Pontifícia Universidade Católica
do Rio de Janeiro

Iuri Martins Santos

Data-driven joint chance-constrained optimization for the workover rig scheduling problem

Tese de doutorado

Thesis presented to the Programa de Pós–graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia de Produção.

Advisor: Prof. Silvio Hamacher
Co-advisor: Prof. Fabricio Oliveira

Rio de Janeiro
December 2022

Pontifícia Universidade Católica
do Rio de Janeiro

Iuri Martins Santos

Data-driven joint chance-constrained optimization for the workover rig scheduling problem

Thesis presented to the Programa de Pós–graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia de Produção. Approved by the Examination Committee.

Prof. Silvio Hamacher
Advisor
Departamento de Engenharia Industrial – PUC-Rio

Prof. Fabricio Oliveira
Co-advisor
Aalto University

Prof. Virgílio José Ferreira Filho
UFRJ

Prof. Paulo Cesar Ribas
Molde University College – Specialized University in Logistics

Prof. Leonardo dos Santos Loureço Bastos
Departamento de Engenharia Industrial – PUC-Rio

Prof. Glaydston Mattos Ribeiro
UFRJ

Rio de Janeiro, December 13, 2022

Iuri Martins Santos

Graduated in Industrial Engineering at the Pontifical Catholic University of Rio de Janeiro in 2015 and obtained his M.Sc. Degree in Industrial Engineering at the Pontifical Catholic University of Rio de Janeiro in 2018.

Bibliographic data

Santos, Iuri Martins

Data-driven joint chance-constrained optimization for the workover rig scheduling problem / Iuri Martins Santos ; advisor: Silvio Hamacher ; co-advisor: Fabricio Oliveira. – 2022.
176 f. : il. color. ; 30 cm

Tese (doutorado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial, 2022.
Inclui bibliografia

1. Engenharia Industrial – Teses. 2. Programação de sondas. 3. Otimização sob incerteza. 4. Otimização orientada sobre dados. 5. Restrições probabílisticas conjuntas. I. Hamacher, Silvio. II. Oliveira, Fabricio. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. IV. Título.

CDD: 658.5

Acknowledgments

To Samir Mohd Azzam and Diego Moah. Rest in peace, my dear friends and colleagues.

To my family: my mother Mariela, my father Saulo, my grandmother Therezinha, my sister Maíra, and my nephew Tom.

To my girlfriend, Carol. Thank you for all the support during my weekends and nights studying.

To my friends, Ernani, Sidney, Lucas, Heloísa, Jéssica, Bernardo, Felipe, Eduardo, and Leon, that I had to miss many appointments during my PhD and Master.

To my advisors Prof. Silvio and Prof. Fabricio. Thank you for the help, knowledge, and opportunities. This wouldn't be possible without your support.

To my colleagues, but also friends, from PUC-Rio and Tecgraf, Leonardo (or should I say teacher and R master?), Victor, Luana, João, Janaína, Leila, Bianca, Amanda, Igor, Gabi, Danuza, Davi, Rachel, Lucas (that was also my colleague in Aalto. Thank you for the hospitality), Pierry, and Flávio.

To my teachers and university staff.

Your support and help were crucial for my success.

This Study was financed in part by the Coordenação de Aperfeiçoaento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Other supporters (financially or operationally) for this study were PUC-Rio, Tecgraf Institute, and Petrobras SA.

Abstract

Santos, Iuri Martins; Hamacher, Silvio (Advisor); Oliveira, Fabricio (Co-Advisor). **Data-driven joint chance-constrained optimization for the workover rig scheduling problem**. Rio de Janeiro, 2022. 176p. Tese de Doutorado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Workover rigs are a crucial resource in petroleum exploration and production, used in the wells' maintenance operations. The Workover Rig Scheduling Problem (WRSP) determines which rigs will serve the wells and when the activities will occur. This decision-making problem emerges in a highly uncertain environment, and most literature approaches are based on deterministic models and heuristics. Aiming to assist the WRSP, this thesis proposes a regression-based data-driven (DD) optimization methodology, applying it in real-life-based instances. This DD optimization approach is composed of three phases: data treatment, where text mining and clustering techniques are used to refine and retrieve information from the data; predictive modeling, using ridge regression to estimate the workover duration and the endogenous uncertainties in the model; optimization, where the regression prediction and random error are inserted in the joint chance-constrained (JCC) models, generating solutions more resilient to the uncertainties. We propose a stochastic JCC formulation based on simulation and Wasserstein distance to generate scenarios and reduce the problem size. This model is compared with four alternatives: a non-stochastic DD, a stochastic integrated CC, a stochastic budget-constrained model, and the company's current approach. For small and medium-sized instances, the stochastic JCC model guarantees a feasibility confidence level with an error of approximating lower than 5%. However, the stochastic JCC model does not close the GAP in large instances. For these instances, the non-stochastic DD model is a good alternative with disturbances not greater than 10%. Overall, the DD optimization methodology finds schedules that are more often feasible and with lower costs compared with the company's method.

Keywords

Rig scheduling; Optimization under uncertainty; Data-driven optimization; Joint chance constraints; Regression models

Resumo

Santos, Iuri Martins; Hamacher, Silvio; Oliveira, Fabricio. **Otimização com restrições conjuntas probabilísticas orientada por dados para o problema de programação de sondas de intervenção**. Rio de Janeiro, 2022. 176p. Tese de Doutorado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

As sondas de intervenção são um recurso crucial na exploração e produção de petróleo, sendo utilizadas nas operações de manutenção de poços. As empresas de petróleo planejam quais sondas atenderão os poços. O Problema de Programação de Plataforma de Trabalho (WRSP) determina quais sondas atenderão os poços e quando as atividades ocorrerão. Com o intuito de auxiliar o WRSP, esta tese propõe uma metodologia de otimização orientada por dados (DD) baseada em regressão, aplicando-a em instâncias reais. Essa abordagem de otimização DD é dividida em três fases: tratamento de dados, onde técnicas de mineração de texto e agrupamento são usadas para refinar e recuperar informações dos dados; modelagem preditiva usando regressão de cume para estimar a duração do workover e as incertezas endógenas do modelo; otimização, onde a previsão da regressão e seu erro aleatório são inseridos nos modelos de restrições probabilísticas conjuntas (JCC), gerando soluções mais resilientes às incertezas. Propomos uma formulação estocástica de JCC baseada em simulação e distância de Wasserstein para gerar cenários e reduzir o tamanho do problema. Esse modelo é comparado com quatro alternativas: um DD não estocástico, um CC integrado estocástico, um modelo estocástico com restrição orçamentária e a abordagem atual da empresa. Para instâncias de pequeno e médio porte, o modelo estocástico JCC garante um nível de confiança de viabilidade e um erro de aproximação inferior a 5%. No entanto, o modelo estocástico JCC não fecha o GAP em instâncias maiores. Para essas instâncias, o modelo DD não estocástico é uma boa alternativa com perturbações não superiores a 10%. No geral, a metodologia de otimização DD encontra cronogramas que são mais frequentemente viáveis e com custos menores em comparação com o método da empresa.

Palavras-chave

Programação de sondas; Otimização sob incerteza; Otimização orientada sobre dados; Restrições probabílisticas conjuntas; Modelos de regressão

Table of contents

List of figures

List of tables

List of abbreviations

AC – Ant Colony
AC-PR – Ant Colony with Path-Relinking
ALNS – Adaptive Large Neighborhood Search
AM – Adaptive Memory
ARO – Adjustable Robust Optimization
ARRH – Aggregated Rank Removal Heuristic
ASAS - Automated Services Assignment System
BCS – *Bombeio Centrifugo Submerso*
BPC – Branch-Price-and-Cut
BP -– British Petroleum
BS – Bubble Swap
CAPEX – Capital Expenditure
CART – Classification and Regression Trees
CC – Chance-Constrained or Chance Constraint
CP – Constraint Programming
CPVNS – Cooperative Parallel Variable Neighborhood Search
CS – Clustering Search
CVaR – Conditional Value-at-Risk
DAH – Dynamical Assemble Heuristic
DCC – Data-driven Chance-Constrained
DD – Data-Driven
DD-JCC – Data-Driven Joint Chance-Constrained
DP – Dynamic Programming
DRO – Distributionally Robust Optimization
DRSP – Drilling Rig Scheduling Problem
D-WRRSP – Dynamic Workover Rig Routing and Scheduling Problem
E&P – Exploration and Production
EEV – Expected Value Solution
EOA – Approximation Error or Error of Approximating
ESP – Electrical Submersible Pump
FP – Field Planning
GA – Genetic Algorithm
GDP – Generalized Disjunctive Programming
GH – Greedy Heuristic
GLM – Generalized Linear Model
GRASP – Greedy Randomized Adaptive Search Procedure
HGA – Hybrid-GA
HH – Hyper-Heuristic
ICC – Individual Chance-Constrained or Individual Chance Constraint
ILS – Iterated Local Search
IWLS – Iteratively re-Weighted Least Squares

JCC – Joint Chance-Constrained or Joint Chance Constraint

JCC-WRSP – Joint Chance-Constrained Workover Rig Scheduling Problem

KDE – Kernel Density Estimation

K–L – Kullback–Leibler

kNN – k-Nearest Neighbors

LB – Lower Bound

LCS – Longest Common Substring

LCV – Light Construction Vessel

LHS – Left-Hand Side

LOESS – Locally Weighted Least Squares

LV – Levenshtein

LWIV – Light Well Intervention Vessel

M – Million

MA – Memetic Algorithm

MCS – Monte Carlo Simulation or Monte Carlo Sampling

MILP – Mixed-Integer Linear Programming

MINLP – Mixed-Integer Non-Linear Programming

ML – Machine Learning

MLE – Maximum Likelihood Estimation

MPTH – Maximum Priority Three-criteria Heuristic

NLP – Natural Language Processing

NPV – Net Present Value

OF – Objective Function

OLS – Ordinary Least Squares

OR – Operations Research

OSV – Offshore Support Vessel

OVRP – Open Vehicle Routing Problem

P&A – Plug and Abandonment

PAE – Evolutionary Algorithm for Planning

PDF – Probability Density Function

PLSV – Pipe Laying Support Vessel

PR – Path-Relinking

PRISM – Plan Resource Implement Schedule Manage

R^2 – R-Squared

RHS – Right-Hand Side

RMSE – Root-Mean-Square Error

RP – Resource Planning

RQ – Research Question

RR – Ridge Regressions

RSP – Rig scheduling Problem

SA – Simulated Annealing

SAA – Sample Average Approximation

SLR – Systematic Literature Review

SS – Scatter Search

TA – Transgenetic Algorithm

TS – Tabu Search

UB – Upper Bound

USA – United States of America

VRP – Vehicle Routing Problem
VNS – Variable Neighborhood Search
VND – Variable Neighborhood Descent
WOS – Web of Science
WP – Workover Planning
WRSP – Workover Rig Scheduling Problem
WRRSP –Workover Rig Routing and Scheduling Problem
WS – Wait and See

List of Algorithms

1
Introduction

The exploration and production (E&P) of oil and gas involves several complex, expensive, and high-risk operations (Suslick and Schiozer 2004). One of the most critical phases of the E&P is well construction and maintenance, which relies mainly on oil rigs. Such rigs are typically costly and scarce resources, for which daily rates can vary between US$ 50,000 and US$ 700,000, depending on the rig type, market, and operational specifications (Kaiser and Snyder 2013, Osmundsen et al. 2010). Companies generally hire rigs to perform essential well operations, such as drilling, evaluation, completion, and workover. An undersized fleet of rigs can result in oil production delays that affect the profitability of the wells. On the other hand, 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 at the lowest cost possible.

The decision-making comprising the rig scheduling problem (RSP) is highly complex, as it involves numerous widely varied tasks from different phases of the E&P (Tavallali et al. 2016). Operations are subject to risks in geological prospects, economic evaluations, development, and production (Suslick et al. 2009). Moreover, several techniques from interdisciplinary areas can be applied to the RSP (Khor et al. 2017). A large number of studies aiming at supporting decision-making in this context have been proposed using quantitative and analytical methods, such as mathematical programming, heuristics, and data science techniques. However, few studies consider the uncertainties inherent within the RSP, and those that consider it, lack data-driven considerations.

Aiming to fulfill this gap, this thesis addresses a particular case of the RSP, known as the workover rig scheduling problem (WRSP). In this problem, wells require workovers (interventions to correct the oil flow) over the planning horizon. These interventions are performed by oil rigs that need to be hired, and the wells requiring workover have oil production loss associated with the waiting time. The WRSP's goals are to determine the fleet of rigs to be hired, select the wells that will be attended to, and schedule the wells on the rigs (when and by which rigs the wells will be served), minimizing the rig fleet costs and the oil production loss of the wells. These operations are subjected to disruptions, which calls for the use of optimization models under uncertainty to deal with duration variability.

Scheduling and sizing models under uncertainty are frequently subjected to low-probability extreme cases that lead to over-conservative solutions. Chance-constrained models allow some infeasible solutions to occur within a confidence level, enabling robust solutions that are more realistic, considering the uncertainty but less biased to extreme scenarios with low-probability. The chance-constrained optimization, also known as probabilistic programming, is divided into two types: individual and joint chance-constrained models (ICC and JCC, respectively). The first refers to models in which the confidence level is individual for each constraint, and the second, to cases in which multiple constraints must respect a joint probability together. Several linear formulations exist for the ICC models, but most JCC formulations often lead to non-linear programming models. One of the goals of this thesis is to propose a data-driven optimization methodology using regression models to obtain a linear programming representation for joint chance-constrained models.

Data-driven optimization is a new trend in the Operations Research (OR) community that combines data-science, statistics, and machine learning algorithms with optimization models to properly estimate and account for the uncertainty and errors within the data. This thesis proposes a data-driven joint chance-constraint methodology based on text mining, clustering algorithms, regression models, and scenario approximation algorithms, and applies the methodology to the workover rig scheduling problem (WRSP). Our goal is to find answers to the following research questions (RQs):

- *RQ 1.* What trends, gaps, and insights can be observed in the rig scheduling problem literature from the academic and industrial perspectives?
- *RQ 2.* How can the optimization under uncertainty be applied in problems such as the workover rig scheduling problem to mitigate issues emerging from uncertainties in the problem? Which techniques are available and suitable for the problem?
- *RQ 3.* What are the main assumptions of the workover rig scheduling problem from an offshore industry perspective? Which uncertainties affect the WRSP, and how can they be represented in the mathematical modeling?
- *RQ 4.* How can the mathematical programming be used to support the workover rig scheduling problem and address the offshore operational demands? In other words, how can we model the WRSP for a practical

case, *i.e.*, with realistic assumptions, but still implementable in an oil company.

- *RQ 5.* How can we combine chance-constrained models and data-driven optimization to solve large and realistic problems such as the workover rig scheduling problem?

Inspired by the aforementioned research questions, the thesis is divided into seven sections, illustrated in Figure 1.1, where each dashed box is a chapter.

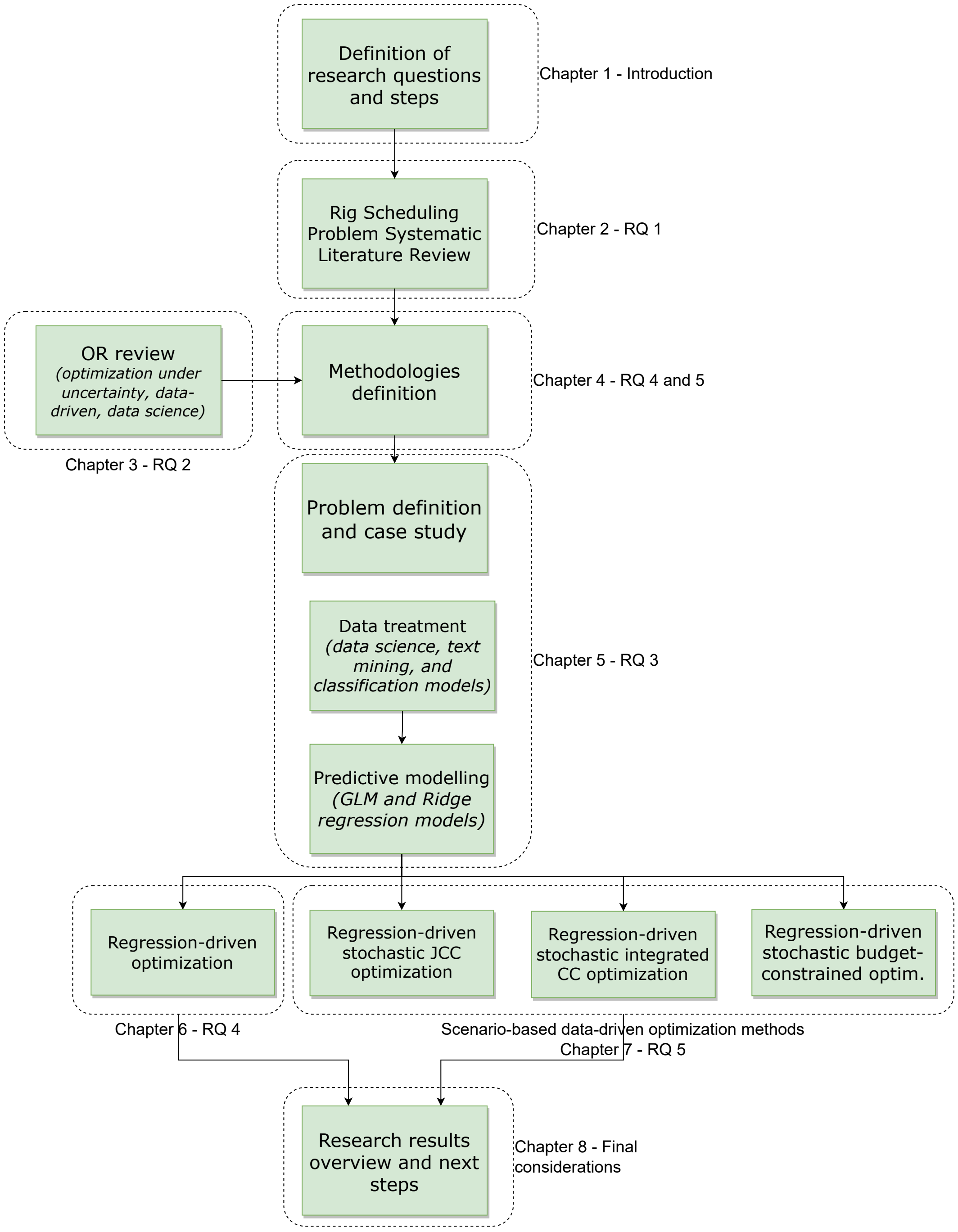

Definition of research questions and steps
Chapter 1 - Introduction
Rig Scheduling Problem Systematic Literature Review
Chapter 2 - RQ 1
OR review (optimization under uncertainty, data-driven, data science)
Chapter 3 - RQ 2
Methodologies definition
Chapter 4 - RQ 4 and 5
Problem definition and case study
Data treatment (data science, text mining, and classification models)
Predictive modelling (GLM and Ridge regression models)
Chapter 5 - RQ 3
Regression-driven optimization
Chapter 6 - RQ 4
Regression-driven stochastic JCC optimization
Regression-driven stochastic integrated CC optimization
Regression-driven stochastic budget-constrained optim.
Scenario-based data-driven optimization methods
Chapter 7 - RQ 5
Research results overview and next steps
Chapter 8 - Final considerations

Figure 1.1: Thesis research structure and main phases.

This thesis begins with this introductory chapter, where the research questions are defined. Chapter 2 addresses the **RQ 1**. Not only a classification and taxonomy for the rig scheduling problem are proposed, but also an extensive and systematic literature review (SLR) on the RSP is performed, analyzing all its dimensions and perspectives. A trend for data-driven optimization models and more realistic assumptions is noticed. Furthermore, some gaps are detected, including the application of more optimization under uncertainty techniques, such as joint chance-constrained formulations, although very few of these studies are being implemented or verified within the industry.

These trends of optimization models under uncertainty and data-driven observed during the SLR contribute to **RQ 2** regarding the available techniques in the Operations Research and statistic literature that could be useful for problems such as the WRSP, and what are their pros and cons. This second question is approached in Chapter 3, where chance-constrained optimization and data-driven techniques are studied. Using these techniques, Chapter 4 proposes different data-driven methodologies for these problems that will also be useful in answering **RQ 4-5**.

Another question raised by the SLR is related to the modeling assumptions and uncertainties of the workover rig scheduling problem that should be considered to meet the industry demands (**RQ 3**). Chapter 5 approaches this research question by describing the offshore workover rig scheduling problem of a Brazilian oil company (*Petrobras SA*) and analyzing the impact of the workover duration disturbances on the schedule. A data-driven methodology is proposed to estimate the uncertainties in Section 5.3.1, which is divided in two parts: uncertainty treatment and joint chance constrained models. First, it presents the text mining, clustering, and regression algorithms used to treat and estimate the uncertainty. The definition of which uncertainties affect the WRSP the most leads to searching for the best ways of representing it.

Aiming to properly represent these uncertainties and assumptions in a mathematical model that addresses the offshore industry demands (**RQ 4**), Chapter 6 applies some of the techniques studied in Chapter 3 in the real-life-based WRSP instances presented in Chapter 5, following the regression-driven optimization methodology from Chapter 4. Two alternatives of mixed-integer linear programming models are proposed and compared with the current methodology of the studied company. A sensitivity analysis with simulation raises the concern of how the workover duration disturbances on the schedule feasibility, motivating a search for a more suitable methodology.

