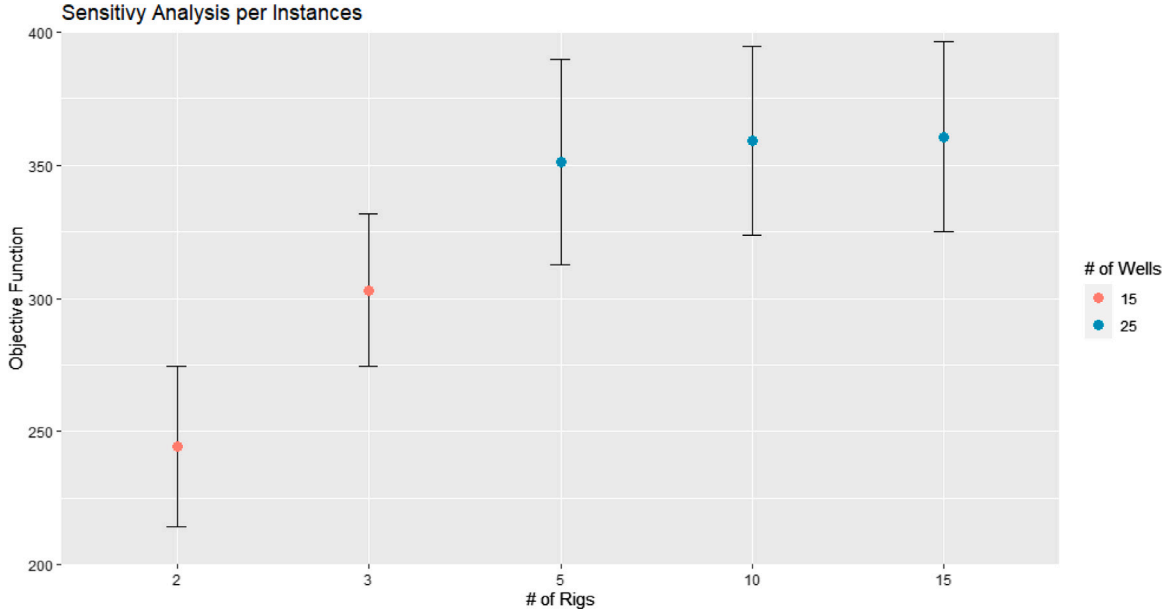


Fig. 6. Confidence intervals of the objective function due to the regression error.

$$X1_i^k \leq Z^k \qquad \forall i \in J, k \in K \quad (23)$$

$$X_{ij}^k \in \{1,0\} \qquad \forall i \in J, j \in J, k \in K | i \neq j \quad (24)$$

$$X1_i^k \in \{1,0\} \qquad \forall i \in J, k \in K \quad (25)$$

$$X2_i^k \in \{1,0\} \qquad \forall i \in J, k \in K \quad (26)$$

$$W_i \in \{1,0\} \qquad \forall i \in J | i \neq 0 \quad (27)$$

$$S_i \in \mathbb{Z}^+ \qquad \forall i \in J \quad (28)$$

$$Z^k \in \{1,0\} \qquad \forall k \in K \quad (29)$$

New constraints were added to define the auxiliary variables and simplify the equations. Constraints (16) are flow balance rules. The new auxiliary variables ($X1_i^k$, $X2_i^k$, and W_i) are defined in constraints (17), (18), (19), and (20), and they guarantee that a well i can only be served once. Constraints (21) calculate the starting time (S_i) of each task i according to the previous service of the rig ($S_i + d_i^k$). Constraints (22) guarantee that the task i starting time (S_i) satisfies its release date (a_i). Constraints (23) connect variables Z^k and $X1_i^k$, forcing the model to hire a rig (Z^k) in order to execute a task i with this rig k ($X1_i^k$). The other constraints (24) to (29) state the domains of the variables.

6.2. Computational experiments

To test the proposed data-driven optimization methodology for the workover rig scheduling problem, data from a major Brazilian oil company were gathered and structured. A total of 74 real-life based

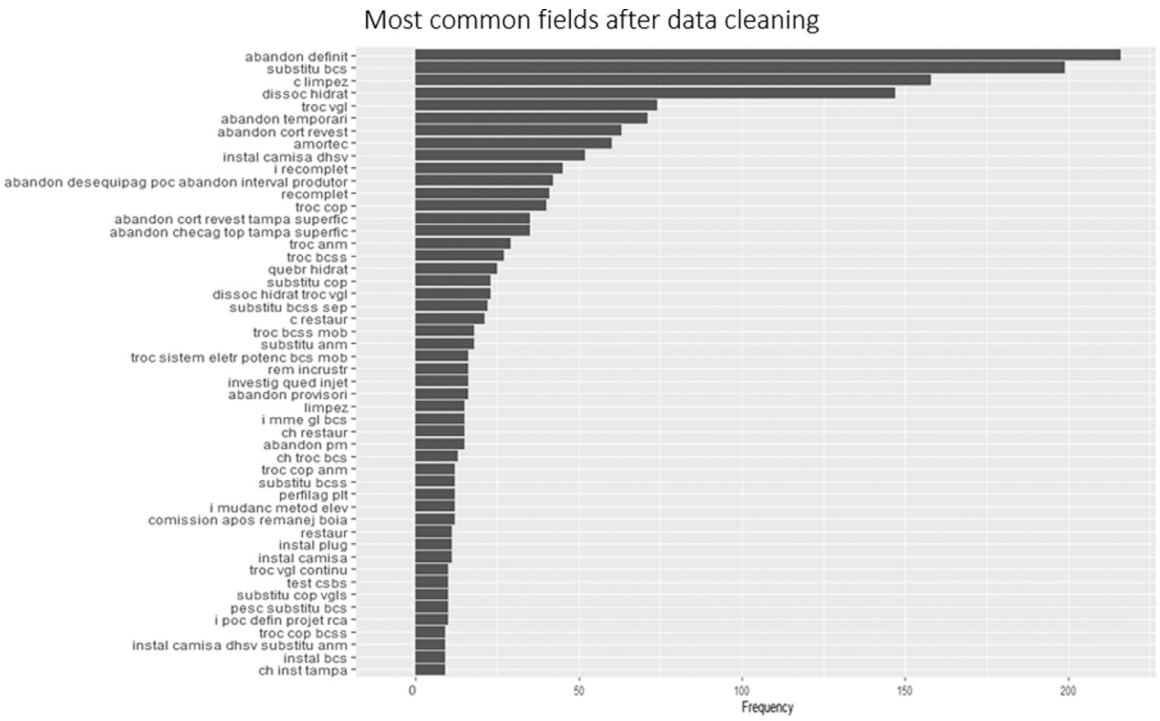


Fig. 7. Bar-plot of the most common fields in the treated task description.