During the optimization under uncertainty review, the data-driven chance-constrained models stood out as a reliable alternative for dealing with solution feasibility uncertainty, which motivates the **RQ 5** of how the data-driven optimization and chance-constraints could be combined to solve realistic problems such as the WRSP. This research question is tackled in Chapter 7, where the data-driven stochastic chance-constrained optimization models proposed in Section 4.1.2.2 are adapted to the offshore WRSP. Four regression-driven models are proposed in this chapter: a joint chance-constrained (JCC) deterministic-equivalent, a stochastic JCC, a stochastic integrated chance-constrained, and a stochastic budget-constrained. From those models, only the regression-driven JCC deterministic-equivalent requires mixed integer non-linear programming (MINLP). The stochastic regression-driven models are mixed-integer linear programming (MILP) reformulations based on scenarios and multi-stage decision variables, using Monte Carlo sampling for the scenario generation and a Wasserstein-distance-based method for the scenario reduction. These different methods are compared with each other and other alternatives mentioned earlier.

Last, Chapter 8 is dedicated to the final considerations of this thesis, overviewing the main findings, analyzing the strengths and weaknesses of the proposed methodology, and suggesting future studies to be performed.

2
Systematic literature review

The lack of a standardized classification for the RSP is a critical shortcoming in the available literature. Against this backdrop, this thesis proposes a new taxonomy and presents an extensive and systematic literature review of the RSP. With the purpose of supporting future studies, this analysis also describes the current gaps in the RSP literature and trends. By assembling this literature review, our goal is to find answers to **RQ1** *"What trends, gaps, and insights can be observed in the rig scheduling problem literature from the academic and industrial perspectives?"*. To classify the different dimensions of the RSP it is important to understand the exploration and production (E&P) of oil and gas and how the rig scheduling problem is inserted in it, Section 2.1.

2.1
The exploration and production of oil and gas

The E&P plays a central role in the decision-making of the supply chain of the oil and gas sector. E&P can take many years, and it is a crucial part of the process for the company's profitability. As mentioned by Baker (1996), IFP School (2015), and Pereira (2005), the E&P can be separated into five main phases:

1. *Discovery phase*: mapping of possible oil fields with geological and seismic studies and the drilling of exploratory wells to confirm the presence of hydrocarbons;

2. *Appraisal phase*: after the oil presence confirmation, delineation wells are drilled and reservoir modeling studies are performed to estimate the properties, size, value of the reserve, and its techno-economic development feasibility;

3. *Development phase*: comprising essential activities and production decisions, including field design (well location and type, surface network and facilities design), field operation planning (well drilling schedule, flow scheduling, rig fleet scheduling, and offshore logistics), and field construction (facilities fabrication and installation, well construction, drilling, and completion);

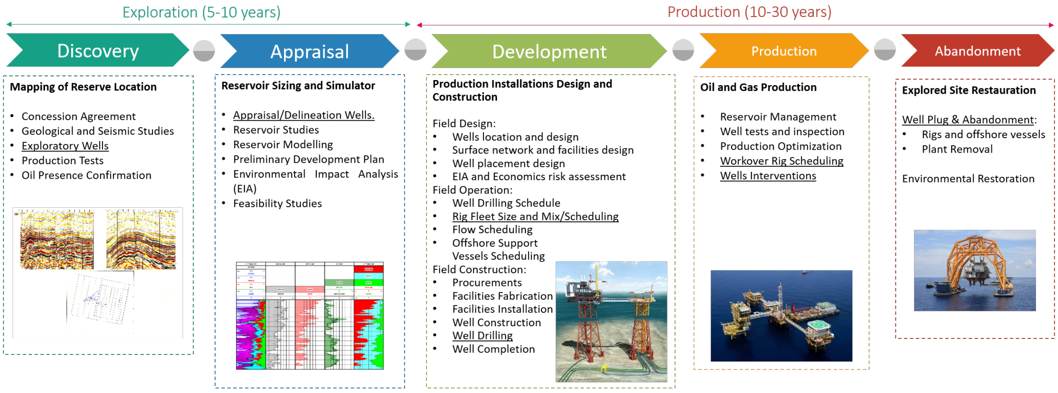

Exploration (5-10 years)
Production (10-30 years)
Discovery
Appraisal
Development
Production
Abandonment
Mapping of Reserve Location
• Concession Agreement
• Geological and Seismic Studies
• Exploratory Wells
• Production Tests
• Oil Presence Confirmation
Reservoir Sizing and Simulator
• Appraisal/Delineation Wells.
• Reservoir Studies
• Reservoir Modelling
• Preliminary Development Plan
• Environmental Impact Analysis (EIA)
• Feasibility Studies
Production Installations Design and Construction
Field Design:
• Wells location and design
• Surface network and facilities design
• Well placement design
• EIA and Economics risk assessment
Field Operation:
• Well Drilling Schedule
• Rig Fleet Size and Mix/Scheduling
• Flow Scheduling
• Offshore Support Vessels Scheduling
Field Construction:
• Procurements
• Facilities Fabrication
• Facilities Installation
• Well Construction
• Well Drilling
• Well Completion
Oil and Gas Production
• Reservoir Management
• Well tests and inspection
• Production Optimization
• Workover Rig Scheduling
• Wells Interventions
Explored Site Restauration
Well Plug & Abandonment:
• Rigs and offshore vessels
• Plant Removal
Environmental Restoration

Figure 2.1: Exploration and Production of oil and gas - phases and decisions. Adapted from Santos (2018), Agarwal et al. (2016), TOYO (2019), NORWEP (2019).

4. *Production phase*: can extend through decades and has many different successive phases to increase productivity, correct oil flow loss, and solve mechanical failures;

5. *Abandonment phase*: when the hydrocarbon production rate becomes economically infeasible and the reservoir is abandoned.

Figure 2.1 describes the aforementioned phases and their main decisions, in which the underlined ones are related to the RSP.

From the oil discovery to the field abandonment, rigs are used to drill, complete, maintain, and abandon wells, as shown in Figure 2.1. In the discovery phase, exploratory wells are drilled using rigs to confirm hydrocarbon presence. Then, in the appraisal phase, delineation wells are drilled to estimate the reserve properties. In the development phase, injector and production wells are drilled and completed also using rigs. In this phase, the rig fleet size is decided, as well as the operation schedules of the rigs. While in the production phase, completed wells often require workover rigs to perform an intervention aiming to increase productivity, correct oil flow losses, and solve mechanical failures. Last, wells must be plugged and abandoned (a process in which wells are abandoned and cannot be used again after the installation of well barriers or plugs), which is often performed using rigs. Each of these operations requires a specific rig type (Bakker et al. 2017). In what follows, we describe the different types of oil rigs and their purpose.

As pointed out earlier, the rigs are one of the primary resources used in the exploration of oil and gas. These highly complex and expensive structures are used in critical activities such as evaluation, drilling, completion, and workover. There are several types of oil rigs, each one with a specific purpose in terms of operations it can perform and its own technical specifications. Figure 2.2 presents the different types of oil rigs.

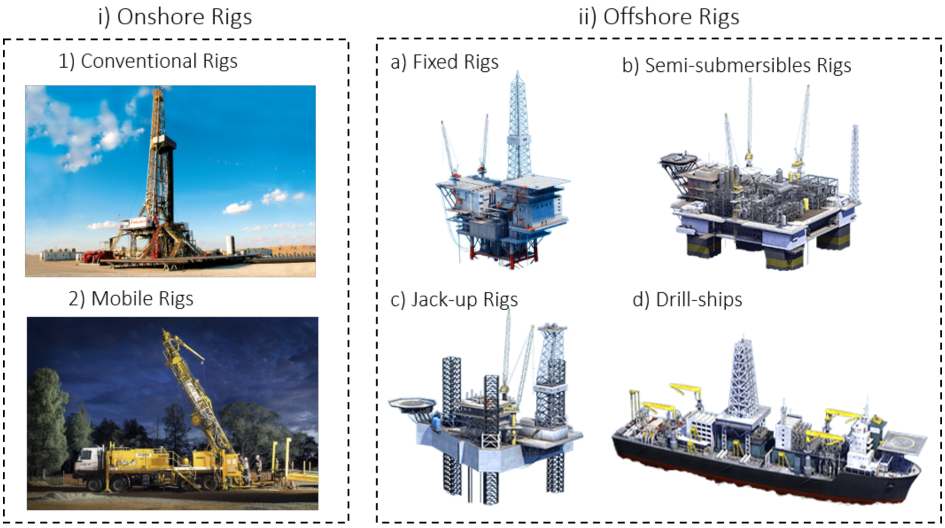


Figure 2.2: Different types of onshore (left) and offshore (right) rigs. Adapted from Santos et al. (2017), Petrobras (2014), IADC (2015), Khodro Diesel (2019).

As shown in Figure 2.2, the main classification is according to the location of the well, which can be onshore or offshore (Al-Azani 2014). According to Bourgoyne et al. (2016), onshore drilling rigs are mainly classified based on their mobility, being either conventional or mobile rigs. The first is a fixed rig in which the derrick is assembled at the drilling location and is the most frequently used land rig type. The second is a mobile rig coupled on wheeled trucks, allowing it to be moved between drilling facilities more efficiently. On the other hand, the most common offshore rigs are: fixed rigs (bottom-supported oil platforms used until 300 meters of water depth); semi-submersibles rigs (floating platforms used up to 2,000 meters of water depth); jack-up rigs (bottom-supported platforms with elevating legs used until 150 meters) and drill-ships (floating platforms constructed in a vessel hull used up to 2,000-meter water depth) (Petrobras 2014, Markit 2019, Al-Azani 2014, Baker 1996). Next, Section 2.2 aims to classify the different dimensions of the RSP under the perspective of available case studies and solution approaches.

2.2 The rig scheduling problem

Generally, an oil and gas company operates many oilfields and wells simultaneously, each one in a different E&P stage. To meet expected production and delivery dates, well's activities must be adequately scheduled and allocated to rigs, aiming to avoid delays and optimize the use of resources (Eagle 1996). This decision-making is called the Rig Scheduling Problem (RSP).

In essence, the RSP considers a set of wells (e.g., onshore or offshore, injector or producer), a set of tasks (drilling, completion, or workover; notice that the terms operation, activity, and job are used interchangeably in this context) that have to be executed on these wells and a set of resources (onshore or offshore rigs) available or to be hired to perform these activities. Usually, the objective is to provide a schedule that minimizes the total costs or the total oil production loss, considering a list of operating and engineering constraints and assumptions (Tavallali et al. 2016).

The RSP can be classified according to several characteristics involving the problem setting (oil field location, well operations, field development integration, resources considered) or the approach (how the problem is addressed and solved: modeling, rig fleet, single/multiple jobs, and type of objective functions). The setting attributes refer to characteristics of the problem that may have implications for the approach, attributes connected with the oil and gas exploration and production. And the approach refers to how the problem is being modeled and solved, attributes related to the technique and modeling. Figure 2.3 presents our taxonomy of the RSP.

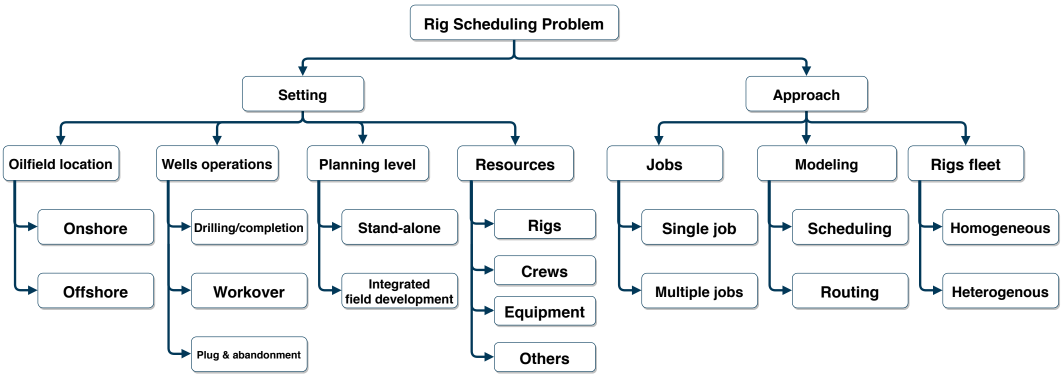


Figure 2.3: Characteristics for the rig scheduling problem.

The first division of the RSP is according to the oilfield location: onshore or offshore. Onshore well operations are generally more straightforward, faster, and less expensive than offshore well operations, for the latter is a more complex and high-risk environment (Bassi et al. 2012). The problems also vary according to the types of rig operations that are scheduled. Some studies consider drilling activities, while others focus on the completion, which are interventions to prepare the well for production after it has been drilled. Also, some papers consider problems with workover activities, which consist of well interventions performed after its completion (Holmager and Redda 2013, IOM3 2015).

Rig scheduling is a multidimensional decision task related to several planning levels of the E&P. Most studies tackle the rig scheduling as an independent decision-making process; we refer to this planning level *stand-alone*. However, some studies have an integrated field development planning, considering other problems such as reservoir modeling, field design, facilities locations, and production scheduling. Similar behavior was noticed by Lasschuit and Thijssen (2004) in planning and scheduling of the downstream of the oil and gas sector. The more decisions are considered, the harder it will be to solve the problem. Rig operations involve multiple resources and, thus, another classification is possible, according to the planned resources. Usually, only rigs are considered in the planning. However, sometimes other resources, such as crews, pieces of equipment, and other vessels, are also considered.

In some cases, the wells are physically close to each other and the operations have processing times that are much longer than the traveling time between the wells, precluding the need of taking into account routing considerations in the planning process. Therefore, the RSP can also be classified according to the employment of routing techniques, dividing the problems into: scheduling-only (or simply scheduling) problems, in which the traveling times between wells are negligible; and routing and scheduling (to which we refer simply as routing hereinafter, presuming that the scheduling activities are implied by context), when the rigs transportation costs are significant and vary between wells, and thus routing decisions have also to be considered (Bissoli et al. 2016). Usually, routing and scheduling problems are classified according to the fleet of vehicles or the set of machines (resources) available (Georgiadis et al. 2019, Eksioglu et al. 2009). The fleet of rigs can be homogeneous when all rigs share the same costs, processing times, technical specifications, and so forth, or heterogeneous, meaning that the rigs possess different characteristics and might be capable of performing specific operations or require particular well conditions. According to the level of complexity of the problem, a well might have more than one operation to be executed, or the rigs may consider operations with different characteristics, as in Hasle et al. (1996) and Fernández Pérez et al. (2018). Therefore, another RSP classification depends on the number of jobs: whether a single job for each well is considered or multiple jobs from the same well are allocated to the rigs.

Another critical attribute is the objective function that will be optimized. The problem can consider a single objective (when there is only one objective to be optimized) or be multi-objective (when there are multiple objectives to be optimized). As to the nature of the objective function, it can be a non-monetary indicator (such as completion time, tardiness, distance traveled, number of rigs used, and wells served), a monetary indicator (total costs, net present value, and cash flow), or an oil-related indicator (oil production, oil production loss, or expected oil recovery).

Analyzing the RSP literature according to the aforementioned characteristics, similarities between the papers were observed, enabling us to gather them in four major classes of problem:

(1) *Drilling Rig Scheduling Problem* (DRSP): refers to drilling and completion rig scheduling problems in a stand-alone planning level, when the rig scheduling is an independent decision from the rest of the field development problems. Drilling and completion usually take place in the development of non-completed (new) wells and the goal is to minimize the rigs fleet costs meeting a due date for the wells to start oil production. Some DRSP problems consider a well to have a single job to be performed while others consider multiple jobs for each well;

(2) *Workover Planning*: problems in which the rigs are used in workover operations. These operations occur in the production phase, when existing and completed wells require maintenance or re-work to be performed. As a result, this decision-making is usually separated from the other rig's decisions and there is only a single job (a single maintenance) to be scheduled for each well. It can be divided according to the use of routing in two sub-groups: workover rig scheduling problems (WRSP) and workover rig routing and scheduling problems (WRRSP);

(3) *Field Planning (FP)*: refers to problems in which the rig scheduling is integrated with other field development decisions, such as reservoir modeling, field design, and production flow scheduling. In these cases, the RSP depends on or influences other decisions related to the oilfield development. For instance, the location of the platforms influences the duration of the drilling and the traveling time between the platforms, which are, in turn, crucial parameters in the RSP. On the other hand, the rig scheduling affects the well production schedule, as the well can only start producing after being drilled and completed. Usually, these problems aim to maximize the net present value (NPV) of the oilfield;

(4) *Resource Planning (RP)*: the rig scheduling problem is integrated with the planning of other resources that also affect the decision or are affected by the RSP, such as offshore support vessels (OSVs), equipment, and crews. For instance, many OSVs are used to lay pipes connecting the wells and platforms. As noted by Abu-Marrul et al. (2020), the connections can only start after the drilling and completion of the well. Also, all

rig operations demand the use of other resources, such as equipment and crews. To assure that the equipment and crews will be available on time to execute the rig tasks, it is important to consider the inventory level (equipment) or schedule (crew) within the RSP. This supply chain perspective usually aims to minimize the logistic costs of the operations.

This classification and the taxonomy in Figure 2.3 allows to classify the different dimensions of the RSP under the perspective of available case studies and solution approaches. Next, we present the methodology used to assemble this literature review of the RSP.

2.3
Systematic literature review's methodology

This systematic literature review (SLR) was executed based on a step-by-step method adapted from Thome et al. (2016), which can be divided into: (i) research delineation; (ii) literature search; (iii) data collection; (iv) data analysis; (v) interpretation. This complete SLR process is illustrated in Figure 2.4, in which green labels represent the outputs of each step.

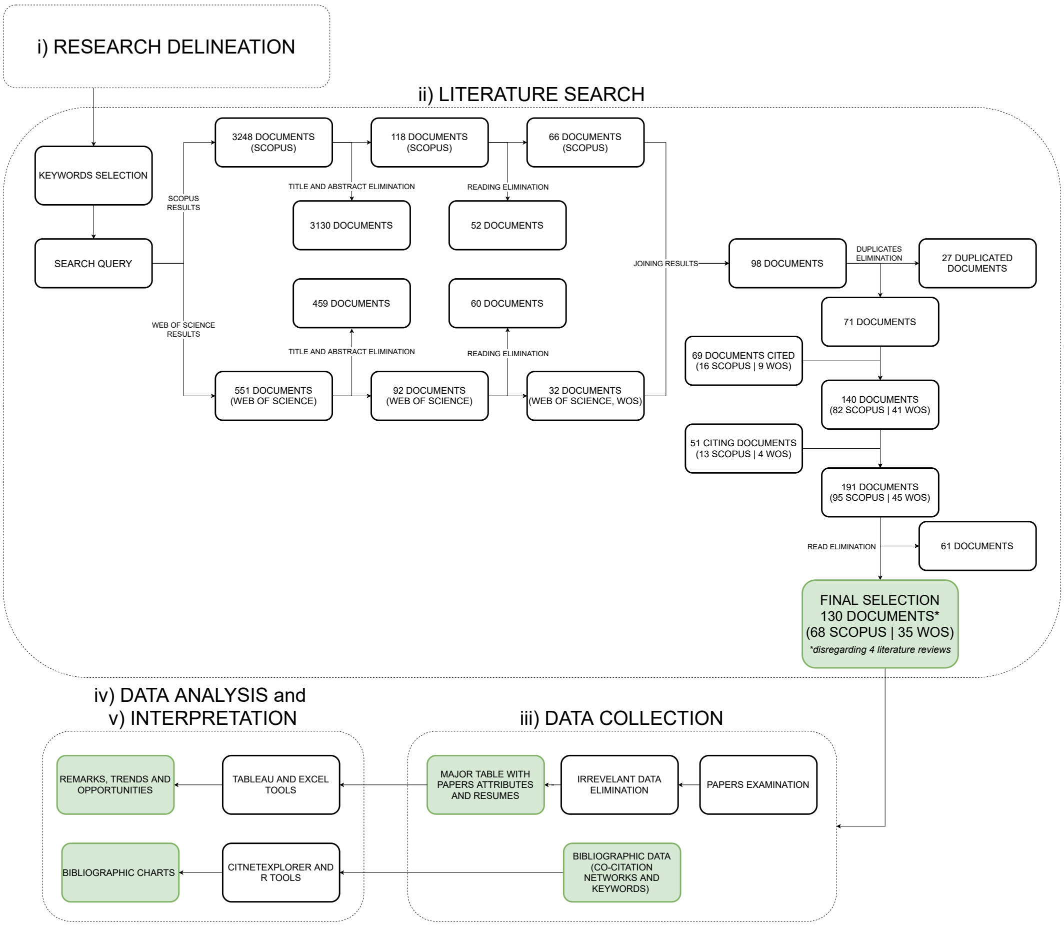

i) RESEARCH DELINEATION
ii) LITERATURE SEARCH
KEYWORDS SELECTION
SEARCH QUERY
SCOPUS RESULTS
WEB OF SCIENCE RESULTS
3248 DOCUMENTS (SCOPUS)
118 DOCUMENTS (SCOPUS)
66 DOCUMENTS (SCOPUS)
TITLE AND ABSTRACT ELIMINATION
READING ELIMINATION
3130 DOCUMENTS
52 DOCUMENTS
459 DOCUMENTS
60 DOCUMENTS
TITLE AND ABSTRACT ELIMINATION
READING ELIMINATION
551 DOCUMENTS (WEB OF SCIENCE)
92 DOCUMENTS (WEB OF SCIENCE)
32 DOCUMENTS (WEB OF SCIENCE, WOS)
JOINING RESULTS
98 DOCUMENTS
DUPLICATES ELIMINATION
27 DUPLICATED DOCUMENTS
71 DOCUMENTS
69 DOCUMENTS CITED (16 SCOPUS | 9 WOS)
140 DOCUMENTS (82 SCOPUS | 41 WOS)
51 CITING DOCUMENTS (13 SCOPUS | 4 WOS)
191 DOCUMENTS (95 SCOPUS | 45 WOS)
READ ELIMINATION
61 DOCUMENTS
FINAL SELECTION 130 DOCUMENTS* (68 SCOPUS | 35 WOS)
*disregarding 4 literature reviews
iv) DATA ANALYSIS and v) INTERPRETATION
iii) DATA COLLECTION
REMARKS, TRENDS AND OPPORTUNITIES
TABLEAU AND EXCEL TOOLS
MAJOR TABLE WITH PAPERS ATTRIBUTES AND RESUMES
IRREVELANT DATA ELIMINATION
PAPERS EXAMINATION
BIBLIOGRAPHIC CHARTS
CITNETEXPLORER AND R TOOLS
BIBLIOGRAPHIC DATA (CO-CITATION NETWORKS AND KEYWORDS)

Figure 2.4: Framework for the systematic selection process.

The first step was the research delineation, the problem description in Section 2.2, and the methodology definition (current section). In the second step, we have chosen to use Scopus and Web of Science (WOS) as the primary search databases. Aiming to define the most common keywords, a collection of benchmark RSP studies was manually selected and their titles, keywords, and abstracts were analyzed. As mentioned earlier, one of the main contributions of this study is to propose a taxonomy and classification for the RSP that can be used in its different dimensions and decision levels. This literature gap can be observed through our SLR and especially after analyzing the selected RSP papers. With the help of a word cloud plot (shown in the Appendix), using titles, keywords, and abstracts, we selected a search query that was simultaneously broad to avoid limiting results and limited to avoid too many unwanted results. This search query combines the following keywords: *scheduling*, *rescheduling*, *routing*, *mobilization*, *move*, *programming*, *simulation*, *optimization*, *model*, *algorithm*, *heuristic*, *procedure*, *technique*, *system*, *well*, *oil*, *petroleum*, *onshore*, *offshore*, *workover*, *drilling*, *completion*, *downhole*, *reservoir*, *evaluation*, *rig*, and *vessel*. No specific time frame was considered, and all articles found have been included as long as an online version of the paper could be found. The complete search query is presented in the Appendix.

This search query was executed on the Scopus and WOS databases on 27^{th} November 2020, resulting in 3248 papers from Scopus and 551 documents from WOS. By reviewing titles, keywords, and abstracts, 3130 and 459 documents were eliminated from Scopus and WOS, respectively. The elimination criteria used was to select only papers that consider routing or scheduling of oil rigs and written in English or Portuguese. This large number of unwanted papers is a direct consequence of the lack of a robust classification and taxonomy system for the RSP, which this SLR tries to address. The 118 (Scopus) and 92 (WOS) resulting articles were read and 52 (Scopus) and 60 (WOS) documents were eliminated from the review. After joining the 98 documents found in Scopus and WOS, 27 duplicated papers were eliminated, culminating in the selection of 71 papers. The articles cited by them were searched using Google Scholar and Scopus databases, and 69 new documents were appended to the list, of which just 16 and 9 were actually in Scopus and WOS, respectively. Then, the papers citing the selected papers in Google Scholar and Scopus databases were read and 50 new studies (13 from Scopus and 4 from WOS) were found. Last, 61 documents were eliminated after a more profound analysis to double-check the selection, resulting in 130 papers, of which 68 and 35 were available in Scopus and WOS Databases, respectively. During this process, some literature reviews were found: Bissoli et al. (2016), Tavallali and Karimi

(2014), Tavallali et al. (2016), and Khor et al. (2017). We chose to exclude these papers from the final selection as they do not specifically propose a new method, model, or case study. Table 2.1 summarizes these literature reviews, their scope, the number of studies revised (RSP and total), and the time frame (the years considered). The literature reviews with an asterisk did not count how many papers were being revised, but the total number was bigger than 100. As none of them did a systematic literature review and many classes of the RSP were not considered, the number of RSPs revised by them is considerably smaller than the number of publications found using our methodology. Their results will be compared with our findings later on.

Literature Review	Main goal	Papers revised (RSP)	Papers revised (total)	Years
Tavallali and Karimi (2014)	Field planning, development decisions, and process system perspectives	9	48	1990-2014
Tavallali et al. (2016)	Field planning, development decisions, and process system perspectives	8	100*	1990-2016
Bissoli et al. (2016)	Workover planning, its drivers, and vehicle routing perspectives	33	64	1977-2015
Khor et al. (2017)	Optimization methods for field development problems	13	100*	1972-2017

Table 2.1: Summary of the others literature reviews found.

For the third step (data collection), the papers were thoroughly examined. With the support of the classification presented in the previous section, the data was gathered in a table with the essential information. Other data, such as co-citation networks and keyword occurrence, were manually generated and assembled in bibliographic citation files.

The fourth and fifth steps, data analysis and interpretation, are presented in the following sections. Qualitative and quantitative content analyses were performed with the assist of the following tools: CitNetExplorer (software for bibliography analysis used for co-citation and keyword analysis of the papers), R (a programming language and free software environment used for statistical computing and graphics), Excel and Tableau (an interactive data visualization software).

2.4 Systematic literature review results

The use of Operations Research and quantitative methods in oil rig scheduling dates back to the 1960s when Aronofsky and Williams (1962) and Aronofsky (1962) proposed two linear models for oil production planning. At that time, these models required substantial computational resources, preventing any practical application to the problems (Pittman 1985). As a result, most developments concerning rig scheduling employed approximation techniques (Barnes et al. 1977) and decision-making rules (Cochrane 1989). The body of research around the topic only started to fully develop in the 1990s with the improvement in computational capacity and optimization methods, as shown in Figure 2.5.

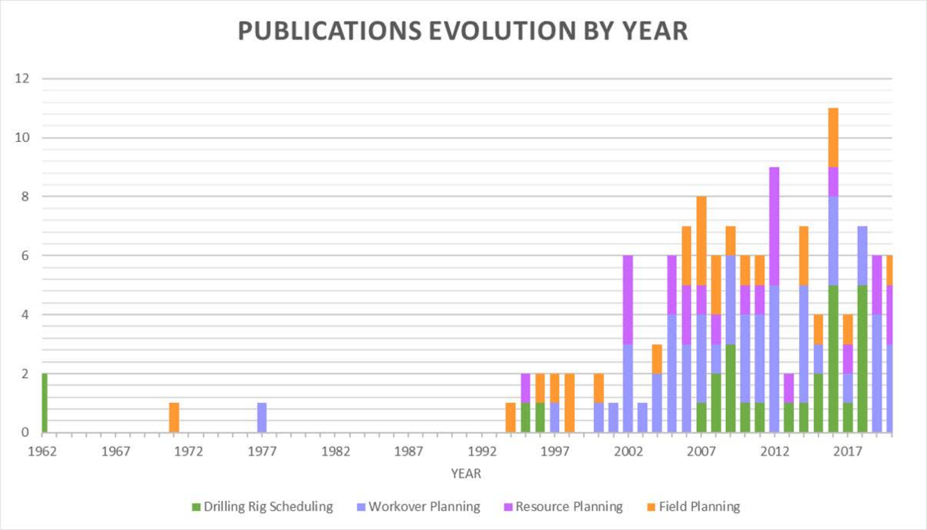


Figure 2.5: Evolution of the Rig Scheduling studies over the years.

It is possible to observe in Figure 2.5 that there was a significant growth of workover rig planning (scheduling/routing and scheduling) research in the early 2000s. This particular growth is mainly due to the interest of Brazilian researchers and industry stakeholders in the problem, for which we can identify three main reasons. First, the Brazilian Petroleum Investment Law mandates that part of the royalties of oil and gas exploration and production is invested in research and development applications (Pessôa Filho et al. 2006). Second, during the 2000s and 2010s, major oil fields were discovered in Brazil, leading to a significant growth in its oil production and, consequently, in investments in research (Iachan 2009). Finally, the Brazilian exploration and production of oil and gas are concentrated within the state-owned company Petrobras, which used to hold a monopoly in the Brazilian downstream and midstream supply chain (Iachan 2009). As a result, Brazil is one of the main rig markets, with

57 active rigs in 2011, of which 10 are owned by Petrobras (Kaiser and Snyder 2013). This large fleet of active rigs poses significant challenges concerning its planning. Consequently, with the massive amount of investments in research and operations optimization, many rig scheduling studies were published by authors affiliated with Brazilian institutions, as shown in Table 2.2. Also, it is possible to observe that other petroleum-producing countries have a large number of publications. The values in Table 2.2 were calculated by summing the number of authorships from each country in each study.

Author's Country	Number of Publications
Brazil	144
USA	78
Norway	24
Argentina	20
United Kingdom	14
Canada	13
Saudi Arabia	11
United Arab Emirates	10
Iran	8
Austria	8
India	6
France	6
Indonesia	5
Others	17

Table 2.2: Author affiliation's country distribution for the rig scheduling problem.

Regarding the type of publication, the majority of studies was from conference proceedings. However, there is a significant volume of papers published in journals (49), as shown in Table 2.3.

Type	Publication
Conference	**56**
Journal	**49**
Master Thesis	10
Doctoral Dissertation	7
Book	4
Graduate	2
Tech Report	1
Book Chapter	1

Table 2.3: The number of rig scheduling publications per type of study

Using the CitNetExplorer software, we developed the citation network of the rig scheduling publications found in the literature review shown in Figure 2.6. Studies without any citation link were omitted from the chart for the purpose of clearance and clarity. The four major groups are marked in the chart: Drilling Rig Scheduling (green labels); Workover Planning (blue labels); Resource Planning (purple labels); and Field Planning (orange labels).

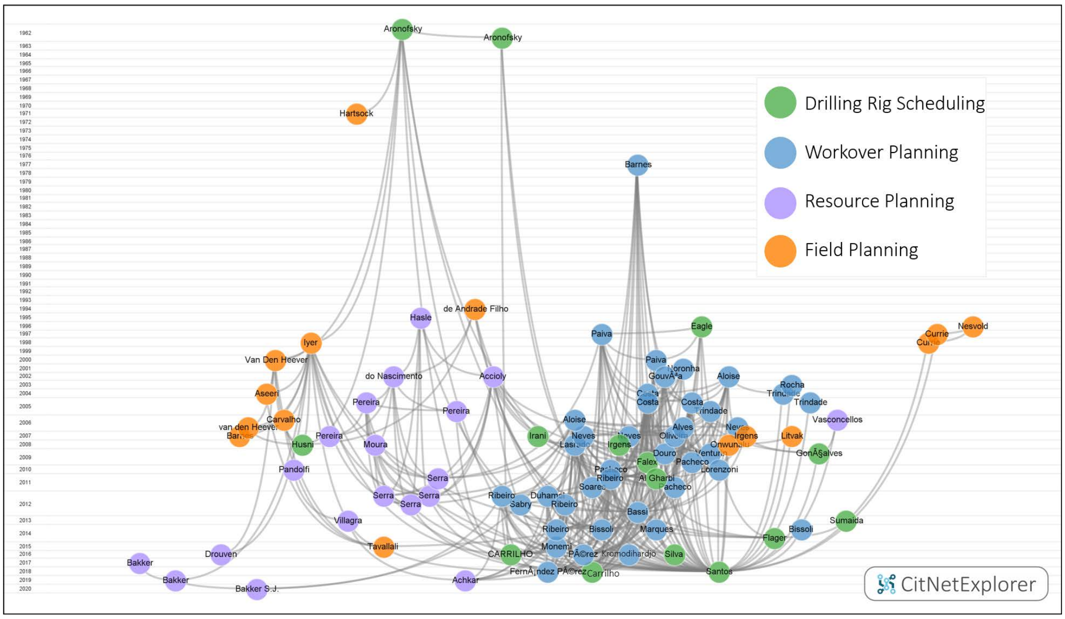


Figure 2.6: The citation network of the RSP studies.

The citation links in Figure 2.6 are also helpful in understanding the connection between these RSP groups. For instance, workover planning problems are highly centralized and connected, suggesting a greater level of discussion inside the literature. As expected, the drilling rig scheduling problems are closer to the workover planning problems, as these two groups share many modeling similarities. The integrated problems, resource planning and field planning, are more scattered, as these problems are more diversified. Next, we study each of these groups, analyzing how the research outputs evolved and understanding how the RSP has been addressed in the literature.

2.4.1 Drilling rig scheduling problem

Aronofsky and Williams (1962) and Aronofsky (1962) were the first known studies addressing the RSP using linear programming. They proposed a model for scheduling the oil production curve under a fixed drilling rig schedule and another one to schedule rigs and drilling tasks under a predefined production. Hartsock and Greaney (1971) developed a mixed-integer non-linear programming (MINLP) inventory model for optimizing the drilling schedule of an oilfield considering the rig's operation and transportation costs. Benefiting from the improved computational resources made available since then, Haugland et al. (1991) proposed a linear programming model for allocating fixed and movable offshore rigs to routes maximizing the net presented value (NPV). Gutleber et al. (1995) presented a fuzzy ranking method used in the drilling schedule. One year later, Eagle (1996) used a simulated annealing (SA) algorithm to schedule drilling rigs and to maximize NPV in a multi-period horizon.

A decade later, Irani (2007) described a system implemented in a Mexican oil company that allows real-time managing and visualization of the rig schedule and the drilling tasks. Irgens and Lavenue (2007) and Irgens et al. (2008) used a stochastic local search to maximize oil production and provide real-time visualization to schedule a heterogeneous fleet of drilling rigs. Meanwhile, a drilling rig fleet sizing model was proposed by Husni (2008) for scheduling oil projects using linear programming and a genetic algorithm (GA) with a greedy heuristic (GH). Glinz and Berumen (2009) presented a mathematical programming model to schedule drilling resources. Falex (2009) proposed a GA to the drilling rig scheduling problem with heterogeneous fleet that minimizes the rig's hires and oil production loss. Addressing a real case study of an Emirati oil company, Sumaida et al. (2013) presented a systematic methodology for manually routing onshore rigs. Amrideswaran et al. (2015) presented a framework for risk assessment in workover and plug and abandonment (P&A) operations in which offshore rigs are scheduled according to a GH, based on a priority ranking matrix. Another systematic approach was later proposed by Arnaout et al. (2017), focusing on the operational perspective of onshore drilling RSP. Amer et al. (2016) described a system for scheduling and managing a fleet of drilling and workover rigs with feasibility validation over a master schedule.

Some authors combined simulation and optimization techniques for RSP. Flager (2014) proposed a multi-objective GA with Monte Carlo simulation to schedule a heterogeneous fleet of onshore drilling rigs maximizing oil production and minimizing its cost. Zahran and Al-Fardan (2014) proposed an automated system for scheduling and routing rigs that used both simulation and optimization algorithms.

Meanwhile, other studies focused on mathematical formulations and solving techniques. Gonçalves (2009) used a GA for the drilling rig routing and scheduling problem, taking advantage of the ease of GA modeling to introduce complex constraints, such as environmental and regulatory laws and rigs displacement costs. Al Gharbi (2011) addressed the routing and scheduling of a homogeneous fleet of onshore drilling rigs and proposed a heuristic based on the Dijkstra algorithm. Haugland and Tjøstheim (2015) presented alternative linear programming formulations for scheduling and routing a heterogeneous fleet of offshore rigs. First, they introduced a model for drilling and location decisions. As an alternative method, the authors proposed a dynamic network flow model for rigs moving and drilling decisions. Chowdhury (2016) optimized the routes and schedules of onshore drilling rigs using the program evaluation and review technique and critical path method techniques. Silva et al. (2016) proposed a mixed-integer linear programming (MILP) model for the rig routing and scheduling problem of a heterogeneous offshore fleet, minimizing production loss and rig utilization costs. They tested their approach considering a small instance with a variety of tasks and realistic assumptions, resulting in non-linear constraints.

Aiming to address technical and economic constraints, Tavallali and Zare (2018) proposed a MILP model for routing and scheduling the drilling activities on a fleet of owned/hired rigs, minimizing drilling costs, rigs movements costs, and hiring costs and considering eligibility, rig's contract length, and others constraints. Using advanced techniques of the VRP literature, Kulachenko and Kononova (2020) presented a Variable Neighborhood Search (VNS) based matheuristic in which a MILP solver is used to optimize the well-drilling work distribution and the VNS solves the routes.

A decision support system was presented by Carrilho and Villas Boas (2016) for the RSP. It uses a MILP model to maximize the tasks allocation to the rigs already hired and minimize the fleet of rigs to be hired. To reduce computational requirements, the authors considered the jobs in blocks and the time horizon in weeks. Using this block structure, Santos (2018) presented two models for the offshore RSP, one for minimizing the fleet of rigs and another for minimizing its costs. Some local searches, constructive heuristics, and matheuristics were also developed and tested in real-life instances. However, most methods required considerable computational effort as a result of the time-indexed formulation considering long-term planning horizons. An alternative formulation for the time-indexed parallel machine scheduling was introduced by Carrilho et al. (2018) based on bucket-indexing. This time formulation divides the planning horizon into periods of equal length (buckets with a size that can assume values between 1 and the shortest processing time of jobs) and achieved promising results that enable realistic rig scheduling models for large instances with long planning horizons.

A trend that has become popular since 2015 is the use of machine learning and data science techniques to support optimization algorithms, also known as data-driven optimization. Ma et al. (2018) proposed a method that uses a data mining system to extract key information from daily drilling reports and historical data, convert and aggregate it in a database, identify drilling opportunities, and use it to optimize the short-term rig schedule. Castiñeira et al. (2018) used machine learning and natural language processing (NLP) for an automated analysis of drilling data. The historical data was then used to optimize the rig schedule through heuristics, maximizing NPV and oil production. Both studies have used advanced machine learning techniques, but the optimization methods and formulation are not defined in sufficient detail, preventing us from comparing them with others formulations. A summary of the total of the DRSP studies discussed in this section is illustrated in Figure B.1 in the Appendix.

2.4.2 Workover rig scheduling problem

Barnes et al. (1977) investigated the workover rigs scheduling problem (WRSP) and proposed two approximate techniques to minimize the oil production loss, testing it on a small and short-term instance. Decades later, several other papers addressing the WRSP were published. Noronha and Aloise (2001) presented a GH for planning operations in onshore rigs that minimizes not only the oil production loss but also the environmental risks. Aloise et al. (2002) tested variations of ant colony with path-relinking (AC-PR) against a GA and a greedy randomized adaptive search procedure (GRASP). Gouvêa et al. (2002) proposed two evolutionary heuristics for scheduling workover operations in a homogeneous fleet of onshore rigs: a transgenetic algorithm (TA) and a memetic algorithm (MA). Maia et al. (2002) compared a simplified tabu search (TS)-based heuristic with the heuristics from Gouvêa et al. (2002) and

Aloise et al. (2002). The AC-PR remained achieving the best results. According to Bassi (2010), this study was related with a geo-referenced computational system for workover onshore rigs management that was also discussed in Maia et al. (2002), Gouvêa et al. (2002), Aloise et al. (2002), and Aloise et al. (2006).

After modeling the WRSP as a binary integer linear model, Costa and Ferreira Filho (2004) created a maximum priority three-criteria heuristic (MPTH), whose simplicity allows it to be implemented in simulations and sensitivity analysis. Later, Costa and Ferreira Filho (2005) tested a dynamical assemble heuristic (DAH) that overperformed the MPTH even in large examples. These two methods can also be found in Costa (2005), which also presents a GRASP and 300 real-life instances for the problem.

Several other authors used these instances later and tested them with different solution algorithms: a GA (Alves and Ferreira Filho 2006); a scatter search (SS) by Oliveira et al. (2007) and Lorenzoni and Polycarpo (2010); a bubble swap (BS) by Pacheco et al. (2009a); a GA-2opt (Douro and Lorenzoni 2009); a GRASP-PR from Pacheco et al. (2009b) and Pacheco et al. (2010); a simulated annealing (SA)-based heuristic from Ribeiro et al. (2011); a memetic algorithm (MA) from Pacheco (2011). The greedy randomized adaptive search procedure with path-relinking (GRASP-PR) from Pacheco et al. (2009b) and Pacheco et al. (2010) was based on Costa (2005)'s GRASP. To achieve better solutions for the WRSP, Lorenzoni and Polycarpo (2010) enhanced the SS (Oliveira et al. 2007) using MPTH as a solution generator and found 11 new best solutions. Ribeiro et al. (2011) tried to solve the WRSP's harder instances, proposing a simple, yet robust, simulated annealing (SA)-based heuristic to generate an initial solution and apply the SA iterative, ultimately outperforming DAH, GRASP, GRASP-PR, SS, BS, GA-2opt and MA.

A few variations of the WRSP can be found in the literature. Specifically, Lasrado (2008) created an application based on a manual methodology that adapted the reservoir simulation technique from de Andrade Filho (1994) to generate schedules and minimize the number of rigs and traveling distances, reducing transportation and contract costs. Marques et al. (2014) presented a decision support system that uses a MILP model to size and schedule homogeneous offshore rigs, minimizing the fleet size and maximizing its utilization.

Meanwhile, Monemi et al. (2015) addressed the heterogeneous WRSP, proposing a new MILP model based on arc-time-indexed formulations and two solution techniques: branch-price-and-cut (BPC) and hyper-heuristic (HH), which found near-optimal solutions with just a few seconds. Danach (2016) also tackled this problem with a (1,0)-linear programming model and a HH using algorithms for construction, local search, perturbation, and reconstruction.

The HH was tested in a real case and faced difficulties in solving large instances. Hence, future works on the mathematical formulation were suggested by the authors to improve its efficiency.

Aiming to achieve better solutions, Pérez et al. (2016) proposed a decomposed reformulation of the (1,0)-linear model from Costa and Ferreira Filho (2004) for the WRSP of homogeneous onshore rigs that had fewer variables and constraints and was tested in the instances of Costa (2005), finding new exact solutions for large instances and outperforming the heuristic methods. Based on their model, Fernández Pérez et al. (2018), Pérez et al. (2019) presented deterministic and stochastic models for the WRSP to fleet size and to minimize the oil production loss and rig fleet costs. The authors adapted the instances from Costa (2005), Paiva (1997), Soares et al. (2011), Ribeiro et al. (2012a), and Bissoli (2014), testing it with several scenario generation methods, including Monte Carlo simulation, scenario reduction, and Quasi-Monte Carlo, and achieving robust solutions even for large instances. A summary of the number of WRSP studies discussed in this section is illustrated in Figure B.2 in the Appendix.