instances were created based on these data. A detailed description of the instance generator is provided in Appendix E. Instances in this study vary according to the number of rigs (2, 3, 5, 10, and 15), the number of wells (15, 25, 50, and 75), the release date density (0.1, 0.5, and 0.9), and the random seed used for drawing numbers and replicating the instance. These instances were used to compare the two formulations, analyze their robustness and the impact of the regression error on the mathematical models, and compare the trade-off between the proposed data-driven model and the current technique used by the company. The computational experiments were performed in a computer with Intel® Core ™ i7-8565U CPU and a 20.0 GB RAM memory. The models were implemented using the Julia programming language (Bezanson et al., 2012) and optimized with Gurobi solver v. 9.1.2 (Gurobi Optimization, 2018).

Table 5 presents a solution comparison between both models, the original model (I) and the reformulated model (II), for different instances with a scheduling horizon of 360 days and using the different seeds (1019, 2657, and 3229) in the instance generator. The terms "UB" and "LB" are acronyms for "Upper Bound" and "Lower Bound", respectively, both in million (M) dollars. A time limit of 3600 s was also enforced to solve the models.

The results in Table 5 show the gap and the computational time difference between the two mathematical models. Model I (the original formulation) requires, in most instances, a longer time than Model II (the reformulation with auxiliary variables) to obtain optimal solutions. In the larger instances, both models started reaching the 3600-second time limit, but the GAPs from the original model are consistently higher than those from the reformulated model. These results indicate that, despite the more significant number of constraints and variables in the reformulated model, the auxiliary variables reduce the computational effort required and enable the model to obtain better solutions.

Another important analysis is to compare the proposed data-driven optimization methodology with the current approach of the company. As mentioned in Section 3.1, the company uses the average duration based only on the type of workover. Using instances generated with out-of-sample records, the reformulated model was used to generate optimal schedules according to a given duration. Three types of duration are used: the average planning duration ($\bar{d}_i$), which is the current strategy used by the studied company and does not consider any information about the rig; the regression estimation ($\hat{d}_i^k$), which is the proposed strategy using the data-driven model and depends on the rig and the well; the actual duration of the well i ($\tilde{d}_i$), which was obtained from the out-of-sample historical data (as the optimization cannot guarantee that the rig performing the workover is the same from historical records, this duration is not influenced by the rig in this case).

Aiming to analyze the robustness and the flexibility of the model's solution, i.e., the capacity to accomplish what was planned by the model, we performed the following experiment. We considered the actual workover duration for each well i ($\tilde{d}_i$), the rig fleet, and the list of served wells obtained when performing the schedule using estimated workover durations (average duration, $\bar{d}_i$, as currently done by the studied company, or $\hat{d}_i^k$, estimated using the proposed data-driven method). This analysis emulates the process of planning the workover resources beforehand in terms of defining which rigs will be hired and how contracts (i.e., which wells are to be served by the hired rigs) are designed in advance. This comparison is presented in Fig. 5. The vertical axis in Fig. 5 is the percentage of deviation of each comparison in terms of the objective function value.

Clearly, the solutions using the duration estimation through the proposed regression model are closer to the "best possible" solutions (obtained with the actual duration of the workover) than the solutions generated with the current approach of the studied company (average duration). Furthermore, the regression solutions fit better with the real duration of the workover activities, as the rescheduling not only is closer to the "best possible", but also varies much less than the solutions using the average duration.

6.3. Sensitivity analysis

In the previous section, the robustness of the data-driven optimization model was tested against the non-data-driven model. Undoubtedly, the data-driven approach generates solutions closer to the "best possible" solutions (obtained with the actual duration of the workover) than the traditional method. Nonetheless, this solution is dependent on