2.4.3 Workover rig routing and scheduling problem

Building upon advances in vehicle routing problem formulation and solution techniques, Paiva et al. (2000) proposed a SA, based on Paiva (1997), for the workover rig routing and scheduling problem (WRRSP), minimizing rig expenses and oil production losses. Since then, many other authors have tackled the homogeneous WRRSP with several heuristics. Rocha et al. (2003) presented 3 variations of variable neighborhood search to the WRRSP, obtaining the best results with a cooperative parallel VNS (CPVNS) with PR. Trindade and Ochi (2004) proposed 6 variations of GRASP-PR, later enhanced by Trindade (2005) and Trindade and Ochi (2005) to a hybrid GRASP-PR. To improve the GRASP-PR efficiency, Neves (2007) and Neves and Ochi (2006, 2007) presented a GRASP with adaptive memory (GRASP-AM) and tested it against other heuristics such as TS and iterated local search (ILS). Ribeiro et al. (2012b) compared this ILS with a clustering search (CS) and an adaptive large neighborhood search (ALNS). In this study, the ALNS outperformed the other methods. Finally, Shaji et al. (2019) proposed a new aggregated rank removal heuristic (ARRH) to the ALNS (Ribeiro et al. 2012b) and compared it with other heuristics: VNS (Aloise et al. 2006); GA; GA with VNS (GA+VNS) and ALNS (Ribeiro et al. 2012b). These heuristics were tested in some theoretical instances in which the ARRH based ALNS outperformed the other methods.

Some authors focused on new formulations for the homogeneous WRRSP. Sabry et al. (2012) proposed a new MILP formulation minimizing oil production and rig operation costs for a company that owned a dedicated rigs fleet and could hire additional rigs. The authors tested their model considering a short-term theoretical instance using a MA and GRASP. Duhamel et al. (2012) proposed three models and hybrid methods for onshore workover rigs aiming to minimize the total production loss: a MILP model based on Aloise et al. (2006); an open vehicle routing problem (OVRP) strategy with lifted constraints and better bounds; and a set-covering formulation, obtained through a Dantzig-Wolfe decomposition of the OVRP and enhanced using column generations with GRASP and VND. Last, Kromodihardjo and Kromodihardjo (2016) used a discrete simulation software to propose exhaustive search and combinatorial algorithms for the WRRSP, obtaining near-optimal or optimal solutions in small instances based on real data.

Another variation of the WRRSP is to consider a heterogeneous fleet of rigs. Aloise et al. (2006) addressed this problem using a VNS that mixes several swap and insert moves (e.g., changing the wells allocated to a rig or allocating different rigs to a well). The problem was tested in real-life instances and later implemented in a Brazilian oil company, generating potential savings of US$2.5M per year. Soares et al. (2011) analyzed the characteristics of the WRRSP and proposed constructive heuristics and a new objective function minimizing the rig's fleet cost. Meanwhile, Ribeiro et al. (2012a) tried to find exact solutions with a branch-price-and-cut (BPC) approach (based on TS, column generation, ng-path-relaxation, and subset-row inequalities) that enabled it to solve real-life examples with up to 200 wells and 10 rigs. Later, Ribeiro et al. (2014) presented a hybrid-GA (HGA) to heterogeneous WRRSP and compared it with three other methods: VNS (Aloise et al. 2006), branch-price-and-cut (Ribeiro et al. 2012a), and ALNS (Ribeiro et al. 2012b). The BPC, ALNS, and HGA were consistently superior to the VNS, having the first, faster solutions than the alternative methods, but with lower qualities than the ALNS and HGA, which in turn was the method that found all the best solutions. Last, Bissoli et al. (2014) and Bissoli (2014) also addressed the WRRSP using a bi-objective ALNS that minimizes the production loss and the onshore rig fleet size, which, according to the authors, also minimizes the total costs. However, this assumption is a simplification as, in reality, a minimal fleet does not mean that chartering costs are optimal.

A different approach to the WRRSP was proposed by Vasconcelos et al. (2017). The authors developed a genetic algorithm that uses operational historical data integrated into an optimization workflow to minimize the total non-productive time of the offshore wells served by a heterogeneous fleet of vessels with different load capacities and limited abilities. The proposed algorithm was tested on real data of a petroleum company and improved, in terms of navigation and operation time, between 20-40% of the original plan. Later, Tozzo et al. (2020) proposed a hybrid GA for the WRRSP with heterogeneous fleet minimizing oil production loss and rig fleet costs from a multi-objective perspective.

A trend in the optimization models is to consider uncertainty in the decision-making. This trend is important to RSP, which emerged from the need to consider the uncertainties related to geological concepts (structure, reservoir seal, and hydrocarbon charge), economic evaluations (costs, probability of finding, and producing economically viable reservoirs, technology and oil price), development and production (infrastructure, production schedule, quality of oil, operational costs, and reservoir characteristic), traveling time between wells and well service time (especially in the offshore fields that are subjected variable conditions such as weather and sea state) (Suslick et al. 2009).

Following this tendency, Bassi (2010) and Bassi et al. (2012) proposed a simulation–optimization approach to minimize production loss for heterogeneous offshore rigs in two phases: simulation of well service times and oil potentials; and optimization using GRASP. These phases were repeated for a significant number of times, instances, and fleet sizes, enabling to make scheduling decisions under uncertainty and unveiling the trade-off between fleet size and oil loss as a larger number of rigs results in better performance measures and higher operating costs. Bassi et al. (2012) also provide a literature review concerning workover rigs with fruitful discussions.

Most WRRSP studies consider that the decision-maker knows beforehand which wells will require intervention. However, in reality, often, one cannot know with certainty which and when a well will be due maintenance. To tackle this problem, Silva and Silva (2018) proposed a dynamic approach to the WRRSP, minimizing the total oil production loss of wells that are revealed along the planning horizon, called as D-WRRSP (dynamic workover rig routing and scheduling problem). The model was based on the formulation of Ribeiro et al. (2012a) and was tested considering new small and short-term instances adapted from Costa (2005). We provide a summary of the total of the WRRSP studies discussed in this section in Figure B.3 in the Appendix.

2.4.4 Integrated problems

The interdependence of operations in the oil and gas sector requires that oil and gas companies plan and optimize their processes on an enterprise-wide level (Oliveira et al. 2013). As a result of the rig scheduling being a multifaceted decision process, many studies approach this problem by integrating this rig scheduling decision with others. We divide these studies into two classes of integration: field planning and resource planning.

2.4.4.1 Field planning

Before scheduling the rigs to drill new wells, field design and planning decisions are needed, such as well drilling schedule, well placement, facility design, and flow scheduling. Ideally, these problems should be solved together from a field planning perspective.

Using the Eclipse reservoir simulator and a polytope search optimization algorithm, de Andrade Filho (1994) addressed the drilling RSP by deciding the main development dates combined with the drilling rigs allocation and schedules. Since then, other studies also integrated reservoir simulator models with optimization algorithms, such as: linear programming and simulation approaches (Nesvold et al. 1996); a procedure for BP's (British Petroleum) top-down reservoir modeling tool using an enhanced GA by Litvak et al. (2007), Litvak and Angert (2009) and Litvak et al. (2011); a GA (Litvak et al. 2007) with statistical proxies procedures using clustering-based techniques from Onwunalu et al. (2008); design space exploration (Cong et al. 2008); automated decision-making system with a first-in-first-out algorithm (Davidson et al. 2009); MINLP model and GA (Tavallali et al. 2015); approximation algorithms with optimization and local searches (Tavallali et al. 2016).

Several authors proposed formulations for field design and planning considering rigs. Iyer et al. (1998) proposed a multi-period MILP model with branch and bound to maximize the NPV. Currie et al. (1997a,b) presented a simplified MIP model for the redevelopment and reservoir management of wells, deciding the projects, wells, and drilling rigs to be used annually. Van Den Heever and Grossmann (2000, 2006) presented a MINLP model maximizing the NPV, which was solved using a dynamic programming (DP) approach with an iterative aggregation/disaggregation algorithm. Carvalho and Pinto (2006) proposed a model and a decomposition method to determine the drilling rigs (platforms) locations, the well to be drilled by each one, and the drilling schedule. Later, Barnes and Kokossis (2007) proposed mathematical models for an integrated field development when deciding the location, drilling schedule, and production rate of the wells. Similar models were also proposed by Wang et al. (2019) and Ondeck et al. (2019). Still, none of these previous studies considered the rig scheduling as a decision but only as something that affects their decisions. Ondeck et al. (2019) performed a sensitivity analysis demonstrating the impact of the drilling rig fleet and crew mobilization costs in their model.

A system for field planning was presented by Martin et al. (2010), mainly for well design. The system designs the well-pad layout, determines the production facilities, allocates wells to rigs considering the fleet availability, rig locations, and other related attributes, and returns the rig schedule as one of its outputs. Omosebi et al. (2014) proposed a methodology based on project management to correctly plan drilling projects and rigs schedules. However, the authors did not mention the employment of optimization methods. Lange and Lin (2014) and Dewan et al. (2016) presented solutions that model the well scheduling process as a multi-agent system to allow optimal decisions, including the rig schedule, for all parts involved in the scheduling process. As mentioned by Neiro and Pinto (2004), this modeling strategy allows to integrate the business entities involved in the supply chain management. Kelly et al. (2017) introduced a MILP model for well startup considering each well as a batch process subject to resource availability constraints (processing plants, drilling rigs, and crew).

Also addressing the planning of an offshore oilfield infrastructure, Aseeri et al. (2004) proposed a sample average approximation (SAA) algorithm to maximize the NPV considering constraints of budget and the availability of one rig to drill the wells, optimizing the flow balance between production platforms, wells platforms, and the reservoirs. As showed by Smith (1956), rig availability constraints for the case of one rig can be considered as a RSP. The authors also considered the travel time between the wells when scheduling the rigs operations. Barnes and Kokossis (2007) introduced a MILP model for the analysis, design, and scheduling of offshore oilfields, considering the drilling schedule, platform locations, and a single rig available. Last, Calderón and Pekney (2020) focused on the field planning decisions related to the enhanced oil recovery to reduce gas flaring in shale oil development. The authors proposed a sophisticated model that optimizes drilling rig schedules, workover decisions, pipeline and facilities infrastructure, location of wells, and injection rates. A summary of the total of field planning studies discussed in this section is illustrated in Figure B.4 in the Appendix.

2.4.4.2 Resources planning

The E&P operations need several resources, which are often planned separately. However, as their decisions might affect each other, resource usage should ideally be integrated with the E&P operations planning. This is the case for the rigs and the other resources involved in the E&P, such as crews and vessels. For instance, Hasle et al. (1996) used constraint reasoning for the well activity scheduling problem, where drilling rigs and wire-line cranes are appointed for drilling, completion, perforation, and logging activities. Horton and Dedigama (2006) presented a resource scheduling system used by an Australian oil company to schedule several operations, such as drilling, completion, and workover, and others that do not involve rigs, such as interconnections.

After completing an offshore well with a rig and before starting its production, offshore support vessels (OSV) are used to interconnect wells, manifolds, and platforms. Focusing on integrating the rig scheduling with the OSVs' decisions, Accioly et al. (2002) used a constraint programming (CP) model to maximize oil production considering drilling, completion, workover, and pipelines connecting activities, priorities, precedence, wells, and ship characteristics. To enable the use of optimization solvers, the authors used different search heuristics to explore the solution space. Since then, several authors tested formulations and solution methods: CP model solved with hybrid TS algorithm (Nascimento 2002); GRASP (Pereira 2005, Pereira et al. 2005a,b, Moura et al. 2008); GA (Vasconcellos and Ferreira Filho 2006); CP model (Serra et al. 2011, Serra 2012, Serra et al. 2012b,a); continuous-time MILP formulation with upper bound relaxations (Serra et al. 2012c). Pereira (2005) and Pereira et al. (2005a,b) formulated a CP model for scheduling drilling, completion, and interconnection operations in offshore wells with rigs, OSVs, and production units. The authors tested a GRASP to solve it in real instances of the Brazilian oil company Petrobras, which resulted in considerable savings for the company and, according to the authors, was implemented in a system called *ORCA*. Moura et al. (2008) adapted the formulation to consider resource displacement, proposing a GRASP to solve it. Based on the model proposed in Serra et al. (2011), Serra (2012), and Serra et al. (2012b,a) proposed a CP model for offshore resource scheduling of a heterogeneous fleet of rigs and PLSVs, aiming to maximize production.

Other studies have tackled the RSP considering its equipment requirements. Drouven and Grossmann (2016) proposed a MINLP based on generalized disjunctive programming (GDP) for the shale gas development. The presented model maximizes the NPV defining which wells will be drilled, when they will be drilled, which rigs, crews, and equipment will perform the drilling, and the layout of the gathering pipelines. Mazzini et al. (2002) proposed a MILP model that decides rigs equipment and drilling/completion rigs schedules, minimizing the costs associated with resources contracts and rigs tardiness. Finally, Marchesi et al. (2019) proposed a MILP model for the construction of wells considering rigs and equipment aiming to minimize tardiness and earliness.

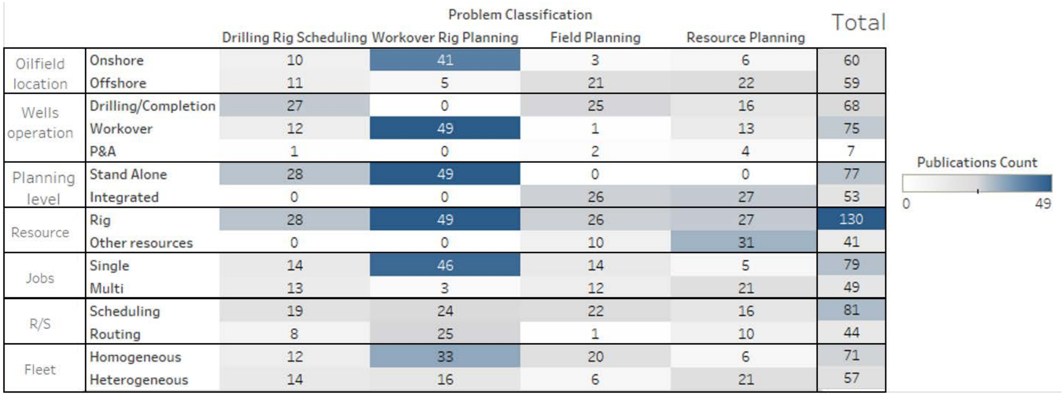
There are several types of workover operations, and, as a result, different types of equipment are needed. With this assumption, McKechnie et al. (2002) presented a management system for workover operations that allows control of the rig schedule and the required equipment. Later, a different problem was introduced by Pandolfi et al. (2010). The authors described a system called PAE (evolutionary algorithm for planning), applied in an extension of the WRRSP considering other resources such as crews and equipment used to service onshore wells. Later, Villagra et al. (2013) adapted the model proposed by Pandolfi et al. (2010) to consider penalty functions and repair algorithms to transform infeasible solutions into feasible ones in a constrained version of the problem. Achkar et al. (2019a,b) proposed a MILP model for this extended WRRSP, considering a heterogeneous rig fleet, precedence constraints, crew shifts, failure risks, and minimizing production loss and costs. Finally, Aurachman et al. (2020) used an influence diagram analysis to model a variation of the WRRSP considering equipment decisions. According to the authors, they are developing a dynamic programming model in which the oil production loss changes with the waiting time.

The decision-making integrated with other resources is especially important for offshore P&A campaigns, which relies on rigs and lighter vessels, such as light well intervention vessels (LWIVs) and light construction vessels (LCVs). Bakker et al. (2017) approached the planning of the offshore P&A campaign, presenting a MILP model based on the VRP that aims to minimize costs of a heterogeneous fleet of semi-submersible rigs, mobile offshore drilling units, and light well intervention vessels. Adapting the model from Bakker et al. (2017), Bakker et al. (2019) developed a commodity flow type formulation for the P&A planning that allowed to tackle larger instances and consider different assumptions such as multiple routes per ship and reduced operability of lighter vessels in the winter. Bakker et al. (2021) adapted these formulations to consider learning curve effects on the vessel's processing times. All these P&A models were part of Bakker (2020), in which the author also presented a study problem using stochastic dual dynamic integer programming for the development of a mature offshore oilfield.

A summary of the total of resource planning studies discussed in this section is illustrated in Figure B.5 in the Appendix, together with further details on all of the literature presented.

2.5 Remarks, trends, and opportunities

The papers found in the literature review were classified according to the taxonomy proposed in Section 2.2. These results are presented in the supplementary files and analyzed with data visualization tools in what follows. Figure 2.7 contains the number of papers found for each taxonomy (rows) and classification group (columns), summarizing how the RSP has been most often addressed in the literature, both from academic and industrial perspectives. Note that "Scheduling" refers to problems that only consider scheduling, while "Routing" refers to problems that consider routing integrated with scheduling.



		Problem Classification				Total
		Drilling Rig Scheduling	Workover Rig Planning	Field Planning	Resource Planning	
Oilfield location	Onshore	10	41	3	6	60
	Offshore	11	5	21	22	59
Wells operation	Drilling/Completion	27	0	25	16	68
	Workover	12	49	1	13	75
	P&A	1	0	2	4	7
Planning level	Stand Alone	28	49	0	0	77
	Integrated	0	0	26	27	53
Resource	Rig	28	49	26	27	130
	Other resources	0	0	10	31	41
Jobs	Single	14	46	14	5	79
	Multi	13	3	12	21	49
R/S	Scheduling	19	24	22	16	81
	Routing	8	25	1	10	44
Fleet	Homogeneous	12	33	20	6	71
	Heterogeneous	14	16	6	21	57

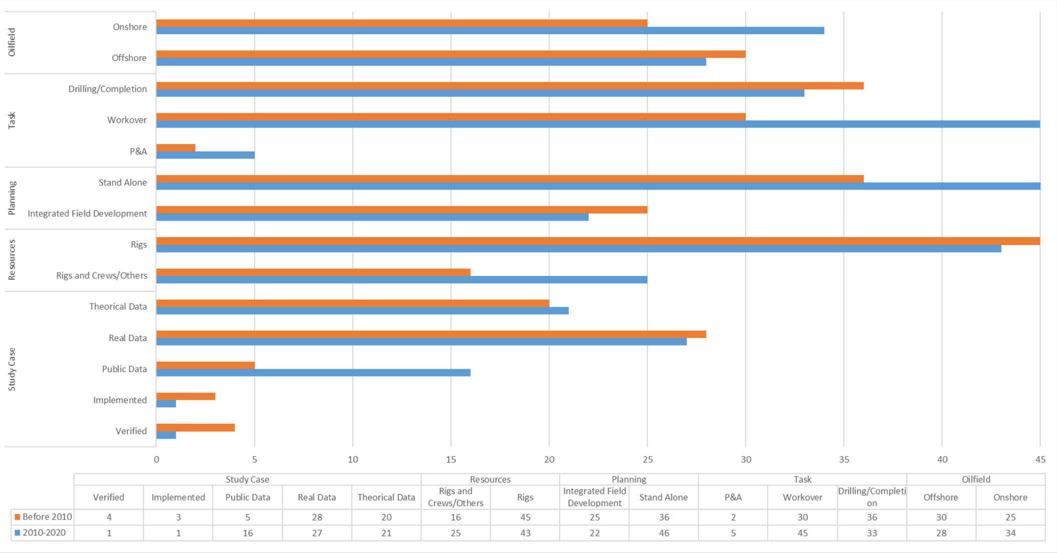
Figure 2.7: Problems count according to their taxonomy and classification group.

Clearly, there is a pattern between the problem classification groups and the taxonomy. DRSPs tackle drilling/completion in stand-alone planning, considering only the rigs. In some cases, they might consider other operations (workover or P&A) as well. Usually, the duration of the drilling operations is long and the distance (specifically the travel time) between wells is short. As a result, these studies are usually modeled as scheduling problems. The workover rig planning problems (WRSPs and WRRSPs) are also set in a stand-alone planning level considering only the rigs. However, differently from the DRSP, workover rig planning focuses on workover operations on wells, usually considering single jobs representing interventions that are often unplanned and can be of much shorter duration than drilling operations. Therefore, the distance between the wells becomes a relevant aspect that is considered by modeling the problem as a routing problem (*i.e.*, WRRSP). In addition, most workover problems consider a homogeneous fleet of rigs.

Field planning problems are those problems that tackle rig operations at an integrated planning level and are usually for drilling operations. Due to its complexity, most of these RSPs are modeled as scheduling problems and consider single jobs and homogeneous fleet. Resource planning problems are related to those RSPs that consider others resources when planning well operations (usually drilling or workover). As there are multiple types of resources, these problems consider multiple jobs per well and a heterogeneous fleet of rigs. Figure 2.7 also reveals some literature gaps: few DRSPs modeled as routing problems; even fewer RSP considering P&A operations and workover rig planning problems for offshore wells; and a lack of field planning and resource planning for onshore oilfields or considering routing approaches.

2.5.1 Problem setting

Aiming to analyze the problem setting (oilfield and tasks type, planning level, resources considered, and case study presented) evolution, the studies were separated according to the publishing date in two groups: *Before 2010* (orange bars) and *2010-2020* (blue bars). Figure 2.8 contains the number of papers found in each group and problem characteristic.