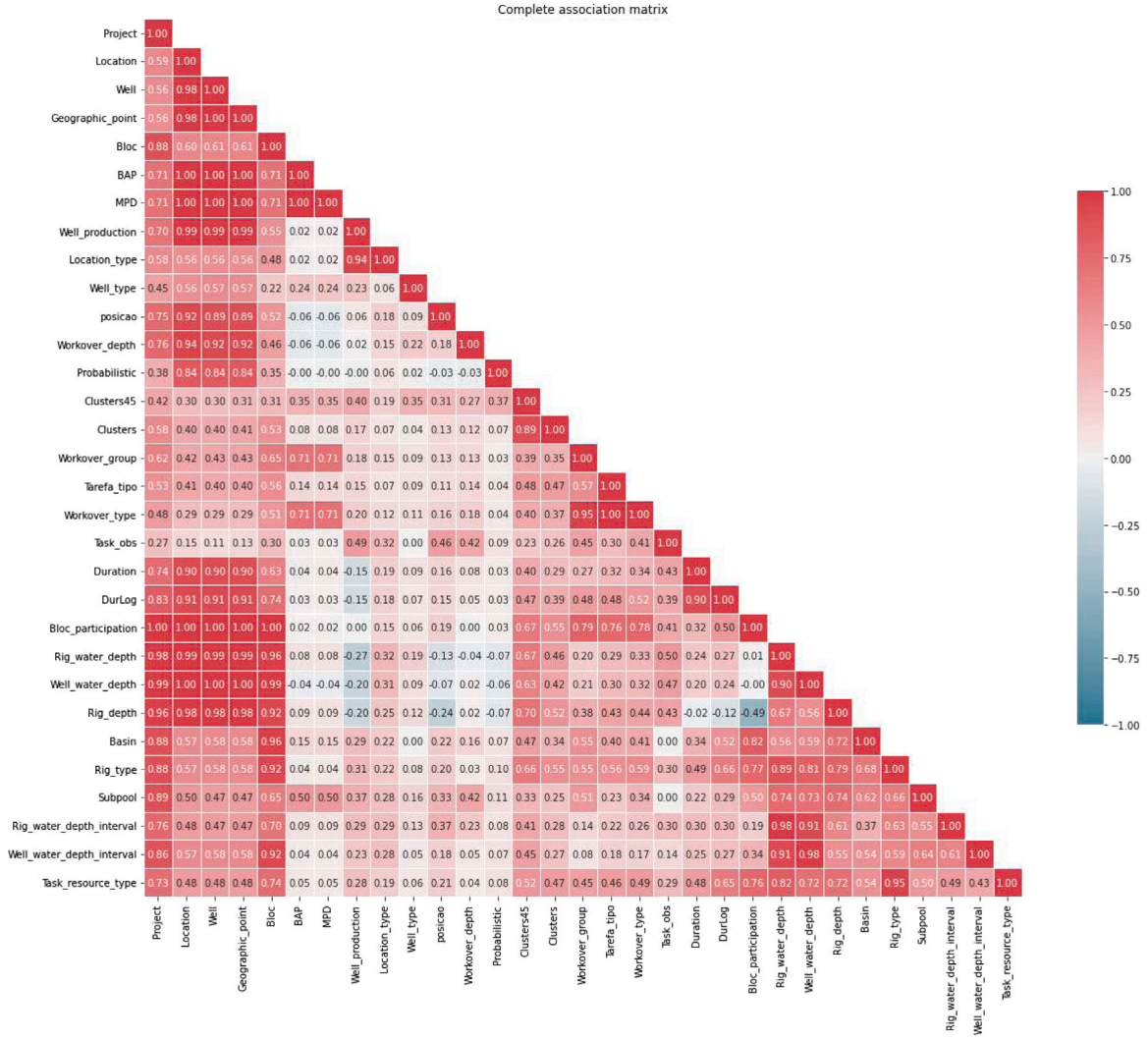


Fig. 8. Associations between features and the workover duration and its logarithmic scale.

the quality of the regression model selected. The duration predictions can vary due to the error associated with the regression, which might impact the data-driven model results.

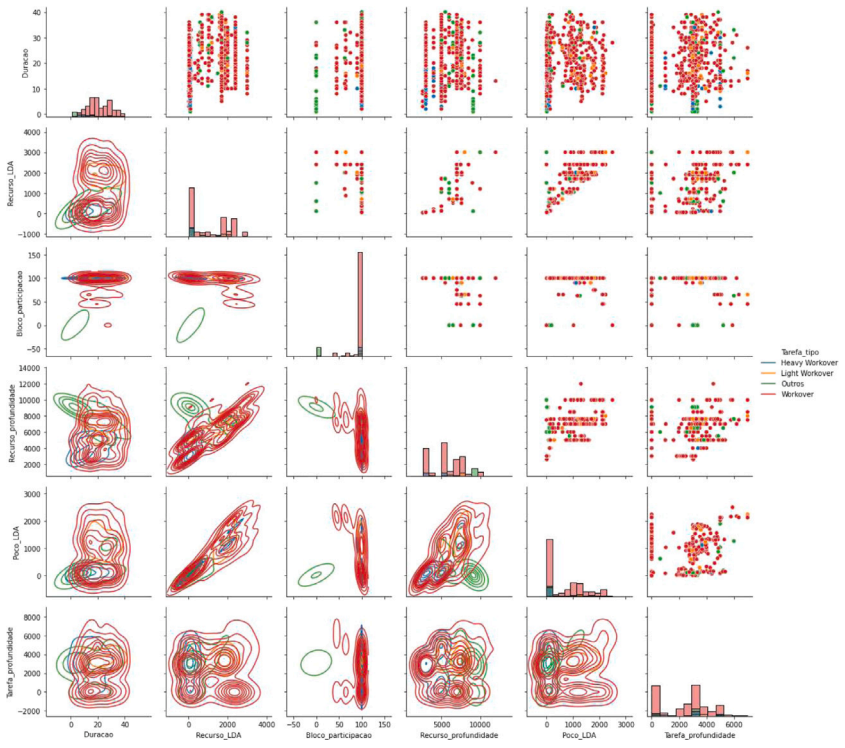
To check the impact of the regression error component on the objective function value of the data-driven model, a sensitivity analysis was performed, simulating the workover duration estimated by the regression. As mentioned earlier, the regression estimation and the actual duration of the workover differ from each other according to regression error, i.e., $\tilde{d_i^k} = \hat{d_i^k} + \epsilon$, where $\tilde{d_i^k}$ is the actual workover duration, $\hat{d_i^k}$ is the regression estimation, and ϵ represents the regression error, which follows a normal distribution estimated as $N \sim (\mu = 1.054672, \sigma = 7.438810)$. In this sensitivity analysis, the WRSP is optimized using the duration estimated by the regression ($\hat{d_i^k}$). The solution of each optimal schedule is fixed in the number of rigs, and the wells that can be attended to and 500 simulations of the regression error (ϵ) are made by sampling from the normal distribution $N \sim (\mu = 1.054672, \sigma = 7.438810)$, to determine the actual duration of the workover ($\tilde{d_i^k}$) for each simulation. With this duration using the regression error, a reschedule is generated according to the rigs and wells selected in the first schedule. The rescheduled solutions are used to obtain a confidence interval for our data-driven optimization model. Fig. 6 presents this sensitivity analysis for each instance according to the number of rigs (horizontal axis), and the number of wells (color labels). The objective function is given by the markers' relative position on the vertical axis, and the error bar represents the confidence interval of this objective function calculated using the t-score ($t_{fracalpha2,N-1}$), where alpha is 5%, and N is the sample size of 500 replications.

Analyzing Fig. 6, we can observe that the objective function and its variability are highly influenced by the instance size. The larger the number of wells needing intervention, the larger the costs associated, as expected. The number of rigs is also important; a small number of rigs reduces the solution flexibility, and when the number of rigs is sufficiently large, increasing the selection of available rigs allows the model to select cheaper and better rigs, reducing the costs.

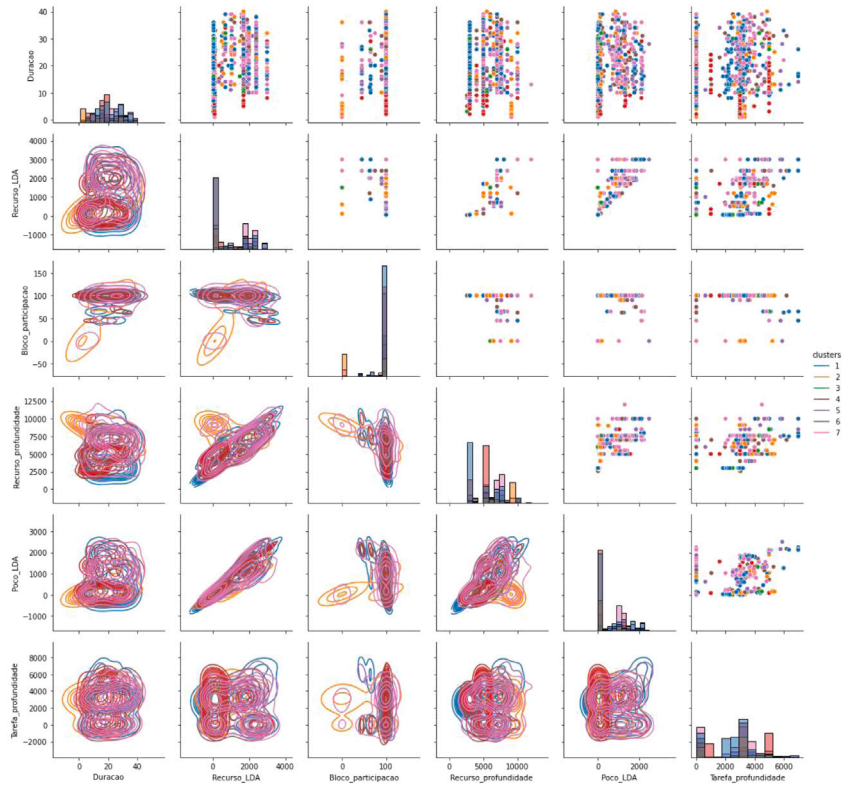
Overall, these results indicate that, with 95% of confidence, the decision maker will not observe losses greater than 15%, and that is for the most uncertain case, which despite providing some comfort in terms of reliability, does indicate potential benefits from additional uncertainty mitigation measures.

7. Conclusions

Oil rigs are an expensive and scarce resource used in critical oil-and-gas production operations, such as workover activities, which are intervention services made on the wells to recover productivity or correct oil flow losses. To support the planning of these operations, we studied a real-life workover rig scheduling problem (WRSP) in which a company needs to decide which wells must be served by which rig in the scheduling horizon, minimizing the costs associated with hiring the rigs and the oil production loss of the wells waiting for



(a)



(b)

Fig. 9. Pair plots for workover type (a) and clustering (b) features.

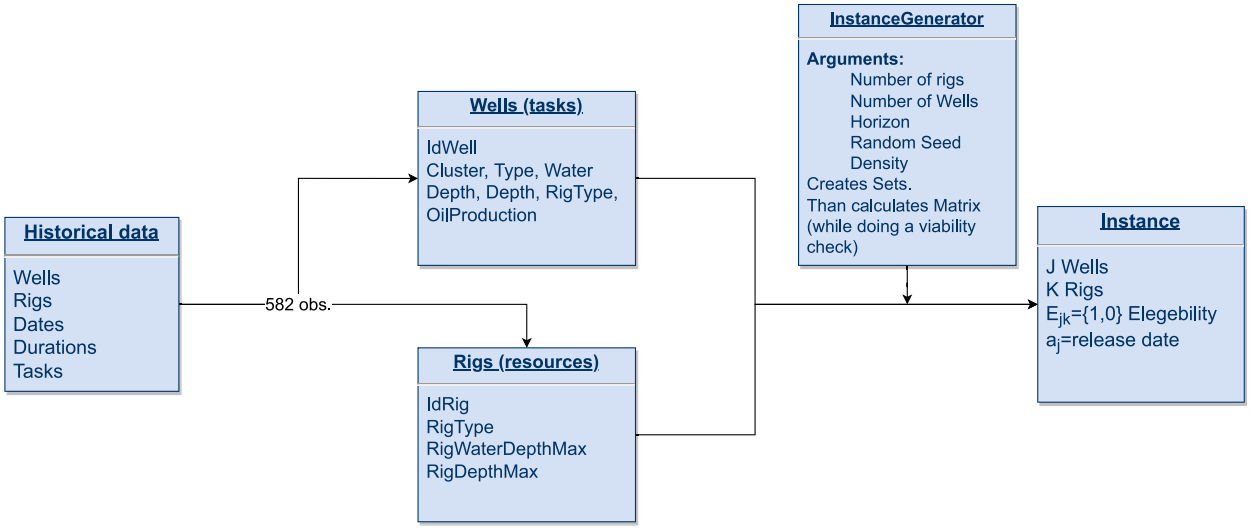


Fig. 10. Data structure and instance generation frameworks.

Table 5
Comparison between models (I) and (II).

Jobs	Rigs	Instance density	UB (avg. in M.)		LB (avg. in M.)		GAP (avg.)		Time (avg. in s)	
			I	II	I	II	I	II	I	II
15	2	0.1	222.5	222.5	222.5	222.5	0%	0%	1	0
		0.5	220.2	220.2	220.2	220.2	0%	0%	1	0
		0.9	221.5	221.5	221.5	221.5	0%	0%	2	0
	3	0.1	258.8	258.8	258.8	258.8	0%	0%	0	0
		0.5	275.7	275.7	275.7	275.7	0%	0%	0	0
		0.9	291.7	291.7	291.7	291.7	0%	0%	0	0
20	5	0.1	338.5	338.5	338.5	338.5	0%	0%	1	0
		0.5	359.1	359.1	359.1	359.1	0%	0%	2	0
		0.9	371.9	371.9	371.9	371.9	0%	0%	1	0
	10	0.1	290.0	290.0	290.0	290.0	0%	0%	1	0
		0.5	302.1	302.1	302.1	302.1	0%	0%	1	1
		0.9	310.7	310.7	310.7	310.7	0%	0%	1	1
	15	0.1	307.1	307.1	307.1	307.1	0%	0%	1	1
		0.5	314.2	314.2	314.2	314.2	0%	0%	1	1
		0.9	314.7	314.7	314.7	314.7	0%	0%	1	1
25	5	0.1	308.0	308.0	308.0	308.0	0%	0%	1	0
		0.5	313.6	313.6	313.6	313.6	0%	0%	2	0
		0.9	319.1	319.1	319.1	319.1	0%	0%	5	2
	10	0.1	320.5	320.5	320.5	320.5	0%	0%	2	0
		0.5	323.9	323.9	323.9	323.9	0%	0%	2	1
		0.9	325.0	325.0	325.0	325.0	0%	0%	2	1
	15	0.1	317.0	317.0	317.0	317.0	0%	0%	3	1
		0.5	327.8	327.8	327.8	327.8	0%	0%	2	1
		0.9	333.9	333.9	333.9	333.9	0%	0%	5	1
50	5	0.1	457.2	457.2	448.2	454.9	2%	1%	1231	1204
		0.5	448.8	448.8	448.8	448.8	0%	0%	22	3
		0.9	447.7	447.7	447.7	447.7	0%	0%	46	11
	10	0.1	416.4	416.4	416.4	416.4	0%	0%	74	72
		0.5	413.6	413.6	413.6	413.6	0%	0%	22	6
		0.9	417.5	417.5	417.5	417.5	0%	0%	44	29
	15	0.1	412.9	412.9	412.9	412.9	0%	0%	288	257
		0.5	414.9	414.9	414.9	414.9	0%	0%	23	10
		0.9	413.6	413.6	413.6	413.6	0%	0%	69	42
75	5	0.1	494.6	494.6	479.8	484.9	3%	2%	3600	3600
		0.5	481.6	481.6	481.6	481.4	0%	0%	1179	1232
		0.9	475.7	475.7	475.7	475.7	0%	0%	228	69
	10	0.1	630.0	495.9	455.9	478.3	21%	3%	3600	3600
		0.5	762.3	486.0	456.1	479.0	32%	1%	3600	2441
		0.9	472.5	472.4	472.5	472.4	0%	0%	1076	59
	15	0.1	484.1	484.3	469.4	475.6	3%	2%	2452	1503
		0.5	474.7	474.7	474.7	474.7	0%	0%	367	133
		0.9	465.6	465.6	465.6	465.6	0%	0%	147	27