	Study Case					Resources		Planning		Task			Oilfield	
	Verified	Implemented	Public Data	Real Data	Theorical Data	Rigs and Crews/Others	Rigs	Integrated Field Development	Stand Alone	P&A	Workover	Drilling/Completion	Offshore	Onshore
■ Before 2010	4	3	5	28	20	16	45	25	36	2	30	36	30	25
■ 2010-2020	1	1	16	27	21	25	43	22	46	5	45	33	28	34

Figure 2.8: Problem setting evolution.

In the past, the majority of studies were related to offshore wells, but onshore problems have gained more attention in the last decade. Another notable point regarding the type of operations considered is that drilling RSP studies have decreased over the last decade. Meanwhile, the WRSP and

WRRSP have increased considerably. However, the P&A field remains with few studies and it might be thus an important direction for future works.

We can observe that the number of studies using public data has increased significantly in the last decade. The majority of the research uses real or public data, which means that there are studies with a practical perspective fostering the exchange of knowledge between academy and industry. Nonetheless, few studies were verified or implemented by companies, highlighting a gap in the literature. Futures work should therefore focus more closely on meeting the industry demands.

Furthermore, there was a decrease in integrated field development problems and an increase in stand-alone problems as the RSP gained priority in the decision-making. In contrast, we can observe a slight growth in studies considering other resources besides rigs, such as offshore support vessels (OSVs), lighters vessels, crews, equipment, and wire-line cranes, as shown in Figure 2.9.

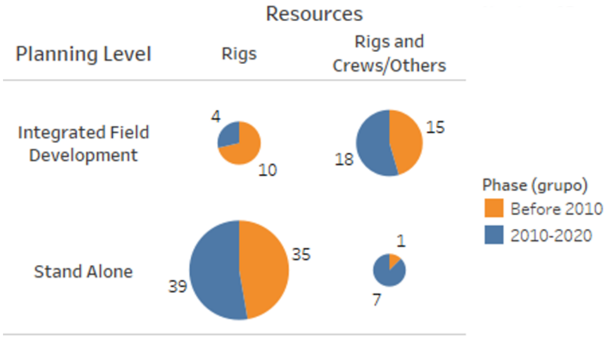


Figure 2.9: Evolution of the study planning level and the resources considered in the study.

With the purpose of understanding more about the relationships underneath the problem characteristics, Figure 2.10 presents a classification of the papers according to their oilfield, task types, and planning levels.

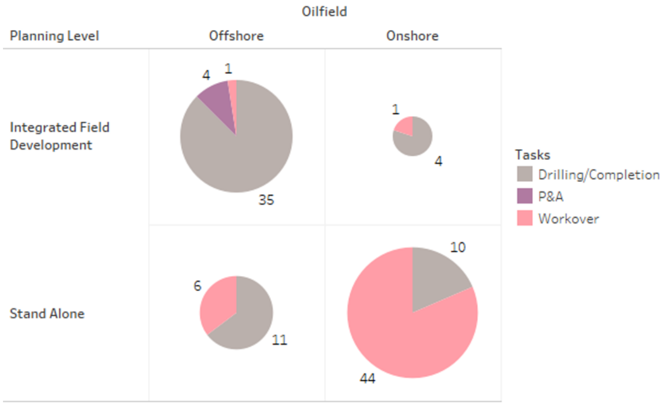


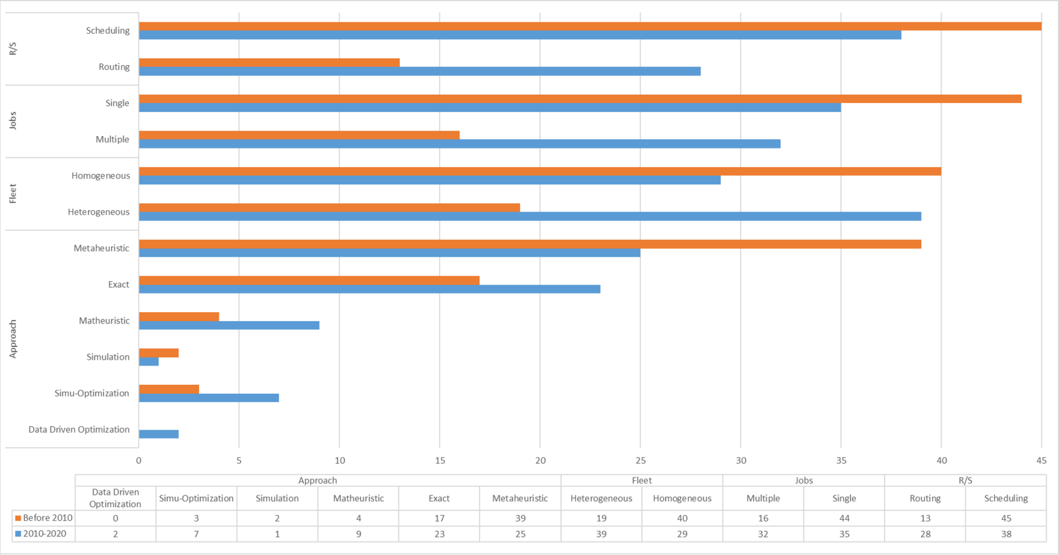
Figure 2.10: Relationship between oilfield, task type, and the study planning level.

First, we observe in Figure 2.10 that the problems considering P&A are more relevant for offshore wells and require an integrated field development perspective. Second, as mentioned by Tavallali et al. (2016), workover planning is a field operation (production phase) decision and is usually separated from field development. Therefore, studies addressing workover planning are usually stand-alone problems. Third, most of the attention of the workover problems has been to onshore wells, so there is an opportunity for WRSP and WRRSP for offshore oilfields. Last, we can observe that most of the drilling RSPs were for offshore wells and integrated field development.

According to Suslick et al. (2009), the offshore environment is immersed with uncertainties, high investments, and high-risk operations, which makes the offshore RSP more complex, requiring the drilling RSP to be treated as an integrated field development decision. Also, these results show new possibilities for studies related to applying an integrated field development for onshore wells and workover operations. Bissoli et al. (2016) suggested that in real situations, the WRSP and WRRSP models should consider all the possible elements that affect the optimization.

2.5.2 Approach

Aiming to observe the evolution of problem approaches (routing/scheduling, jobs, fleet type, and method) over time, the studies were separated according to the publishing date in two groups: *Before 2010* (orange bars) and *2010-2020* (blue bars). Figure 2.11 contains the number of papers found in each group and problem characteristic. The row *R/S* stands for *Routing/Scheduling*. Note that "Scheduling" refers to problems that only consider scheduling, while "Routing" refers to problems that that consider routing integrated with scheduling.



	Approach						Fleet		Jobs		R/S	
	Data Driven Optimization	Simu-Optimization	Simulation	Matheuristic	Exact	Metaheuristic	Heterogeneous	Homogeneous	Multiple	Single	Routing	Scheduling
🟧 Before 2010	0	3	2	4	17	39	19	40	16	44	13	45
🟦 2010-2020	2	7	1	9	23	25	39	29	32	35	28	38

Figure 2.11: Problem approaches evolution.

One can observe in Figure 2.11 that there were some changes in the modeling approach over the last decade. Before 2010, most of the studies focused on scheduling, but since then, the number of studies incorporating routing has grown considerably and the scheduling-based approaches alone have fallen proportionally. A similar pattern is observed in the way that the jobs (operations) are modeled. Studies considering multiple jobs to be optimized have gained more attention. Meanwhile, single jobs have become less frequent. Finally, just as the previous attributes, the RSP considering a heterogeneous fleet has grown and was accountable for the majority of the studies in the last decade. In summary, there seems to be a trend towards turning modeling assumptions more realistic, and thus rig scheduling studies started to consider more complex assumptions, such as considering routing, multiple jobs, heterogeneous fleet, and other more realistic aspects, as mentioned by Santos (2018).

As to the solution methods (the approach row in Figure 2.11), there was a reduction in the use of heuristics and metaheuristics, even though they still represent the most common type of method used. On the other hand, there was an increase in other approaches: exact (mathematical programming), matheuristic (hybrid methods combining heuristic and mathematical programming), simulation, simu-optimization (a hybrid approach combining simulation and optimization), and data-driven optimization (an emerging approach that uses machine learning methods applied to the optimization). This pattern follows Khor et al. (2017) literature review of the optimization methods used in field development problems, where sophisticated methods are being employed more often, in particular matheuristics and models that consider uncertainties in costs, geological aspects, processing and traveling times, tasks occurrence, and rig availability.

We also highlight that a dynamic programming approach for the WRRSP (Silva and Silva 2018) was found during our review. Bissoli et al. (2016) and Bassi et al. (2012) suggested, as future works, dynamic models allowing real-time optimization and rescheduling. Another crucial point is the employment of data-driven optimization (Ma et al. 2018, Castiñeira et al. 2018), which is a new trend in the areas of Operations Research and Management Science and entails using big data and machine learning (ML) techniques to devise improved models and/ or solution methods. Both authors focused on using ML to support the mathematical modeling, but there is the possibility of using ML to improve the performance of the solution method as well. Next, Figure 2.12 presents the number of publications found of each modeling approach (routing or scheduling), the task type studied (drilling/completion, P&A or workover), and the objective function used (time indicator, rig fleet size, oil production, oil production loss, costs, multi-objective or economic indicator).

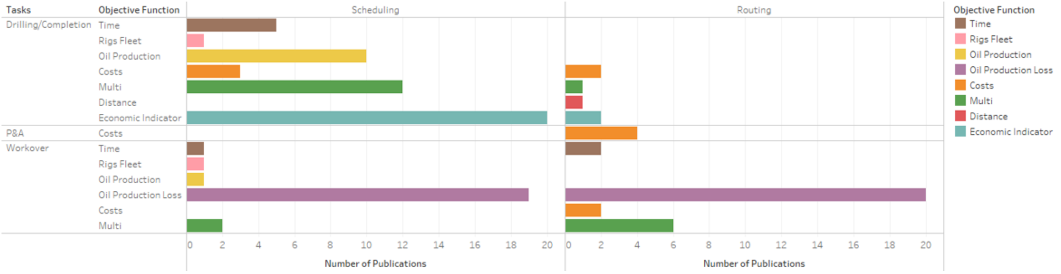


Figure 2.12: Relationship between objective function and task types.

Figure 2.12 allows us to observe the most common objective function type for each problem. Drilling rig scheduling problems usually consider economic (NPV, cash flow, CAPEX, *i.e.*, capital expediture, etc.), time (makespan, tardiness, or completion time) or oil production indicators, or a combination of these indicators using a multi-objective approach. However, drilling rig routing and scheduling problems usually consider a monetary objective function (cost or an economic indicator), as a routing problem often has transportation costs associated. As mentioned earlier, few studies were published with P&A operations. We can observe that all these use a routing approach and consider objective functions associated with costs, which is expected, as P&A operations involve more than one vessel and the distance between wells and ports in these operations is not negligible. Finally, most of the WRSP and WRRSP focus on minimizing oil production loss and some WRRSP might consider objective function with costs or multiple objectives. As mentioned by Attia et al. (2019), the oil and gas upstream is a multi-dimensional supply chain that requires multi-objective models to represent it properly. Therefore, future multi-objective approaches should also be considered in other RSPs.

Figure 2.13 presents the evolution of the heuristics in the RSP. The greedy randomized adaptive search procedure (GRASP) and the genetic algorithm (GA) are common heuristic methods for the RSP due to their general nature of implementation and flexibility. These are powerful well-known methods that, if appropriately designed, can provide good solutions reasonably fast. In some studies published between 2010-2020, the adaptive large neighborhood search (ALNS) metaheuristic has also achieved outstanding results for rig scheduling problems: Ribeiro et al. (2012b), Bissoli et al. (2014), Bissoli (2014), and Shaji et al. (2019).

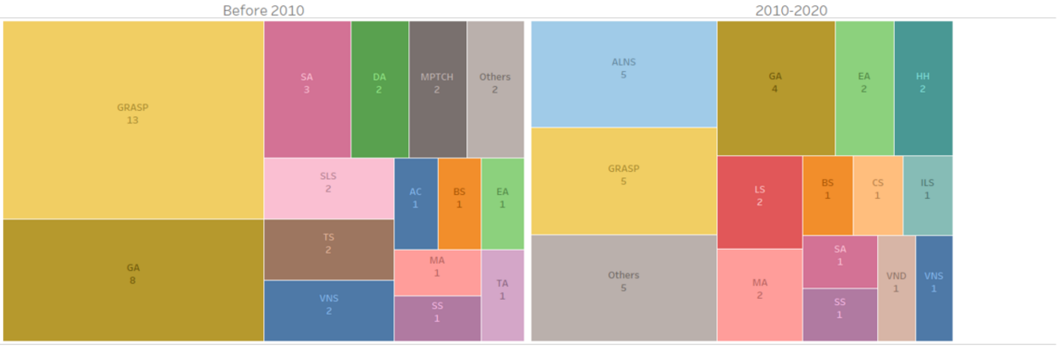


Figure 2.13: Evolution of the heuristics and metaheuristics methods for the rig scheduling problem.

In summary, refined techniques, such as hybrid optimization methods, sophisticated metaheuristics, and uncertainty models, are being used to approach complex problems involving multiple tasks, multiple types of resources, and realistic assumptions.

2.6 Perspectives and insights

As mentioned earlier, the RSP has several dimensions and levels. Therefore, it is important to analyze the previous trends and opportunities from different perspectives. In this section, we present insights according to the perspectives of uncertainty, data-driven, integrated field planning, and collaboration between academia and industry.

2.6.1 Uncertainty models

As mentioned by Suslick et al. (2009) and Santos et al. (2017), the RSP has emerged in a risky environment, with uncertainties related to geological aspects (structure, reservoir seal, and hydrocarbon charge), economic evaluations (rigs, transportation, and operation costs, chances of finding hydrocarbonates and producing economically viable reservoirs, and oil prices), development and production (infrastructure, production schedule, oil quality, operation times, and reservoir characteristic), and logistics (travel time, resource availability, inventories, and weather). Therefore, it is paramount that any analytical approach for the RSP considers the uncertainty that affects its decision-making.

Uncertainty in this context can be presented in the optimization method (e.g., simulation-optimization, when the uncertain parameter is simulated and then the simulation result is used as an input in the optimization) or in the modeling approach (e.g., optimization under uncertainty). According to Diwekar (2008), optimization under uncertainty can be divided into: (i) “wait and see”: when the decision is made only after the observation of the random values; (ii) “here and now”: optimization over some probabilistic measure, which includes stochastic and robust optimization; and (iii) “chance-constrained optimization”: when constraints that are not expected to be always satisfied. Chaari et al. (2014) propose a classification for scheduling under uncertainty, dividing the approaches into proactive (robustness measures, probabilistic methods, simulation, and optimization under uncertainty), reactive (priority rules), and hybrid (rescheduling or offline planning with real-time scheduling).

The most common approach in the RSP to assess the impact of uncertainties is through simulation models. de Andrade Filho (1994), Cong et al. (2008), Lasrado (2008), Litvak et al. (2011) used reservoir simulation tools to model the uncertainty associated with the reservoir, Gutleber et al. (1995) simulated the oil production for a given rig fleet availability, and Zahran and Al-Fardan (2014) used simulation to assemble scenarios for a rig scheduling system.

Hybrid methods combining simulation models and heuristics in a "here and now" approach are also standard. Onwunalu et al. (2008) simulated the oil production of drilling rig schedules, aiming to maximize the NPV in a hybrid GA. Bassi (2010) and Bassi et al. (2012) approached the WRSP using a GRASP heuristic in which the job processing time was estimated with Monte Carlo simulation. Flager (2014) also proposed a method for the onshore RSP that simulated the job processing times.

Other "here and now" approaches were presented by Fernández Pérez et al. (2018), Pérez et al. (2019). The authors proposed stochastic and robust optimization models for the WRSP considering the uncertainty in the processing time. The main difference between these models was the use of stochastic programming techniques to obtain globally optimal solutions. Bakker (2020) also uses linear programming methods to acquire optimal solutions, but there were no specific details about the uncertain variables considered. Silva and Silva (2018) presented a dynamic programming model for the WRRSP that deals with the uncertainty in the workover occurrences, that is, which wells will need workover and when.

Despite the number of RSP studies considering uncertainty, this trend has several gaps and opportunities. For instance, no chance-constrained optimization model has been proposed for the RSP, despite their potential of generating robust solutions. Furthermore, costs-related uncertainties are a critical feature for industrial stakeholders and should receive more attention. Lastly, there is the opportunity of employing data-driven optimization under uncertainty, which will be further discussed in the next section.

2.6.2 Data-driven models

Data-driven techniques are a recent trend in the literature that relies on the intelligent use of data. These techniques use machine learning, big data, and data science to analyze data and extract relevant, accurate, and valuable information to ease knowledge discovery and decision-making (Ning and You 2019). As a multifaceted problem, the RSP relies heavily on data from multiple sources. Data-driven techniques become crucial to extract information from this vast amount of data.

For instance, machine learning can be used for predicting when an intervention will become necessary or how long workovers will last. As mentioned by Carvalho et al. (2019), machine learning techniques (e.g., support vector machine, random forests, adaptive neural networks, deep learning, and k-means) have been successfully applied to design predictive maintenance applications on others fields. These promising developments involving predictive maintenance could be adapted to the context of RSP.

Ma et al. (2018) proposed a method that uses a data mining system to extract key information from daily drilling reports and historical data, identifies drilling opportunities, and uses it to optimize the short-term rig schedule. Castiñeira et al. (2018) used machine learning and natural language processing (NLP) for the automated analysis of drilling data. The historical data was then used to optimize the rig schedule through heuristics, maximizing NPV and oil production. Both studies have used advanced machine learning techniques to support the optimization, but none consider the uncertainty.

According to Ning and You (2019), data-driven optimization under uncertainty extracts rich information about uncertainty data in an automatic and smart data-driven process. It can be divided in: (i) data-driven stochastic programming and distributionally robust optimization (models the uncertainty using a family of probability distributions); (ii) data-driven chance-constrained optimization (focuses on chance-constraint satisfaction under the worst-case probability); (iii) data-driven robust optimization (a particular case of robust optimization); and (iv) data-driven scenario-based optimization (does not require knowledge of the probability distribution and uses a discrete uncertainty set). Data-driven optimization under uncertainty is a trend in the field of Operations Research that has not been employed in the RSP and is a clear opportunity for future studies.

2.6.3 Integrated planning

As mentioned earlier, the RSP is a multifaceted problem that depends on and affects other E&P decisions and resources, such as OSVs, equipment, and crews. According to Tavallali et al. (2016), only a few references addressed the integration of well placement with drilling scheduling. On the other hand, in Section 2.4.4, 51 studies were found with integrated decisions, of which 24 were from field planning and the others 27 integrating rig scheduling with decisions comprising other resources. This finding suggests that there has been a trend for models with an integrated planning perspective.

In Section 2.5, these studies were analyzed, and some trends and opportunities were detected. For instance, optimizing the RSP with other resources is a trend that has gained more traction between 2010 and 2020, as shown in Figure 2.8. 2.10 showed that very few studies had an integrated perspective for onshore fields, which could be an opportunity for future studies. However, it is important to notice that most field planning studies presented in Section 2.4.4.1 tackled the RSP with simplified assumptions. Tavallali et al. (2015), Martin et al. (2010), and Calderón and Pekney (2020) were the only references to consider the RSP with more realistic assumptions and integrated with the field development planning. Future opportunity still exists in tackling the field planning problem with constraints more closely representing industrial stakeholders' goals. This collaboration between the academia and industry will be discussed in the next section.

2.6.4 Industry and academia collaboration

Oil companies often collaborate with academic institutions to develop advanced decision-support frameworks and achieve better results in their decision-making. As many others decisions in the E&P, the RSP was born from this type of collaboration. However, while the number of studies using public data by academic stakeholders has grown exponentially, the collaboration between the industry and academia has reduced in number considerably between 2010 and 2020 (Figure 2.8). This cutback suggests a widening gap between the industry demands and the academia research agenda. Aiming to understand this gap better, this section analyzes RSP studies from both industrial and academic perspectives.

Eagle (1996) implemented a SA algorithm to schedule drilling rigs maximizing the NPV for BP (British Petroleum, Alaska), which led to savings of 30 million dollars. Hasle et al. (1996) developed a system with Saga Petroleum and other companies for scheduling well activities on rigs and wire-line cranes. The system used constraint programming and considered some technological precedence constraints in the model. Furthermore, it had an efficient and easy user interface fully integrated with other systems from the company. Currie et al. (1997a) and Currie et al. (1997b) implemented a linear programming model for a Norwegian company to integrate and optimize the development decisions. Another resource planning system was developed by Horton and Dedigama (2006) for the Australian oil company Santos Ltd. According to the authors, the system, named PRISM (Plan Resource Implement Schedule Manage), was validated and resulted in gains of approximately 450 million Australian dollars per year. Irgens and Lavenue (2007) presented a system called Aris that was implemented in Saudi Aramco. This software consisted of an interactive application for drilling rig scheduling minimizing costs that reduced traveling costs by 35%. Another system with an integrated field planning approach was discussed by Davidson et al. (2009). The authors presented Exxon Mobil's simulator EM^{power}, which has a module for drilling rig scheduling using first-in-first-out rules. Another system used by Saudi Aramco was reviewed by Amer et al. (2016), ASAS (Automated Services Assignment System) was used for scheduling and monitoring the fleet of drilling and workover rigs with potential savings of 15% in contrast with the previously employed process. Lastly, Ma et al. (2018) presented a decision support technique that was successfully applied in several major oil fields from the Middle East, North America, and South America. This tool used natural language processing and deep neural network models to extract information from daily drilling reports and historical data and to detect non-productive time in the short-term rig scheduling, reducing it between 14-30%.