workover service. This oil company made available historical data on rig schedules and needs to accurately predict the duration of the new workover activities using the new rig.

Two mathematical programming formulations are developed, and a data-driven optimization methodology is proposed for the WRSP. Basically, the proposed approach is separated into data treatment (cleaning, simplification, and labeling of the workover historical data through text mining and clustering), predictive models (estimation of the workover duration according to a well and a rig using historical records), and optimization (definition of an optimal workover rig schedule using MILP).

Computational experiments made on real-life-based instances showed the superior performance of a reformulated version of the optimization model (II), obtaining better or equal objective function in 122 of 126 instances when compared with the results of the initial model (I). Furthermore, the use of the auxiliary variables in Model II improves the formulation of the model and, as a result, reduces the computational effort.

Regarding the proposed data-driven approach, the text mining and clustering procedures proved to be an efficient way of labeling historical data and acquiring hidden information. The combination of these data science techniques with the regression model improved the prediction of the workover duration, which is currently poorly estimated by the studied company. As a result, the solution of the proposed data-driven optimization methodology obtained solutions much closer to the "perfect" schedule (optimal schedule with the actual duration) than the schedules generated with the company's current methodology. The proposed approach achieves solutions with a deviation of less than 10% and therefore requires considerably less rescheduling. Meanwhile, the current approach employed by the company usually has deviations of 40 to 80%, requiring more frequent rescheduling. This indicates how well the regression model can represent the uncertain workover duration and its dependency on the rig allocation, which in turn leads to more stable and reliable schedules.

Nevertheless, the R^2 values of the selected predictor are still in the order of 0.5, which should prompt further efforts to improve the prediction accuracy. This can be achieved by, e.g., investigating further feature engineering strategies. Furthermore, alternative prediction methods can be tested and compared with the current data-driven methodology, such as gradient-boosted trees and random forest or support vector clustering.

Every regression model has an error associated with the regression residuals, i.e., the difference between the estimation and the actual value of the predicted variable. A sensitivity analysis performed by simulating the regression error showed a low deviation from the objective function, demonstrating that the proposed data-driven optimization methodology is suitable for the problem. Nonetheless, the uncertainty embedded in the duration estimation could be explored in future studies. For example, stochastic programming could be used to consider scenarios and optimize the best average solution, a direction that the authors intend to pursue in future research. Another possibility is to use robust optimization and minimize worst-case scenarios or to use chance-constrained programming to control the level of feasibility of the schedules, potentially reducing the need for rescheduling. Last, the proposed data-driven methodology could be applied to similar scheduling problems.

CRediT authorship contribution statement

Iuri Martins Santos: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Visualization, Writing. **Silvio Hamacher:** Conceptualization, Validation, Supervision, Writing. **Fabricio Oliveira:** Conceptualization, Software, Validation, Investigation, Writing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgment

The authors thank the anonymous reviewers and the associate editor for their insightful suggestions, which helped us to improve this paper. Iuri Martins Santos acknowledges the support from CAPES (Financial code 001) and Tecgraf. Silvio Hamacher acknowledges the support from FAPERJ (E-26/2-1.138/2021) and CNPq (310940/2019-2). The authors also thank Professor Leonardo Bastos from PUC-Rio for the discussions and suggestions on how to improve the performance of the regression models.

Appendix A. Portuguese stopwords

The Table 6 presents the complete list of Portuguese stopwords used for the data cleaning process.

Appendix B. Example of cleaned and simplified data

Fig. 7 shows the data simplified after cleaning and simplifying the text, removing stopwords and using the stemming technique (adapted for the Portuguese language).

It can be seen that the data cleaning process has successfully simplified the terms and made the count of the most common descriptions more accurate.

Appendix C. Feature engineering

To support the selection of the features, a correlation matrix was used to observe the relationship between features and avoid redundance. Fig. 4 presents the correlation or strength-of-association between the features in the data-set, using Pearson's R for continuous–continuous cases, correlation ratio for categorical–continuous cases, and Cramer's V for categorical–categorical cases (see Fig. 8).

Another crucial visual analysis in the feature selection was the pair plot, which help to observe multi-dimension relationships. Two important pair plot that are presented in Fig. 9

Finally, Tables 7 and 8 present a resume and some brief statistic of the qualitative and quantitative data, respectively:

Appendix D. Sets, parameter, and variables of the mathematical models

This new model and its sets, parameters, variables, objective function, and constraints are presented next.

Sets:

- $i, j \in \{1, 2, \ldots, J\}$: workover wells (each well represents a single workover task). Well 0 represents a fictional depot node.
- $k \in \{1, 2, \ldots, K\}$: rigs (resources or machines) that are available for hire.

Parameters:

- a_i: release date for workover well i.
- l_i: costs associated with the oil production loss of well i. Equal to the product of the oil price and the oil flow rate in well i. (US$/day)

Table 6
Stopword (in Portuguese) removed from the text.

"de", "a", "o", "que", "e", "do", "da", "em", "um", "para", "é", "com", "não", "uma", "os", "no", "se", "na", "por", "mais", "as", "dos", "como", "mas", "foi", "ao", "ele", "das", "tem", "à", "seu", "sua", "ou", "ser", "quando", "muito", "há", "nos", "já", "está", "eu", "também", "só", "pelo", "pela", "até", "isso", "ela", "entre", "era", "depois", "sem", "mesmo", "aos", "ter", "seus", "quem", "nas", "me", "esse", "eles", "estão", "você", "tinha", "foram", "essa", "num", "nem", "suas", "meu", "às", "minha", "têm", "numa", "pelos", "elas", "havia", "seja", "qual", "será", "nós", "tenho", "lhe", "deles", "essas", "esses", "pelas", "este", "fosse", "dele", "tu", "te", "vocês", "vos", "lhes", "meus", "minhas", "teu", "tua", "teus", "tuas", "nosso", "nossa", "nossos", "nossas", "dela", "delas", "esta", "estes", "estas", "aquele", "aquela", "aqueles", "aquelas", "isto", "aquilo", "estou", "está", "estamos", "estão", "estive", "esteve", "estivemos", "estiveram", "estava", "estávamos", "estavam", "estivera", "estivéramos", "esteja", "estejamos", "estejam", "estivesse", "estivéssemos", "estivessem", "estiver", "estivermos", "estiverem", "hei", "há", "havemos", "hão", "houve", "houvemos", "houveram", "houvera", "houvéramos", "haja", "hajamos", "hajam", "houvesse", "houvéssemos", "houvessem", "houver", "houvermos", "houverem", "houverei", "houverá", "houveremos", "houverão", "houveria", "houveríamos", "houveriam", "sou", "somos", "são", "era", "éramos", "eram", "fui", "foi", "fomos", "foram", "fora", "fôramos", "seja", "sejamos", "sejam", "fosse", "fôssemos", "fossem", "for", "formos", "forem", "serei", "será", "seremos", "serão", "seria", "seríamos", "seriam", "tenho", "tem", "temos", "tém", "tinha", "tínhamos", "tinham", "tive", "teve", "tivemos", "tiveram", "tivera", "tivéramos", "tenha", "tenhamos", "tenham", "tivesse", "tivéssemos", "tivessem", "tiver", "tivermos", "tiverem", "terei", "terá", "teremos", "terão", "teria", "teríamos", "teriam".

Table 7
Quantitative data statistics and summary.

Feature	Count	Mean	Standard Dev.	Min	25%	50%	75%	Max
Probabilistic	582	0.0	0.1	0	0	0	0	1
Duration	582	19.9	9.2	1	13	20	27	40
Rig water depth	582	1000.5	999.3	31	88	555	1900	3000
Rig depth	582	5776.5	2051.8	2600	4000	5000	7500	12 000
Bloc participation	582	90.3	26.7	0	100	100	100	100
Well water depth	582	630.2	682.6	0	80	142	1158.75	2496
Workover depth	582	2546.8	1719.9	0	1000	3000	3444.75	7000
BAP	55	0.0	0.1	0	0	0	0	1
Well production	582	0.9	0.1	0.55	0.9	0.9	0.9	0.9

Table 8
Qualitative data statistics and summary (some names were modified to protect sensitive data).

Feature/Stats	Count	Unique values	Most frequent	Most frequent (count)
Task_id	582	582	*10 076*	1
Project name	582	133	*XX-MPXD*	27
Location name	582	393	*XX-5Y*	6
Well name	582	386	*7-XX-5Y-XXX*	6
Workover group	582	6	*Workover*	531
Workover type	582	4	*Workover*	470
Workover description	582	299	*Substituição de BCS*	92
Resource name	582	74	*P-XX*	50
Resource type	582	3	*SS/NS - Sonda flutuante (semissub ou navio-sonda)*	298
Basin	582	7	*CAMPOS MAR*	395
Bloc	582	53	*CARAPEBA*	42
Subpool	582	11	*SSE*	221
Location type	582	7	*ES*	487
Task resource type	582	4	*SS/NS - Sonda flutuante (semissub ou navio-sonda)*	228
Geographic point	582	388	*7-XX-5Y-XXX*	6
Well type	582	4	*Não definido*	559
Rig water depth interval	582	4	0–700 km	315
Well water depth interval	582	4	0–700 km	334
Workover description (cleaned)	582	269	*substitu bcs*	93
Clusters	582	7	*1*	158
Clusters45	582	40	*43*	108

- e_i^k : binary matrix indicating if rig k is eligible to serve well i.
- d_i^k : duration of the intervention in well i using rig k (in days). The processing time of any rig in the fictional depot node 0 is equal to 0.
- c^k : hiring cost of rig k. (US$/rig)
- H : scheduling horizon (in days).

Variables:

- X_{ij}^k: binary variable that indicates if rig k goes from well (workover task) i to well j.
- S_i: integer variable equal to the starting time of task i in days.
- Z^k: binary variable representing if rig k is hired (used) or not.

Auxiliary variables (only for the reformulated model):

- $X1_i^k$ and $X2_i^k$: respectively, if a rig k arrives at or leaves from well j.
- W_i: if any rig serves (enters and leaves) well i.

Appendix E. Instance generator

The instance generator uses historical data to generate instances for the optimization, as described in Fig. 10

Based on Wigwe et al. (2020), the wells' oil production (in *bbl*, barrels) were generated randomly according to their type, using the Gamma distribution, as follows in Eqs. (30) and (31):

$$p_i = Scale_i^{Operation} \cdot Scale_i^{Well} \cdot p_i^0 \tag{30}$$

$$p_i^0 \sim 10^3 \cdot \Gamma\left(\alpha = 2.3, \beta = 4.2\right), \tag{31}$$

where p_i is the loss of oil production from the well, $Scale_i^{Operation}$ and $Scale_i^{Well}$ are a parameter that makes the oil production loss proportional to the operation type and well type (respectively), and p_i^0 are the random oil production generated using the Gamma distribution

Table 9
Proportional scales of oil production according to the type of well and operation.

$Scale_i^{Operation}$		$Scale_i^{Well}$	
Operation type	Value	Well type	Value
Drilling	1	Producer	1
Workover	0.8	Injector	0.8
Appraisal	0.4	Exploratory	0.3
Abandonment	0.3	Other	0.6

$\Gamma(\alpha = 2.3, \beta = 4.2)$, in which α and β are the shape and the scale of the distribution (respectively). Fernández Pérez et al. (2018) suggested using the price of the oil barrel as 55 $ / barrel. As a result, the oil production loss cost l_i in dollars is equal to $55 \cdot p_i$. Details on the proportional scales values, $Scale_i^{Operation}$ and $Scale_i^{Well}$, are provided in Table 9:

The rig hiring and operation costs were randomly selected from the Markit (2021) database, which has historical information of the rig average day rates according to the type of rig and the market.