In summary, describing the main gaps between the academic literature and industrial practice. From the 130 RSP studies found in this SLR, only nine studies presented a method that was implemented in a real company, and from those nine, only five were verified within the company as a successful tool. Most of these studies were integrating the RSP with other decisions and resources. Not only the approach ought to have an integrated perspective, but when proposing a system, a critical factor for implementation is that the system can be easily integrated with other systems (Hasle et al. 1996). Finally, researchers must use realistic constraints and objective functions for the problem. As mentioned by Tavallali et al. (2016), considering technical constraints and removing unreasonable simplifying assumptions are critical for gaining industrial acceptance.

2.7 Final considerations

To summarize, the RSP has several characteristics that are important when approaching a problem. This study has proposed a methodology that divides problem characteristics into two groups: setting (oil field location, well operations, planning level, and resources considered) and approach (modeling, rig's fleet, single/multiple jobs). According to this taxonomy, a study can be classified into Drilling Rig Scheduling Problem, Workover Rig Scheduling Problem, Workover Rig Routing and Scheduling Problem, Field Planning, and Resource Planning. This taxonomy and classification were used to analyze the papers selected in the review, detecting trends and opportunities.

The first publications were in the 1960s and 1970s and refer to drilling rigs scheduling problems from an oil field development perspective. However, only in the second half of the 1990s, the discussions started to gain traction. New mathematical models were proposed, but they were still not suitable for real-world instances. In the 2000s, there was an increase in the research of new and more efficient methods, mainly focusing on workover rigs. The improvement in the algorithms' and models' performances allowed for approaching more realistic scenarios. We also highlight that there are very few P&A studies.

The current trends identified in the RSP literature take advantage of the sophisticated optimization techniques available in the literature. Developing hybrid methods (combining mathematical programming, metaheuristics, or simulation) and considering multiple types of tasks (drilling, completion, or workover), other resources (OSVs, equipment, and crews), and realistic objective functions and assumptions (such as heterogeneous fleet, variable costs rates over the time horizon, rigs capabilities learning, fleets availability, machine eligibility, net present value, and expected monetary value). Furthermore, the dynamic and high-risk operations of the RSP demand studies considering uncertainty; only a few approaches use stochastic and robust models, dynamic programming, simulation-optimization, or data-driven optimization.

Many of these gaps and trends emerge from the needs of the oil and gas sector. The rig scheduling problem is a crucial decision in the E&P phases, considerably influencing the field profitability. Nevertheless, there were only a handful of studies that were implemented or validated in the industry. Future studies on the RSP ought to apply the advanced quantitative methods available in real instances, validating it with industry stakeholders and integrating academic and practical perspectives.

The SLR and RSP classifications that were presented in this chapter culminated in the publishing of an article "A Systematic Literature review for the rig scheduling problem: Classification and state-of-the-art" in the journal Computers & Chemical Engineering (Santos et al. 2021).

In this thesis, we approach the offshore WRSP and try to address some of these gaps. First, we address real-life instances, using sophisticated optimization techniques with realistic objective functions and assumptions, and validating the results with industry players. Second, we propose creative and innovative data-driven optimization models considering the uncertainty and the particularities, adapted to the industry demands. The methodologies proposed in this thesis are presented in the next section as well as the review of the optimization and data-science concepts involved in these type of models.

3
Technical review and basic concepts

In the previous chapter, the rig scheduling problem was studied through a systematic literature review, which led to several insights. From the problem perspective, the workover planning has received significant attention, but the vast majority of its studies consider only onshore wells. Another important gap detected was the lack of implemented and validated models in the industry. As to the approach perspective, a trend for optimization models considering uncertainty was detected, with very few data-driven optimization models having been made so far and none of them considered chance-constrained approaches.

Aiming to fulfill these gaps, this thesis approaches the offshore workover rig scheduling problem in real-life based instances through data-driven chance-constrained optimization. Four data-driven optimization methodologies for the workover rig scheduling problem are proposed and compared in this thesis. The main methodology is a data-driven joint chance-constrained (DD-JCC) optimization model, which is compared with others data-driven chance-constrained variations (integrated chance-constrained and budget-constrained) and a deterministic data-driven optimization model. All data-driven optimization models use regressions to estimate the unknown parameter based on historical data. These data-driven methodologies are separated in three major phases: data treatment (where data science methods are used to clean and classify the data), prediction (where predictive models are used to estimate uncertain parameters of the optimization, and optimization (where optimization models are generated and solved).

With the purpose of enhancing the reader's understanding of these methodologies, this chapter presents a review of the techniques and basic concepts related to the methodology. This technique overview is divided according to the three phases of the proposed methodology: data treatment, prediction, and optimization. Section 3.1 discuss the data science techniques used to clean and classify the data in data-driven optimization. Section 3.2 overviews the regression models used to predict the uncertainty in the data-driven optimization. Last, Section 3.3 discuss the different ways to optimize under uncertainty, their data-driven optimization ramifications, and the basic concepts and the state-of-art of the data-driven chance-constrained optimization.

3.1 Data treatment and grouping algorithms

Data science is a crucial part of any data-driven model. According to Shcherbakov et al. (2014), data science is a result of the expansion of statistics to practical problems with large volumes of data through computer science methods. Data science can be used to extract information or treat, classify, group, and predict data. Next, we describe some techniques commonly used in text mining and classification algorithms that will be used in this thesis.

3.1.1 Text mining

Text mining is a technique used in unstructured text data to clean the data and extract critical information from it (Shcherbakov et al. 2014). It is separated into two main phases:

- *Data cleaning:* 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:* the 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). 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. An example of this similarity recognition failure is for 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.

Another important concept in text mining is string similarity and distance, which measures how close the sentences of a text data are to each other. Two examples of string similarity measures are the Levenshtein (LV) (Yujian and Bo 2007) and the Longest Common Substring (LCS) (Sun et al. 2015) distances. 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.

3.1.2 Clustering algorithms

Clustering methods divide data into subsets in such a way that similar examples are grouped together and distinct examples are divided into various groups (Rokach and Maimon 2005). Clustering can be used to group data according to a common and symmetric distance measure.

A common clustering technique is the k-means algorithm (Likas et al. 2003b), a partition method that separates the data into a pre-defined 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. 2003c). The k-means algorithm is described in the following pseudo-code:

Algorithm 1: Pseudo-code for the k-means algorithm (Likas et al. 2003a)

```
Data: $D = \{t_1, t_2, \dots, t_n\}$ = Set of n points; k = Number of clusters
Result: K (set of clusters)
$\mu_1, \mu_2, \dots, \mu_k \leftarrow$ initial values;
while convergence criteria not met do
    Assign each $t_n$ to $k = \arg\min distance(t_n, \mu_k)$.;
    Recalculate new means $(\mu_1, mu_2, \dots, \mu_k)$ for each cluster $k$.;
```

A crucial part of the k-means algorithm is defining the number of clusters (k), which is usually performed 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).

3.2 Predictive modeling

Another popular application of data science is to predict the outcome of future events. For that, predictive modeling techniques are used to analyze patterns in a given set of input data and model it as a mathematical process

(Kuhn et al. 2013, Laud and Ibrahim 1995). It is a crucial component of predictive analytics, a type of data analytics that uses current and historical data to forecast activity, behavior, and trend. There are two types of predictive models: linear or parametric models and non-linear or non-parametric regression models. Linear regression models assume that the predicted variable is a linear combination of the input variables. Meanwhile, non-parametric models do not make strong assumptions about the form of the mapping function, allowing non-linear combinations of the input variables.

Linear regression models are parametric statistical models used to determine the relationship between a response variable (Y) and its explanatory variables (X) or to predict its value using other variables. A typical class of these models is the generalized linear model (GLM), which is a generalization of the ordinary linear regression models accepting response variables with errors following an exponential family distribution, not necessarily a normal distribution as the ordinary models (Nelder and Wedderburn 1972). The GLM's predicted value of the observation Y_i is a linear sum of the effects of one or more explanatory variables X_{ij}, as shown in the equation:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_j X_{ij} + \cdots + \beta_J X_{iJ} + \epsilon_i \qquad \forall i, \tag{3-1}$$

where $j = 1, \ldots, J$ denotes the number of explanatory variables (X_{ij}) used and β_j represents their effect on the response variable Y_i (McCullagh and Nelder 2019). As mentioned earlier, in the GLM, the error variable ϵ follows a distribution of the exponential family, which includes the Normal, Poisson, Binomial, and Gamma distributions. The linear coefficients are estimated using the maximum likelihood estimation (MLE) method if the residuals are non-Normal or the ordinary least squares (OLS), if Normal (Yuan and Yang 2005, Yan and Su 2009, Mahmoud 2019). However, if there is a large number of dummy variables; as a result, a large number of coefficients, the model can overfit on 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 (McDonald 2009). One of these techniques is the ridge regression. The ridge regression models are a multiple regression analysis 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 regressions add a degree of bias to the regression estimations, reducing the standard errors. This technique is recommended for regression models with near-linear relationships among the independent variables or many dummies independent variables (Hoerl and Kennard 1970).

To train and test any regression model, it is recommended to separate the data into in-sample and out-of-sample, where the in-sample data is used to train the regression model and the out-of-sample data to predict and evaluate the trained models. Linear regression models can be trained using several methods, such as iteratively reweighted least squares (IWLS) (Street et al. 1988) and K-fold cross-validation (Bengio and Grandvalet 2004). Possible metrics to evaluate the results of a linear regression model in the out-of-sample data are the root-mean-square error (RMSE), R-squared (R^2), and p-value fit for residuals normally distributed.

Kernel smoothing are non-parametric statistical methods used to estimate random variables without specifying their distribution (Racine 2008). An example of kernel smoothing is the kernel density estimation (KDE). This technique is used in mathematical programming applications to represent the uncertainty when its distribution is unknown (Calfa et al. 2015, He and Li 2018, Wang and Li 2017)

Basically, the KDE uses a kernel function $K(u)$ to weight the data points of a histogram and estimate the density function as in Equations (3-2) and (3-3) and Figure 3.1.

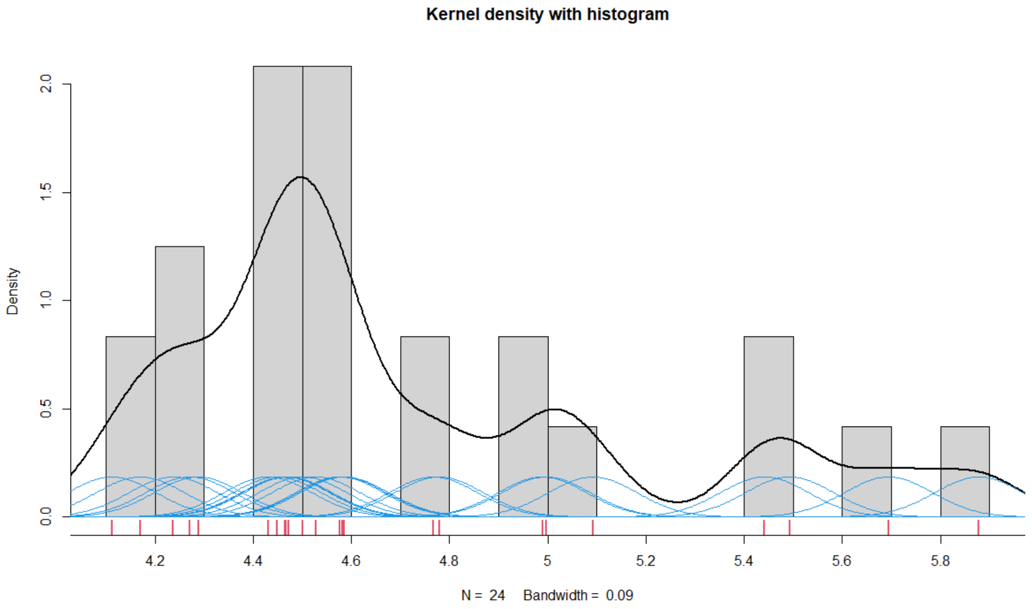


Figure 3.1: Example of a kernel density estimation (black line), the kernel functions (blue lines) for each data point (red marks), and the histogram (grey bars).

Suppose X is a continuous random variable that has an unknown probability density function (PDF) $f(x)$, cumulative probability function (CDF) $F(x)$, and a sample of n data points of (X_i). The KDE can be represented as a mathematical function, (3-2) and (3-3), for a given x:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{3-2}$$

$$\hat{F}(x) = \frac{1}{nh}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x - X_i}{h}\right), \tag{3-3}$$

where h is the bandwidth (a scaling factor), $K(u)$ is the kernel function that weights the data points, and $\mathcal{K}$ is the integrated kernel function. As mentioned, the bandwidth h is an essential parameter of the KDE that sets the smoothness or roughness of the estimation. A larger h results in a smoother curve, and a smaller h makes the density rougher, as illustrated in Figure 3.2.

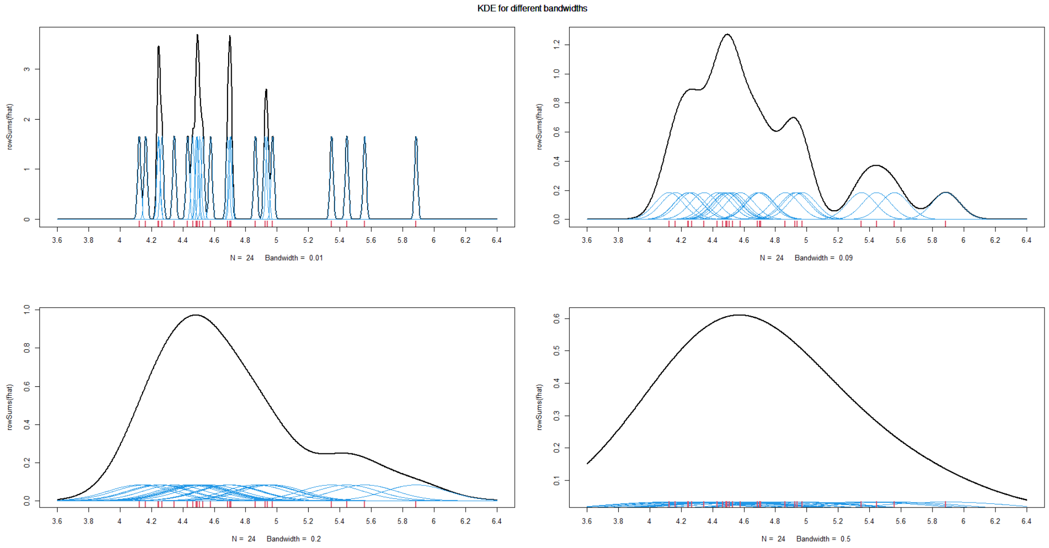


Figure 3.2: Examples of the effect of different bandwidths in the kernel function (blue lines) and the KDE (black lines).

Another important setting in the KDE is the choice of the kernel function $K(u)$ to be used. Several kernel functions are possible, such as: Gaussian, Box, Tri, Triweight, Epnechnikov, and Tri-cube. As mentioned by Calfa et al. (2015), an example of a Gaussian kernel function equal is the standard normal:

$$\mathcal{K}_{Gaussian}(u) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{u}{\sqrt{2}}\right)\right] \tag{3-4}$$

$$\text{where: erf}(u) = \frac{1}{\sqrt{\pi}}\int_{-u}^{u} e^{-\frac{t^2}{2}}\, dt. \tag{3-5}$$

As mentioned by Calfa et al. (2015), the choice of the bandwidth rather than the kernel function type has more effect in obtaining accurate results with the KDE. There are four methods to select the optimal bandwidth: rule-of-thumb, plug-in, least-squares cross-validation, and likelihood cross-validation (Calfa et al. 2015). The cross-validation-based methods are fully automatic or data-driven as they are tailored to the sample data under consideration, being more recommended for data-driven optimization models than the others. Calfa et al. (2015) used the *R np* package from Hayfield and Racine (2008) and Racine (2008) in a data-driven joint chance-constrained optimization model that can be applied to many other problems, including scheduling. Their studies will be discussed later in Section 3.3.6.

3.3
Optimization under uncertainty and data-driven methods

There are several approaches to optimize under uncertainty. According to Diwekar (2008), optimization under uncertainty can be divided into:

- "wait and see": when the decision is made only after the observation of the random values.
- "here and now": optimization over some probabilistic measure, which includes stochastic and robust optimization;
- "chance-constrained optimization": a particular case of robust optimization when constraints are not expected to be always satisfied.

The first approach requires distinguished knowledge of the uncertainty, which is quite rare. The second is divided into stochastic and robust techniques (Wets 2002). Particularly for scheduling and fleet sizing problems with uncertainty in the duration, stochastic programming solutions are usually strongly affected by extreme scenarios with a low-probability, as the model tries to find feasible solutions for all scenarios (Prékopa 2013). Robust optimization tries to generate solutions that are good in most cases, as long as it is feasible for all scenarios (Ben-Tal et al. 2009). Chance-constrained optimization arises as an alternative to the here-and-now standpoint that allows some infeasible solutions within a predefined threshold, which is recommended for fleet sizing and long-term models in which the solution feasibility allows some flexibility (Prékopa 2003, Verderame et al. 2010). This section explains the basic concepts of chance constraints and its main relaxations and reformulations (Section 3.3.1) and reviews the state-of-art of data-driven optimization (Section 3.3.4) from data science algorithms (Section 3.1) to several data-driven chance-constrained optimization approaches (Sections 3.3.5 and 3.3.6).

3.3.1 Chance-constrained optimization: basic concepts

Mathematical programming under probabilistic constraints, also known as chance-constrained programming or probabilistic programming, refers to a type of optimization under uncertainty in which the objective function is subject to at least one probabilistic constraint, *i.e.*, a constraint that must be satisfied within a certain probability. This type of optimization was first introduced by Charnes and Cooper (1959) and Charnes et al. (1958) for the problem of scheduling heating oil production. The authors proposed probabilistic constraints that were imposed individually on each constraint affected by the random variables.

Later, Miller and Wagner (1965) presented a formulation for constraints with joint probability with random variables on the right-hand side of the constraints. Finally, Prékopa (1971, 1973) proposed the general formulations for chance-constrained optimization with joint probabilistic constraints and stochastically-dependent random variables.

Basically, there are two types of probabilistic constraints: *individual chance-constrained (ICC)* and *joint chance-constrained (JCC)*. In the first case, probability thresholds are individual for each constraint. The second type is when at least two constraints must together satisfy a joint probability (Ahmed and Shapiro 2008, Prékopa 2015, 2003, Li and Li 2015). General forms of ICC and JCC are shown in (3-6) and (3-7), respectively:

$$\mathbb{P}\left[g_j(x,\tilde{\xi}_j) \geq 0\right] \geq \alpha_j, \qquad \forall j = 1,...,m \tag{3-6}$$

$$\mathbb{P}\left[g_j(x,\tilde{\xi}_j) \geq 0, \forall j = 1,...,m\right] \geq \alpha, \tag{3-7}$$

where $\mathbb{P}[\cdot]$ is the probability function of the random variables $\tilde{\xi}_j$, x represents the vector of decision variables, $g_j(x,\tilde{\xi}_j)$ is a function of the decision variables and the random variables, and α is the confidence level, reliability level, or risk level. In the ICC, each individual chance constraint is associated to an individual α_j, *i.e.*, the probabilistic constraint must be satisfied individually. Alternatively, all the joint chance constraints must satisfy this confidence level (α) simultaneously. Value-at-Risk (VaR) constraints are an example of chance constraint commonly used in finance and economics (Cui et al. 2013, Zhao and Xiao 2016).

A particular case of chance-constrained models occurs when the uncertainty is on the right-hand side of the equation, which is presented in (3-8) and (3-9):

$$\mathbb{P}\left[g_j(x) \geq \tilde{\xi}_j\right] \geq \alpha_j, \qquad \forall j = 1, ..., m \tag{3-8}$$

$$\mathbb{P}\left[g_j(x) \geq \tilde{\xi}_j, \forall j = 1, ..., m\right] \geq \alpha, \tag{3-9}$$

where $g(x)$ is a function of the decision variables and $\tilde{\xi}$ is the vector of random variables. Note that in these cases the decision variables and random variables are separated on the left-hand side (LHS, the variables side of the constraints) and the right-hand side (RHS, the constant side of the constraints), respectively.

Another common feature in probabilistic programming is to penalize the constraint violations in the objective function. For that purpose, several measures of violations can be used separately or with chance constraints. As mentioned by Prékopa (2003), a hybrid model with probabilistic constraints and penalization of constraints violations would be as follow:

$$\text{Min } c^T x + \sum_{j \in J} q_i \mathbb{E}\left([\tilde{\xi}_j - g_j(x)]_+\right) \tag{3-10}$$

Subject to

$$\mathbb{P}\left[g_j(x) \geq \tilde{\xi}_j\right] \geq \alpha_j, \qquad \forall j = 1, ..., m \tag{3-11}$$

$$Ax \geq b, x \geq 0 \tag{3-12}$$

where q_i are non-negative constants that penalize violations of the probabilistic constraints and $[\cdot]^+$ is the non-negative approximation of a value.

A classic measure of violation is the *integrated chance constraint*. This concept was introduced by Haneveld (1986) and instead of guaranteeing the risk level of the probabilistic constraint, the integrated chance constraints assure that the expected magnitude of violation is lower or equal to a specific bound. Following Prékopa (2003), the integrated chance constraint general form is represented in (3-13) and is used with or without the chance constraint.

$$\mathbb{E}\left(\max_i \left[g_i(x, \tilde{\xi})\right]_+\right) \leq d \tag{3-13}$$

The next section presents common reformulations and relaxations in the literature for representing the probabilistic constraints of the chance-constrained models.

3.3.2 Chance-constrained optimization: reformulations and approximations

As chance-constrained models are generally non-linear, non-convex, and, consequently, extremely difficult to solve, many researchers have analyzed their structural properties, identifying particular cases in which the probabilistic constraints are convex or proposing reformulations, approximations, or relaxations.