Using the wells and rigs data sets, an instance generation algorithm was developed, such that it creates instances for a desirable number of rigs, wells, scheduling horizon, random seed, and density coefficient (represented by ρ). The random seed is a number used to initialize the random number generator and to allow the reproduction of the instance. As to the density coefficient, it is a setting parameter between 0 and 1 that controls the release dates of the workover. A small rho (ρ) tends to result in latter release dates, reducing the feasible windows of the tasks. However, a large ρ would generate smaller release dates, increasing the window of allocation of the workover. The algorithm selects random samples of the set to generate instance sets and parameters. With the sets and parameters selected, the algorithm calculates an eligibility matrix that indicates which rigs from the sample set can serve the sample wells. This eligibility matrix is calculated according to the rig data (the type of rig, the rig's maximum water depth, and the rig's maximum depth) and the well data (the well's water depth, the well's depth, and the rig type that can attend the well). A rig will only be able to serve a well if the well's maximum water depth, depth, and type are within the rigs specifications. During the construction of the eligibility matrix, the algorithm checks the feasibility of the instance, that is, if there is a rig for every well and if all wells have a rig to serve it. In case of infeasibility, new samples are calculated until a feasible instance is found, outputting this instance to the data-driven models.

References

Abu-Marrul, V., Martinelli, R., Hamacher, S., 2020. Scheduling pipe laying support vessels with non-anticipatory family setup times and intersections between sets of operations. Int. J. Prod. Res. 1–15.

Aloise, D.J., Aloise, D., Rocha, C.T.M., Ribeiro, C.C., Ribeiro Filho, J.C., Moura, L.S.S., 2006. Scheduling workover rigs for onshore oil production. Discrete Appl. Math. 154, 695–702.

Aronofsky, J.S., 1962. Linear programming a problem-solving tool for petroleum industry management. J. Pet. Technol. 14 (7), 729–736.

Aronofsky, J.S., Williams, A.C., 1962. The use of linear programming and mathematical models in under-ground oil production. Manage. Sci. 8 (4), 394–407.

Barnes, J.W., Brennan, J.J., Knap, R.M., 1977. Scheduling a backlog of oilwell workovers. J. Pet. Technol. 29, 1651–1653.

Bassi, H.V., Ferreira Filho, V.J.M., Bahiense, L., 2012. Planning and scheduling a fleet of rigs using simulation-optimization. Comput. Ind. Eng. 63, 1074–1088.

Bengio, Y., Grandvalet, Y., 2004. No unbiased estimator of the variance of k-fold cross-validation. J. Mach. Learn. Res. 5 (Sep), 1089–1105.

Bezanson, J., Karpinski, S., Shah, V.B., Edelman, A., 2012. Julia: A fast dynamic language for technical computing. arXiv preprint arXiv:1209.5145.

Bissoli, D.C., Chaves, G.L.D., Ribeiro, G.M., 2016. Drivers to the workover rig problem. J. Pet. Sci. Eng. 139, 13–22.

Chaudhuri, U.R., 2011. Fundamentals of Petroleum and Petrochemical Engineering. Crc Press Boca Raton.

Cochrane, J.E., 1989. Rig performance monitoring and measurement: can it again be useful? In: Proceedings of the SPE/IADC Drilling Conference. New Orleans, United States of America, pp. 597–608.

Costa, L.R., Ferreira Filho, V.J.M., 2004. Uma heurística para o problema do planejamento de itinerários de sondas em intervenções de poços de petróleo. In: Proceedings of the XXXVI Brazilian Symposium on Operations Research. São João del Rei, Brazil, pp. 1844–1853.

Costa, L.R., Ferreira Filho, V.J.M., 2005. Uma heurística de montagem dinâmica para o problema de otimização de itinerários de sondas. In: Proceedings of the XXXVII Brazilian Symposium on Operations Research. Gramado, Brazil, pp. 2176–2187.

Danach, K., 2016. Hyperheuristics in Logistics (Ph.D. thesis). Ecole Centrale de Lille, Lille, France.

de Andrade Filho, A.C.B., 1994. Optimal Scheduling of Development in an Oil Field (Master's thesis). Stanford University, Stanford, USA.

Desai, J.N., Pandian, S., Vij, R.K., 2020. Big data analytics in upstream oil and gas industries for sustainable exploration and development: A review. Environ. Technol. Innov. 101186.

Duhamel, C., Santos, A.C., Guedes, L.M., 2012. Models and hybrid methods for the onshore wells maintenance problem. Comput. Oper. Res. 39 (12), 2944–2953.

Eagle, K., 1996. Using simulated annealing to schedule oil field drilling rigs. Interfaces 26, 35–43.

Engebretsen, S., Bohlin, J., 2019. Statistical predictions with glmnet. Clin. Epigenetics 11 (1), 1–3.

Fernández Pérez, M.A., Oliveira, F., Hamacher, S., 2018. Optimizing workover rig fleet sizing and scheduling using deterministic and stochastic programming models. Ind. Eng. Chem. Res. 57 (22), 7544–7554.

Gurobi Optimization, L., 2018. Gurobi optimizer reference manual.

Hebbali, A., Hebbali, M.A., 2017. Package 'olsrr'.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, Vol. 112. Springer.

Jivani, A.G., et al., 2011. A comparative study of stemming algorithms. Int. J. Comp. Tech. Appl 2 (6), 1930–1938.

Kaiser, M.J., Snyder, B., 2013. The five offshore drilling rig markets. Mar. Police 39, 201–214.

Khor, C.S., Elkamel, A., Shah, N., 2017. Optimization methods for petroleum fields development and production systems: a review. Opt. Eng. 18 (4), 907–941.

Kromodihardjo, S., Kromodihardjo, E.S., 2016. Modeling of well service and workover to optimize scheduling of oil well maintenance. Appl. Mech. Mater. 836, 311–316.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R.C., et al., 2020. Package 'caret'. R J. 223, 7.

Lang, D.T., 2004. Word stemming in R. Department of Statistics, UC Davis.

Lasrado, V.K., 2008. Workover rig scheduling using reservoir simulation. In: Proceedings of the Intelligent Energy Conference and Exhibition, Vol. 1. (February), Amsterdam, Netherlands, pp. 39–49.

Likas, A., Vlassis, N., Verbeek, J.J., 2003. The global k-means clustering algorithm. Pattern Recognit. 36 (2), 451–461.