Special cases in which the chance constraints can be represented with convex reformulations were identified by Charnes and Cooper (1963), Prékopa (1973), and Calafiore and Ghaoui (2006). Charnes and Cooper (1963) used decision rules to obtain, under specific conditions, deterministic equivalent formulations for three classes of optimizing objectives functions with individual chance constraints: maximum expected value (E model), minimum variance (V model), and maximum probability (P model). Meanwhile, Prékopa (1973) studied the convexity of individual and joint chance constraints with log-concave probabilities. Last, Calafiore and Ghaoui (2006) convert individual chance constraints with radial distributions to convex second-order cone constraints.

Some conservative reformulations use p-efficient concepts. The p-efficient concept was introduced by Prékopa (1990) and involves the definition of sufficient conditions for the probabilistic constraint to be feasible. These conditions, or p-efficient points, can be represented as a set of deterministic constraints, allowing deterministic inner approximations of the problem. Other reformulations using this concept for individual and joint chance constraints have been proposed by: Dentcheva et al. (2000), Lejeune and Ruszczyński (2007), Lejeune and Noyan (2010), Saxena et al. (2010), Lejeune (2012), and Dentcheva and Martinez (2013).

There are two other popular reformulations of chance constraints mentioned by Lejeune and Prékopa (2018). The first uses scenario-based reformulation with binary and knapsack constraints (Ruszczyński 2002). Another one is based on pattern-and-boolean programming, as in Lejeune (2012), Kogan and Lejeune (2014), and Lejeune and Margot (2016).

Using the properties of normal and log-normal distributions, Biswal et al. (2005) and Sahoo and Biswal (2005) proposed reformulations for the joint constraint programming when the uncertainty in the RHS follows normal and log-normal distributions. However, these formulations lead to nonlinear joint constraints, as shown in the constraints (3-14)-(3-17) for the case in which the RHS uncertainty is made by independent random variables that follow a

Normal distribution $N(\mu, \sigma)$:

$$\mathbb{P}\left[g_j(x) \geq \tilde{\xi}, \forall j\right] \geq \alpha \tag{3-14}$$

$$\prod_j \mathbb{P}\left[\tilde{\xi} \leq g_j(x)\right] \geq \alpha, \tag{3-15}$$

$$\prod_j \mathbb{P}\left[\frac{\tilde{\xi} - \mu}{\sigma} \leq \frac{g_j(x) - \mu}{\sigma}\right] \geq \alpha, \tag{3-16}$$

$$\prod_j \left[\Phi\left(\frac{g_j(x) - \mu}{\sigma}\right)\right] \geq \alpha, \tag{3-17}$$

where the RHS uncertainty $\tilde{\xi} \sim N(\mu, \sigma)$, $g_j(x)$ is the LHS, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, which is equal to $\Phi(x) = \mathbb{P}(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\inf}^{\inf} e^{-\frac{u^s}{2}} du$.

Most of these reformulations are only for individual chance constraints. Thus, the reformulations that are applicable to joint chance-constrained models require several conditions and complex calculations that reduce the flexibility of the model to be re-adapted when the problem's assumption or instances change. An alternative is through the approximation of the chance constraints.

Nemirovski and Shapiro (2007) proposed a Bernstein-based approximation of the chance constraints, which is convex and efficiently solvable. However, Bernstein-based approximations, despite being easy to use and solve, often generate too conservative solutions, as mentioned by Bertsimas and Sim (2004), Nemirovski (2012), and Zhao and Kumar (2017). Lejeune and Prékopa (2018) review several reformulations of joint chance constraints and proposed approximations using bounding schemes (based on Boole-Bonferroni and product-type inequalities, binomial moments, and step-wise dependence concepts) for problems with continuous random variables.

Another approximation of the chance constraints is through scenario-based reformulations. Calafiore and Campi (2006, 2005) sampled the probabilistic constraints to obtain a standard convex optimization problem that satisfies the probabilistic constraint confidence level. Nemirovski and Shapiro (2007) combined this scenario-based reformulation with the Bernstein approximation in a simulation-based method. As mentioned by Luedtke and Ahmed (2008), these previous approximations are conservative methods that are attractive only when probabilistic constraint confidence level is high ($1-\alpha$ is very small, such as 10^{-6}. Conservative approximations with lower confidence levels, *i.e.*, a bigger $1 - \alpha$ (such as 1%), often cannot to measure how much worse the objective function is relative to the optimal value of that risk level. Aiming to tackle this issue, Luedtke and Ahmed (2008) proposed a sample approximation approach and a scenario-based reformulation for chance-constrained optimization. The authors also proposed a measure for the approximation error to be used to set the scenario sample size of the optimization model. This sample approximation was later used by Nikzad et al. (2019) to propose a two-stage stochastic optimization reformulation for a chance-constrained optimization model in a medical drug inventory routing problem.

3.3.3 Scenario generation and Wasserstein distance

As mentioned earlier, some chance-constraints reformulations use scenarios to represent the uncertainty and solve it as a stochastic programming model. The problem of scenario-based optimization is that the models' size increases considerably with the number of scenarios. To reduce the size of the models, scenario reduction is often used with the scenario generation method. This section focuses on scenario generation using Monte Carlo sampling (MCS) and the scenario reduction technique using the Wasserstein distance. Note that scenario reduction methods are sometimes referred to in the literature as scenario set generation algorithms (Fu et al. 2017). Kaut (2021) suggested several benefits of using Wasserstein distance-based scenario reduction algorithms that will be explained later in this section.

MCS is a sampling technique that generates N replications of a random vector ξ using its probability distribution. This is achieved by generating a random sequence of independent numbers between [0,1] and constructing a sample by applying an appropriate transformation using the probability distribution (Shapiro 2003).

After generating a large number $|S|$ of scenarios, the Wasserstein distances between the scenario sets are calculated and used to select a set of $|S'|$ scenarios that minimizes the Wasserstein distances, *i.e.*, the best set of $|S'|$ scenarios to represent S. This set can be obtained by redistributing the original probabilities from S to S', which is to minimize the amount of moved probability multiplied by the moving distance, represented in Equation (3-18):

$$\text{Min} \sum_{i\in J}\sum_{j\in J} ||D_{ij}||^r \pi_{ij} \tag{3-18}$$

Subject to:

$$\sum_{j\in J} \pi_{ij} = P_i \qquad \forall i \tag{3-19}$$

$$\sum_{i\in J} \pi_{ij} = P_j \qquad \forall j, \tag{3-20}$$

where $||D_{ij}||$ is a distance metric to compare the scenarios, r is the Wasserstein distance order, π_{ij} is the amount of probability moved from i to j, P_i is the probability of each scenario obtained in the scenario generation. Constraints (3-19) and (3-20) are transportation constraints. Note that the Wasserstein distance problem is a linear transportation problem in which there are several algorithms available to solve it in an efficient manner.

As mentioned by Kaut (2021), k-means and Wasserstein-distance-based methods can both be used in scenario generation. To compare and evaluate the scenario generation, Kaut and Stein (2003) and Pflug (2001) suggest using the error of approximating measure, which is the difference between the value of the true objective function at the optimal solutions of the actual process and the approximated problems. Next, we discuss data-driven optimization and review some studies on it.

3.3.4 Data-driven optimization: state-of-the-art

Optimization under uncertainty relies on the quality of the data (Bertsimas and Sim 2004). As mentioned by Ning and You (2019), a wide variety of machine learning and data science methods can be used to analyze uncertainty data and extract accurate, relevant, and useful information for decision-making. The combination of these techniques with optimization is known as data-driven optimization. Data-driven optimization approaches have emerged in recent years aiming to formulate the uncertainty model based on the data (Bertsimas et al. 2018).

Ning and You (2019) reviewed the data-driven optimization under uncertainty literature and classified the studies into four categories:

- Distributionally robust optimization (DRO), also known as data-driven stochastic programming.
- Data-driven chance-constrained programming.
- Data-driven robust optimization.
- Data-driven scenario-based optimization.

Data-driven stochastic programming or distributionally robust optimization model the uncertainty through an ambiguity set, *i.e.*, a family of probability distributions that accurately represent the uncertainty in the data. Distributionally robust optimization models mitigate the effects of a worst-case distribution by constructing an ambiguity set that does not assume a single uncertainty distribution, but instead an uncertainty set of probability distributions estimated through statistical inference and data science (Ning and You 2019).

Data-driven chance-constrained (DCC) models, on the other hand, concentrate on satisfying the chance constraints in their worst-case probabilities. Instead of optimizing the worst-case expected objective as the DRO approaches, DCC models optimize under uncertainty in the probability distributions of the chance constraints (Ning and You 2019).

Data-driven robust optimization combines robust optimization with stochastic programming (Namakshenas and Pishvaee 2019). According to Zhang et al. (2022), robust optimization uses an uncertainty set to represent the uncertain knowledge and is divided into three types: statistic robust optimization (SRO), when all decisions are made at once before the uncertainty realization; two-stage adaptive or adjustable robust optimization (ARO); and multi-stage ARO. Conventional robust optimization models usually define the uncertainty set/model beforehand, without allowing enough flexibility to capture the structure and complexity of uncertainty data. Data-driven robust optimization arises by integrating the uncertainty set definition into the robust optimization model.

Data-driven scenario-based optimization refers to scenario optimization approaches for chance-constrained models. Unlike from the stochastic programming models, scenario-based optimization models do not require explicit knowledge of probability distribution. In the data-driven scenario-based chance-constrained optimization, uncertainty scenarios are used to achieve an optimal solution satisfying the chance constraints (Ning and You 2019). The following sections present some data science algorithms often used in data-driven optimization and details data-driven chance-constrained optimization approaches.

3.3.5 Data-driven chance-constrained optimization

As mentioned by Ning and You (2019), Calfa et al. (2015), and Ben-Tal et al. (2011), the DCC programming models are subjected to chance constraints that need to be satisfied under an ambiguity set (a family of probability distributions representing the uncertainty data). Their general form is the following:

$$\begin{aligned} &\min_{x \in X} f(x) \\ &\text{Subject to: } \min_{\mathbb{P} \in \mathcal{D}} \mathbb{P}\left[G\left(x, \xi\right) \geq 0 | \xi \in \Xi\right] \geq \alpha, \end{aligned} \tag{3-21}$$

where x is the vector of decision variables, ξ represents a random vector that follows a probability distribution $\mathbb{P}$ with an ambiguity set $\mathcal{D}$. The chance constraints are defined by the function $G(g_1, \ldots, g_m)$, using the decision variables X and the random vector ξ, and has to satisfy a risk level α.

Several data-driven chance-constrained approaches have been proposed in the optimization under uncertainty literature. Assuming that all distributions in the ambiguity set had equal and known mean and covariance, Calafiore and Ghaoui (2006) reformulated the distributionally robust individual chance constraints to convex second-order cone constraints. Considering first and second-order moment information of the chance constraint uncertainty, Zymler et al. (2013) used the worst-case conditional value-at-risk (CVaR) approximation to approximate the distributionally robust joint chance constraints. There were also some non-linear approaches for distributionally robust chance-constrained optimization with ambiguity sets constructed according to mean and variance (Yang and Xu 2016); mean absolute deviation (Postek et al. 2018); convex moment constraints (Xie and Ahmed 2018); distributions mixtures (Chen et al. 2018b, Lasserre and Weisser 2021); Kullback–Leibler divergence (Hu and Hong 2013); Wasserstein metrics (Ji and Lejeune 2018, Chen et al. 2018a, Hota et al. 2019, Xie 2019, 2021).

As this study is related to a joint chance-constrained problem, this review is focused more on the data-driven joint chance-constrained approaches. Zhang et al. (2016) used kernel smoothing models to reformulate the JCC model and obtain better solutions for their problems. Zhang et al. (2016)'s reformulation applies kernel smoothing in the robust approximation from Bertsimas and Sim (2004). However, this type of robust approximation leads to strongly conservative solutions. Jiang and Guan (2016) proposed DCC models with ambiguity set constructed using ϕ-divergence. Calfa et al. (2015) extended the models using the kernel smoothing method to avoid conservative solutions, obtaining non-linear programming models for the JCC scenarios.

Another type of data-driven CC approach is the contextual chance-constrained model, which was proposed by Rahimian and Pagnoncelli (2020). Unlike traditional stochastic programming models, contextual chance-constrained programming does not ignore the dependence on multi-dimensional features, *i.e.*, the auxiliary information connected to the random variables. Rahimian and Pagnoncelli (2020)'s formulation can use kernel or machine learning (ML)-weights to consider the auxiliary data. Their kernel-based approach cannot be solved in most cases, especially when continuous and categorical variables are combined in the data. The ML-weights approach is formulated for k-nearest neighbors (kNN), CARTs (classification and regression trees), and random forest techniques. The authors use the proposed contextual CC programming to give emphasis on data points closer to each observation, being recommended in data-driven CC in which the historical data changes the behavior over time. Sometimes, the uncertainty data and its features are subjected to the decision-making, *i.e.*, the realization of the uncertainty is dependent on the decision variables and the auxiliary data. Bertsimas and McCord (2018) proposed an algorithm using machine learning methods for data-driven optimization problems that fall in this case. Finally, Bertsimas and Kallus (2020) studied the formulations, theorems, and properties of the conditional stochastic optimization problems using several ML methods, such as Kernel, kNN, locally weighted least squares (LOESS), and decision trees, suggesting that ordinary linear regression and ridge regression models could also be used to capture the features dependence.

3.3.6 Kernel-based joint chance-constrained optimization

As mentioned in the last section, several data-driven joint chance-constrained approaches have been proposed in the optimization under uncertainty literature. Calfa et al. (2015), Jiang and Guan (2016), and Zhang et al. (2016) used kernel smoothing models to reformulate the model and obtain better solutions for their problems. Zhang et al. (2016)'s reformulation applies kernel smoothing in the robust approximation from Bertsimas and Sim (2004). However, this type of robust approximation leads to strongly conservative solutions. This section details Calfa et al. (2015)'s reformulation, which is based on Jiang and Guan (2016), and does not lead to over-conservative solutions.

Consider the following joint chance constraint:

$$\mathbb{P}\left[g_{ij}(x) \leq \tilde{\xi}_i, \forall j\right] \geq \alpha \qquad\qquad \forall i | i \neq 0. \qquad (3\text{-}22)$$

According to Calfa et al. (2015), if the distributions for $\tilde{\xi}_i$ are independent, uncorrelated, and each one follows a Gaussian model, constraint (3-22) can be reformulated using kernel distribution estimation properties as:

$$\sum_{l=1}^{L} \prod_{j} \left[\mathcal{K}_i\left(\frac{g_{ij}(x) - \hat{\xi}_i^l}{h_i}\right)\right] \geq \alpha'_+ \qquad\qquad \forall i, \qquad (3\text{-}23)$$

where $\hat{\xi}_i^l$ are data points of the uncertainty, α' is a reduced risk level, $\alpha'_+ = \max\{\alpha', 0\}$, h_i is the bandwidth selected for the kernel estimation of uncertainty for i, and $\mathcal{K}_i(\cdot)$ is the kernel or weighting function estimated for i. If we use the Gaussian kernel:

$$\mathcal{K}_{Gaussian}(u) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{u}{\sqrt{2}}\right)\right], \text{ where } \operatorname{erf}(u) = \frac{1}{\sqrt{\pi}}\int_{-u}^{u} e^{-\frac{t^2}{2}} dt. \tag{3-24}$$

This will lead to the following equations:

$$\sum_{l=1}^{L}\prod_{j}\left[\frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{g_{ij}(x) - \hat{\xi}_i^l}{h_i\sqrt{2}}\right)\right] \geq \alpha_+^{'} \qquad \forall i \tag{3-25}$$

According to Calfa et al. (2015), the reduced risk levels $\alpha_+^{'}$ can be estimated through the Kullback–Leibler (K–L) divergence formula: $\alpha_+^{'} = \inf_{x\in(0,1)} \frac{e^{-d}x^{1-\alpha}-1}{x-1}$, which is explained and proved in Jiang and Guan (2016). Last, the bandwidths can be selected using the algorithms available in R packages, as mentioned in Section 3.2.

The advantages of this kernel-based reformulation shown in Constraint (3-25) are that the distributions are estimated based on the available data, and the risk levels are directly associated with the quality and size of the data-set, allowing to generate data-driven solutions.

However, this deterministic-equivalent reformulation of the joint chance-constrained optimization model leads to a non-linear equation that is also non-convex. As a result, the kernel-based JCC becomes a non-convex mixed-integer non-linear programming (MINLP) model, which is known to be one of the most challenging classes of mathematical programming problems to solve (D'Ambrosio et al. 2012). Therefore, other alternative methods should be developed. In the next section, we detail the proposed data-driven joint chance constraints programming methodologies, which use some of the data science algorithms presented earlier and were based on some of the studies described in this section, such as Prékopa (2013), Biswal et al. (2005), Sahoo and Biswal (2005), Nemirovski and Shapiro (2007), Luedtke and Ahmed (2008), and Nikzad et al. (2019).

4
Methodology overview

In this thesis, a data-driven optimization methodology for the workover rig scheduling problem using regression-driven models is proposed and analyzed. Several variations of the regression-driven models are compared to improve the optimization. The final model is a data-driven joint chance-constrained (DD-JCC) optimization model, which is compared with other data-driven chance-constrained variations (integrated chance-constrained and budget-constrained) and a regression-based data-driven optimization model. All data-driven optimization models use regression models to estimate the unknown parameter based on historical data.

The proposed data-driven methodology is divided into three major phases: data treatment (where data science methods are used to clean and classify the data), prediction (where predictive models are used to estimate uncertain parameters of the optimization, and optimization (where optimization models are generated and solved). These three phases are illustrated in Figure 4.1:

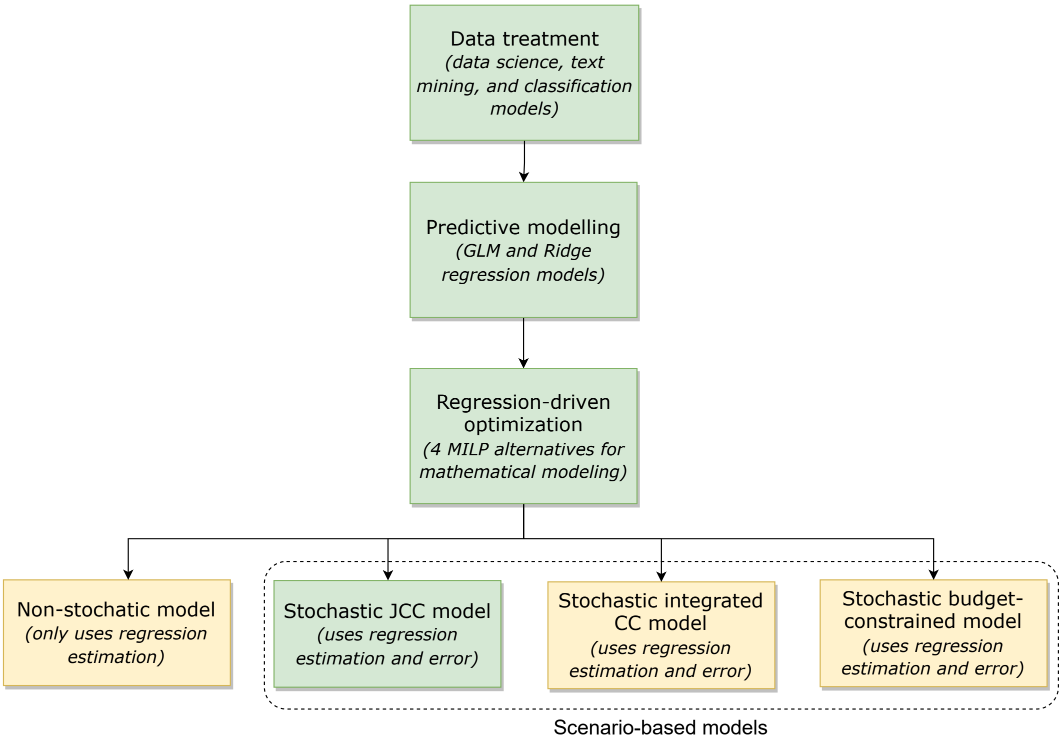

Data treatment
(data science, text mining, and classification models)
Predictive modelling
(GLM and Ridge regression models)
Regression-driven optimization
(4 MILP alternatives for mathematical modeling)
Non-stochatic model
(only uses regression estimation)
Stochastic JCC model
(uses regression estimation and error)
Stochastic integrated CC model
(uses regression estimation and error)
Stochastic budget-constrained model
(uses regression estimation and error)
Scenario-based models

Figure 4.1: Overview of the proposed data-driven optimization methodology.

The techniques involved in these three major phases were discussed in Chapter 3, which included a review of optimization under uncertainty, data-driven optimization approaches, data treatment, and predictive modeling. Next, the different phases and possible models of the proposed methodology are detailed.

4.1 Methodology description

At the beginning of this chapter, a data-driven optimization methodology was proposed (Figure 4.1). After reviewing the basic concepts and techniques involved in these and the state-of-art of the data-driven chance-constrained optimization (Chapter 3), we can now describe in more detail the phases of the proposed methodology, the models, and the software tools used to develop them.

As mentioned earlier, a data-driven methodology is proposed and four alternative optimization models are compared in this study. These methods use regression models to estimate the unknown parameter based on historical data. A regression-driven optimization model is proposed considering only the regression estimations. Other more sophisticated regression-driven models were created using not only the regression estimations but also their associated errors.

First, we develop a regression-driven JCC formulation, but it proves to be non-linear and extremely hard to solve, just as the kernel-based JCC from Calfa et al. (2015) and Jiang and Guan (2016), mentioned in Section 3.3.6. Aiming to obtain a practical JCC formulation, we propose a new linear programming JCC model combining regression models and stochastic programming (the scenario-based regression-driven JCC). As an alternative, we also create two other regression-driven chance-constrained variations: integrated chance-constrained and budget-constrained optimization. Figure 4.2 summarizes the linear variations of the proposed regression-driven methodology (green labels), a non-linear regression-driven model also developed in this thesis (yellow labels), and Calfa et al. (2015)'s non-linear formulation (red labels).