Mahmoud, H.F., 2019. Parametric versus semi and nonparametric regression models. arXiv preprint arXiv:1906.10221.

Markit, I., 2021. Petrodata offshore rig day rate trends. Accesed 10 February 2021 https://www.ihs.com/products/oil-gas-drilling-rigs-offshore-day-rates.html.

Marques, L.C., Machado, F.A.P.P., Oliveira, F.C., Hamacher, S., 2014. Sizing and scheduling resources: a decision support system applied To oil rig scheduling. In: Proceedings of the XLVI Brazilian Symposium on Operations Research. Salvador, Brazil, pp. 2538–2547.

McCullagh, P., Nelder, J.A., 2019. Generalized Linear Models. Routledge.

McDonald, G.C., 2009. Ridge regression. Wiley Interdiscip. Rev. Comput. Stat. 1 (1), 93–100.

Monemi, R.N., Danach, K., Khalil, W., Gelareh, S., Lima, F.C., Aloise, D.J., 2015. Solution methods for scheduling of heterogeneous parallel machines applied to the workover rig problem. Expert Syst. Appl. 42 (9), 4493–4505.

Nelder, J.A., Wedderburn, R.W., 1972. Generalized linear models. J. R. Stat. Soc.: Ser. A (General) 135 (3), 370–384.

Osmundsen, E., Roll, K.H., Tveterås, R., 2010. Exploration drilling productivity at the Norwegian shelf. J. Pet. Sci. Eng. 73, 122–128.

Paiva, R.O., Bordalo, S.N., Schiozer, D.J., 2000. Optimizing the itinerary of workover rigs. In: Proceedings of the 16th World Petroleum Congress. Calgary, Canada, pp. 11–15.

Pérez, M., Oliveira, F., Hamacher, S., 2016. A new mathematical model for the workover rig scheduling problem. Pesquisa Operacional 36 (2), 241–257.

Pittman, J., 1985. Computer speeds offshore well planning, rig scheduling. Oil Gas J. 83, 84–97.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0, URL http://www.R-project.org/.

Ribeiro, G.M., Desaulniers, G., Desrosiers, J., 2012a. A branch-price-and-cut algorithm for the workover rig routing problem. Comput. Oper. Res. 39 (12), 3305–3315.

Ribeiro, G.M., Desaulniers, G., Desrosiers, J., Vidal, T., Vieira, B.S., 2014. Efficient heuristics for the workover rig routing problem with a heterogeneous fleet and a finite horizon. J. Heuristics 20 (6), 677–708.

Ribeiro, G.M., Laporte, G., Mauri, G.R., 2012b. A comparison of three metaheuristics for the workover rig routing problem. European J. Oper. Res. 220 (1), 28–36.

Ribeiro, G.M., Mauri, G.R., Lorena, L.A.N., 2011. A simple and robust Simulated Annealing algorithm for scheduling workover rigs on onshore oil fields. Comput. Ind. Eng. 60 (4), 519–526.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Santos, I.M., Hamacher, S., Oliveira, F., 2021. A Systematic Literature review for the rig scheduling problem: Classification and state-of-the-art. Comput. Chem. Eng. 153, 107443.

Sarica, S., Luo, J., 2021. Stopwords in technical language processing. Plos One 16 (8), e0254937.

Shaji, N., Sundar, C., Jagyasi, B., Dutta, S., 2019. An aggregated rank removal heuristic based adaptive large neighborhood search for work-over rig scheduling problem. In: Deka, B., Maji, P., Mitra, S., Bhattacharyya, D., Bora, P., Pal, S. (Eds.), Pattern Recognition and Machine Intelligence. PReMI 2019. Lecture Notes in Computer Science, Vol. 11941. Springer, Cham, Switzerland, pp. 385–394.

Shcherbakov, M., Shcherbakova, N., Brebels, A., Janovsky, T., Kamaev, V., 2014. Lean data science research life cycle: A concept for data analysis software development. In: Joint Conference on Knowledge-Based Software Engineering. Springer, pp. 708–716.

Silva, F.T., Silva, R.P., 2018. Roteamento dinâmico de sondas de intervenção para otimização da prodção de poços de petróleo: um modelo matemático para o PRSI dinâmico. Braz. J. Prod. Eng. 4, 169–184.

Srnka, K.J., Koeszegi, S.T., 2007. From words to numbers: how to transform qualitative data into meaningful quantitative results. Schmalenbach Bus. Rev. 59 (1), 29–57.

Street, J.O., Carroll, R.J., Ruppert, D., 1988. A note on computing robust regression estimates via iteratively reweighted least squares. Amer. Statist. 42 (2), 152–154.

Sun, Y., Ma, L., Wang, S., 2015. A comparative evaluation of string similarity metrics for ontology alignment. J. Inf. Comput. Sci. 12 (3), 957–964.

Tavallali, M.S., Karimi, I.A., 2014. Perspectives on the design and planning of oil field infrastructure. Comput. Aided Chem. Eng. 34, 163–172.

Tavallali, M.S., Karimi, I.A., Baxendale, D., 2016. Process systems engineering perspective on the planning and development of oil fields. AIChE J. 62 (8), 2586–2604.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.

Tozzo, E., Costa, A.P.S., Lins, I.D., 2020. A hybrid multi-objective genetic algorithm for scheduling heterogeneous workover rigs on onshore oil fields. J. Pet. Sci. Eng. 195, 107935.

Vasconcelos, D., Nogueira, E., Sousa, S., Charrouf, R., 2017. A solution to optimize the logistics of a fleet of workover vessels applied to offshore operations in the Gulf of Mexico. In: Proceedings of the OTC Brasil 2017. Rio da Janeiro, Brazil, pp. 1705–1713.

Wigwe, M.E., Bougre, E.S., Watson, M., Giussani, A., 2020. Comparative evaluation of multi-basin production performance and application of spatio-temporal models for unconventional oil and gas production prediction. J. Petrol. Explor. Prod. Technol. 10 (8), 3091–3110.

Yan, X., Su, X., 2009. Linear Regression Analysis: Theory and Computing. World Scientific.

Yuan, Z., Yang, Y., 2005. Combining linear regression models: When and how? J. Amer. Statist. Assoc. 100 (472), 1202–1214.

Yujian, L., Bo, L., 2007. A normalized levenshtein distance metric. IEEE Trans. Pattern Anal. Mach. Intell. 29 (6), 1091–1095.

Zou, H., Zhang, H.H., 2009. On the adaptive elastic-net with a diverging number of parameters. Ann. Statist. 37 (4), 1733.