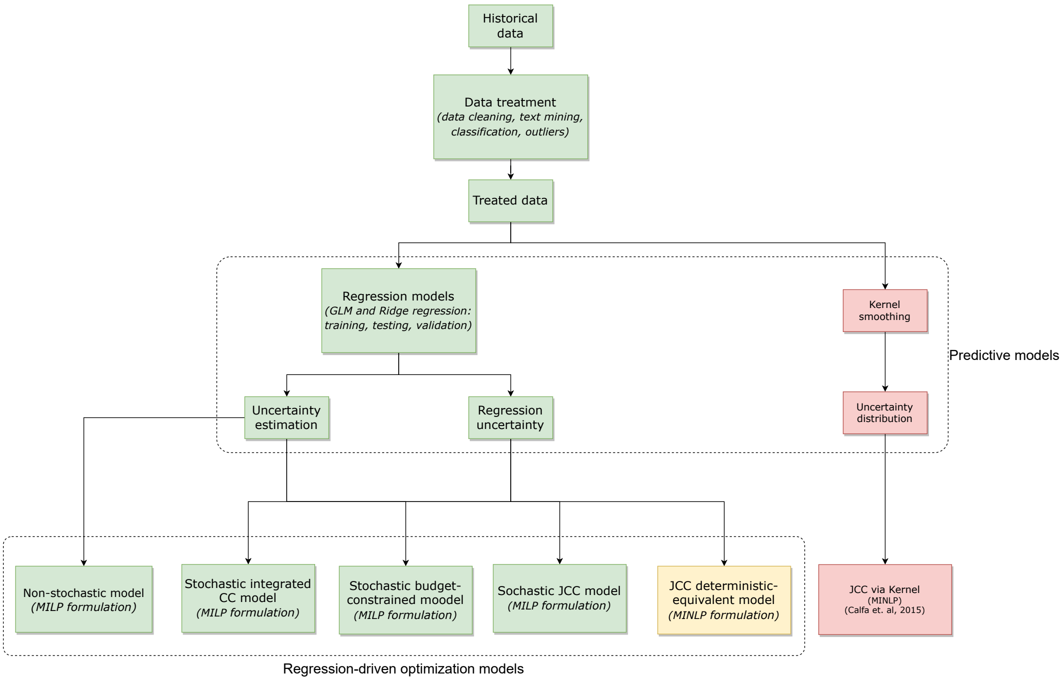


Figure 4.2: Proposed data-driven methodology (green labels are variations using linear models and yellow label is non-linear) and Calfa et al. (2015)'s non-linear formulation (red).

The first two stages of the methodology are related to the uncertainty treatment, which will be presented in Section 4.1.1. In the proposed regression-driven optimization methodology (green and orange labels), the data is first prepared and analyzed using text mining techniques and the k-means clustering algorithm for classification. This treated data is then used in regression models, such as generalized linear models and ridge regression, to estimate the uncertainty. Later, the regression estimators are used to represent the unknown parameter in the data-driven optimization and the regression errors are used in the JCC models as a new uncertainty. An alternative would be to apply Calfa et al. (2015)'s method (red labels), which instead of using parametric regression models to predict the uncertainty, uses kernel soothing models to estimate the uncertainty density distribution and later apply it in the optimization. However, as mentioned earlier in Section 3.3.6, this approach would eventually lead to a challenging non-linear programming model.

The last stage of the methodology is the optimization, which augments the model with the regression to consider uncertainty related to prediction errors using the proposed method based on chance constraints (detailed in Section 4.1.2). Two formulations with regression are proposed for the JCC optimization: one with non-linear programming (yellow label, Section 4.1.2) and another with stochastic linear programming (green labels, Section 4.1.2.2). Two other chance-constraints variations are also suggested as alternatives for the regression-driven stochastic models (green labels, Section 4.1.2.2). These stochastic linear programming approaches use stage decisions and scenarios sets to make the chance constraints linear.

The data analysis, treatment, and prediction of the workover rig schedule data are presented in Chapter 5. The regression-driven model is applied in the workover rig scheduling problem in Chapter 6. Meanwhile, the data-driven chance-constrained models are implemented in the workover rig scheduling problem in Chapter 7. Next, we detail the proposed methodology for analyzing the data and estimating uncertainty.

4.1.1 Data analysis methodology

Focusing on the data analysis underneath the data-driven optimization models, this section is dedicated to the data treatment and regression algorithms used to estimate the uncertainty, proposing a methodology and comparing it with others in the literature. The data analysis is separated into two parts: data treatment and predictive models.

The data treatment uses text mining and clustering methods to simplify and classify the qualitative data, which are used with the quantitative data as input of predictive models that predict the distribution of the duration of the residuals. The data treatment methodology is illustrated in Figure 4.3.

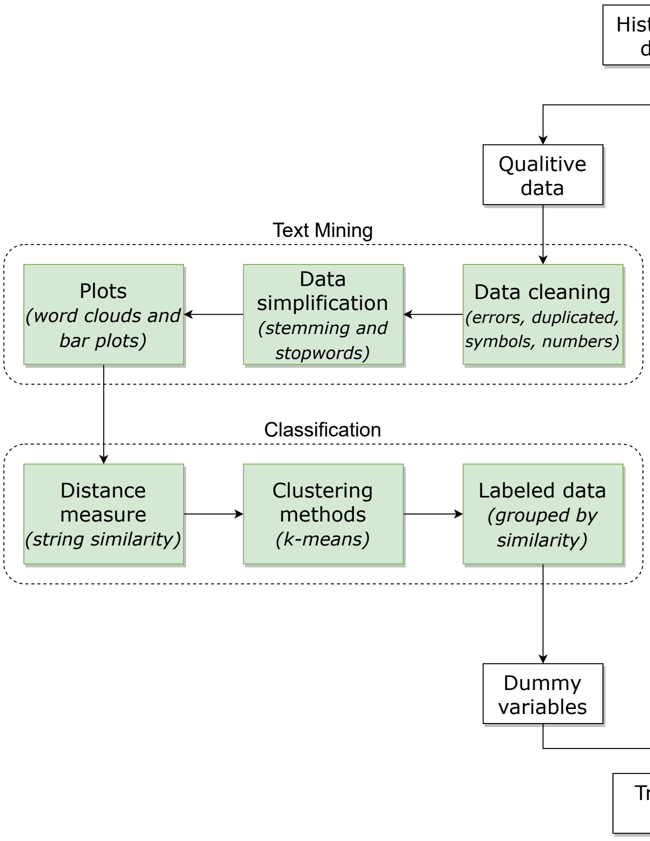

Historical data
Qualitive data
Quantitative data
Text Mining
Plots (word clouds and bar plots)
Data simplification (stemming and stopwords)
Data cleaning (errors, duplicated, symbols, numbers)
Data cleaning (errors, duplicated, empty fields)
Classification
Distance measure (string similarity)
Clustering methods (k-means)
Labeled data (grouped by similarity)
Plots (box plots and histograms)
Outliers elimination
Dummy variables
Numerical variables
Treated data

Figure 4.3: Framework with the data treatment methodology.

The text mining (stemming and string similarity) and the clustering algorithms (k-means) used were presented in Sections 3.1.1 and 3.1.2, respectively. These methods were applied based on the data science framework from Shcherbakov et al. (2014), which 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 box plots and histograms, outliers are eliminated, generating numerical variables for the predictive models. The qualitative data is treated with text mining techniques (responsible for cleaning the data) and classification 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", and "wordcloud" include: data cleaning, and data simplification using stopwords and stemming,

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

- *Distance measure:* uses string similarity and distance tools to measure how close the sentences of the qualitative data are to each other. After several preliminary tests, a custom string similarity measure was employed 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:

$$\text{String Similarity}\,(s1, s2) = \frac{LV\,(s1, s2) + LCS\,(s1, s2)}{2}, \tag{4-1}$$

  where $s1$ and $s2$ in Equation (4-1) refer to "String1" and "String2", respectively. As mentioned earlier, both similarity measures are efficient for short strings and the combination of the two resulted in suitable matches combining an edit-based similarity with a sequence-based measure (KDnuggets 2019).

- *Clustering methods:* applies the previous string similarity measure as a distance measure in a k-means algorithm (Likas et al. 2003b) to group textual data according to their similarities, obtaining richer labels for the categorical data.

Two types of predictive models are employed to estimate the uncertainty: linear regression models and ridge regression models. Both model types were introduced earlier in Section 3.2. The data prediction methodology is illustrated in Figure 4.4.

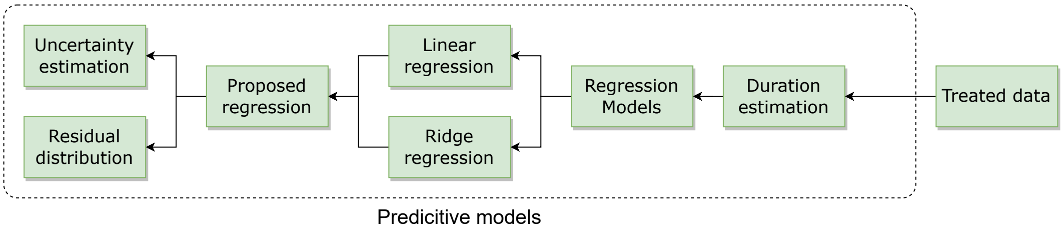


Figure 4.4: Framework with the predictive modeling phase of the proposed methodology.

However, the linear regression models (GLM) might overfit on the training data if there is a large number of dummy variables, and, as a result, a large number of coefficients. Therefore, the model can overfit and might not perform appropriately on an out-of-sample data set. As this study proposes to use qualitative data as an input to predicting the uncertainty, a large number of dummy independent variables might be generated. Therefore, the ridge model has been chosen as an alternative testing method.

Using the previous libraries for GLMs and ridge regression, a procedure was created to exhaustively test all possible combinations of response variables to predict each of the regressions mentioned above. Based on the hold-out validation, the procedure separates the 80% of the data in the in-sample set and the 20% left in the out-of-sample set. 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 by predicting 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, a low error, and possibly low complexity and residuals normally distributed. Last, the selected model is used to predict the uncertainty parameter affecting the optimization.

The combination of the regression model and the optimization variables will be presented in Sections 4.1.2.1 and 4.1.2.2.

Several packages are available in the R programming language for estimating generalized linear models (GLMs). In this study, we used the native library *Stats* (R Core Team 2013) and the package *olsrr* (Hebbali and Hebbali 2017) for the GLMs. These packages allow estimating the coefficients of the model that minimize the loss function. The ridge regression models were estimated using the *glmnet* (Engebretsen and Bohlin 2019) and the native library *Stats* (R Core Team 2013) libraries for the R programming language.

4.1.2 Regression-driven optimization: methodologies and formulations

As mentioned earlier, the optimization phase uses regression estimators for the uncertain parameter. A possible regression-driven formulation would be to consider these estimators as fixed and not subjected to any error associated with the regression method. However, we do know that the regression is highly dependent on the data quality, the features selected, and its training. As a result, the regression has an error associated with it. To obtain more robust solutions that consider the regression uncertainty and have some level of flexibility in the solution to absorb these deviations, we propose formulations with regression-driven chance constraints. These alternative formulations are presented in Figure 4.5:

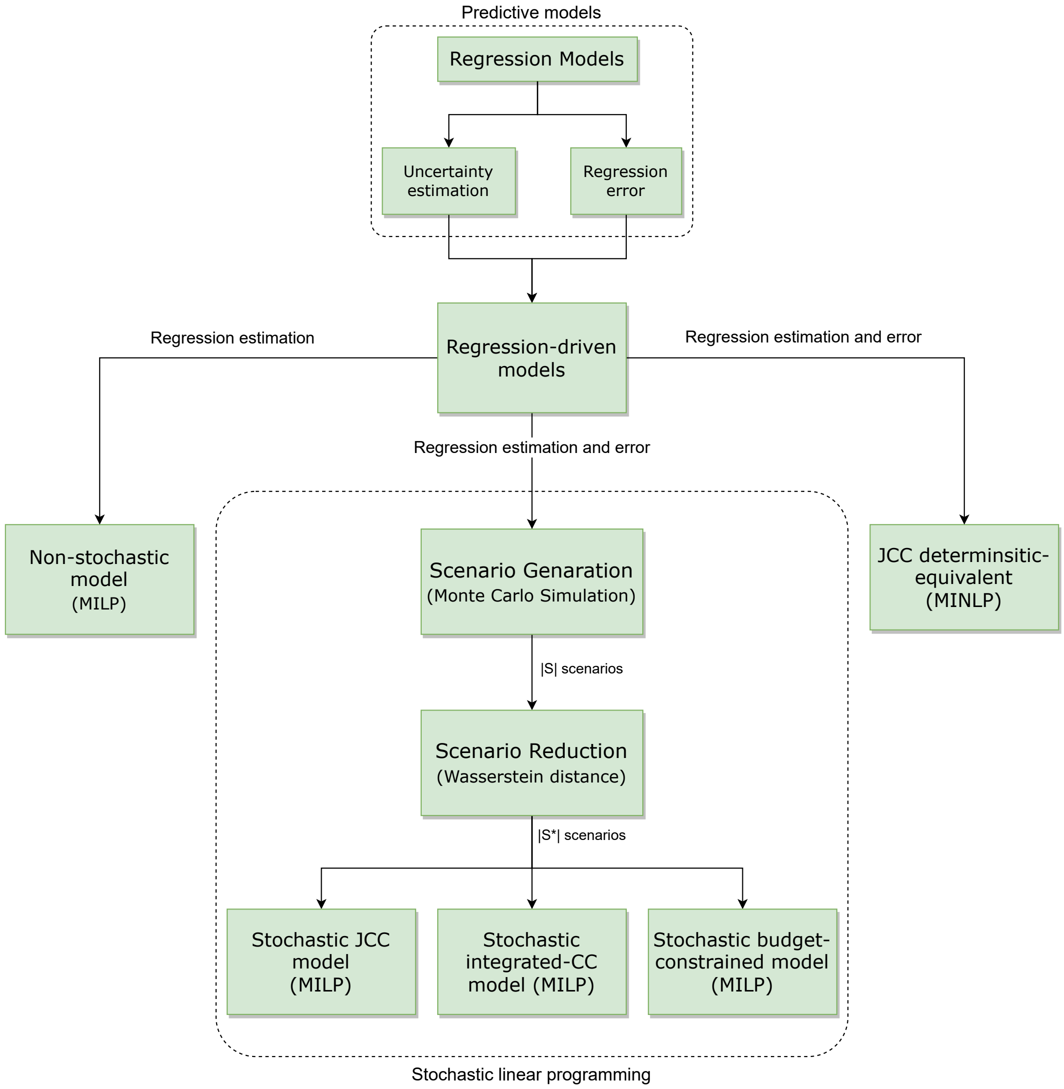

Predictive models
Regression Models
Uncertainty estimation
Regression error
Regression estimation
Regression-driven models
Regression estimation and error
Regression estimation and error
Non-stochastic model (MILP)
Scenario Genaration (Monte Carlo Simulation)
JCC determinsitic-equivalent (MINLP)
|S| scenarios
Scenario Reduction (Wasserstein distance)
|S*| scenarios
Stochastic JCC model (MILP)
Stochastic integrated-CC model (MILP)
Stochastic budget-constrained model (MILP)
Stochastic linear programming

Figure 4.5: Methodology for regression-driven optimization.

The regression-driven optimization methodology shown in Figure 4.5 use text mining, classification algorithms, and regression models to estimate the uncertain parameter (as proposed in Section 4.1.1). Equation (4-2) presents an example of a regression output for a parameter ξ_i:

$$\xi_i \sim \tilde{\xi}_i = \hat{\xi}_i + \varepsilon, \tag{4-2}$$

where ξ_i is the actual value of the parameter, $\tilde{\xi}_i$ is its approximation, $\hat{\xi}_i$ is its prediction from the regression model, and the distribution of ε can be estimated using the residuals from the regression.

The regression-driven non-stochastic model assumes that $\tilde{\xi}_i \simeq \hat{\xi}_i$. Meanwhile, the other more sophisticated methods use the regression error distribution as an uncertainty parameter in its optimization. This thesis opts for the chance-constrained approach. As mentioned earlier in Section 3.3, the chance-constrained optimization allows a certain level of probability of infeasibility, which enables us to find less conservative solutions than classic robust optimization and can be more accurate to the reality of fleet sizing and scheduling problem.

However, chance-constrained optimization usually leads to non-linear formulations. This was observed in the regression-driven deterministic-equivalent JCC formulation (Section 4.1.2.1 details this formulation we developed). To tackle that, a scenario-based modeling approach and a stochastic JCC formulation are proposed (which is described in Section 4.1.2.2). Two other chance-constraints variations based on stochastic programming are also proposed as alternative formulations for the problem (integrated-CC and budget-constrained) and are presented in Section 4.1.2.2. Next, we detail these different regression-driven chance-constrained optimization models proposed in this thesis.

4.1.2.1 Regression-driven JCC: general-form and deterministic-equivalent

This section proposes a regression-driven JCC formulation based on the ridge regression output from Section 4.1.1 assuming that $\tilde{\xi}_i = \hat{\xi}_i + \varepsilon$ (4-2), where $\hat{\xi}_i$ are the ridge regression estimations and ε the residuals associated with it.

First, we apply this equation to the probabilistic joint constraints, obtaining the general-form for the proposed regression-driven JCC:

$$\mathbb{P}\left[g_{ij}(x) \leq \hat{\xi}_i + \varepsilon, \forall j\right] \geq \alpha \qquad \forall i, \tag{4-3}$$

where $\varepsilon \sim N(0, \sigma)$, the residuals error follows a Normal distribution, and $\hat{\xi}_i$ are the regression estimations of the uncertain parameter.

As mentioned in the literature review, Biswal et al. (2005) and Sahoo and Biswal (2005) used the properties of normal and log-normal distributions to reformulate the joint constraint programming when the uncertainty in the RHS follows normal and log-normal distributions (as shown in Equations (4-4)-(4-9)). Based on their reformulations, we can modify the regression-driven JCC (4-3) for the case where the residuals are independent random variables and follow a Normal distribution $N(\mu, \sigma)$.

$$\mathbb{P}\left[g_{ij}(x) - \hat{\xi}_i \leq \varepsilon, \forall j\right] \geq \alpha, \qquad \forall i \tag{4-4}$$

$$= \prod_j \mathbb{P}\left[g_{ij}(x) - \hat{\xi}_i \leq \varepsilon\right] \geq \alpha, \qquad \forall i \tag{4-5}$$

$$= \prod_j \mathbb{P}\left[\varepsilon \geq g_{ij}(x) - \hat{\xi}_i\right] \geq \alpha, \qquad \forall i \tag{4-6}$$

$$= \prod_j \mathbb{P}\left[\varepsilon \geq g^{'}_{ij}(x)\right] \geq \alpha, \qquad \forall i \tag{4-7}$$

$$= \prod_j \mathbb{P}\left[\frac{\varepsilon - \mu}{\sigma} \geq (\frac{g^{'}_{ij}(x) - \mu}{\sigma}\right] \geq \alpha, \qquad \forall i \tag{4-8}$$

$$= \prod_j \left[1 - \Phi\left(\frac{g^{'}_{ij}(x) - \mu}{\sigma}\right)\right] \geq \alpha, \qquad \forall i, \tag{4-9}$$

where $\varepsilon \sim N(\mu, \sigma)$, $g^{'}_{ij} = g_{ij}(x) - \hat{\xi}_i$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard Normal distribution, which is equal to $\Phi(x) = \mathbb{P}(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\inf}^{\inf} e^{-\frac{u^s}{2}} du$.

Additionally, if we consider $\mu = 0$, then (4-9) can be simplified into:

$$\prod_j \left[1 - \Phi\left(\frac{g_{ij}(x) - \hat{\xi}_i}{\sigma}\right)\right] \geq \alpha, \qquad \forall i, \tag{4-10}$$

Nonetheless, this proposed regression-driven joint chance-constrained deterministic equivalent reformulation obtained using the JCC reformulation from Biswal et al. (2005) and Sahoo and Biswal (2005) is still an MINLP model. As mentioned earlier, MINLP models are complex and computationally hard to solve. A possible solution for this is to use the regression-based formulation of the JCC and adapt it to a stochastic formulation, which will be presented in the next section.

4.1.2.2
Regression-driven stochastic chance-constrained optimization

The regression-based joint chance-constrained models presented in the last section resulted in the non-linear constraints, (4-9) and (4-10). An alternative solution that has linear constraints is to represent the uncertainty with scenarios. In this section, we propose a scenario-based reformulation for the regression-driven JCC model, named regression-driven stochastic joint chance-constrained model.

As mentioned in Figure 4.5, the stochastic approaches of this study are basically divided into three procedures:

- *Scenario Generation:* Using the distribution estimated within the regression residuals ($\varepsilon \sim N(\mu, \sigma)$). A large and sufficient number of scenarios is generated using Monte Carlo Simulation (several runs of simulations are made until the average mean and standard deviation converges to respectively distribution's mean and standard deviation).

- *Scenario Reduction:* A algorithm using the Wasserstein distance (explained in Section 3.3.3) is used to reduce the number of scenarios. Two measures are considered while evaluating the best number of scenarios to be used in the optimization: the error of approximating and the expected feasibility ratio. The error of approximating was explained in Section 3.3.3 and was proposed by Kaut and Stein (2003), Pflug (2001), and Luedtke and Ahmed (2008). The expected feasibility ratio was calculated by simulating the solution and the JCC for the entire set of scenarios and measuring the proportion of feasible solutions (Hong et al. 2015).
- *Stochastic linear programming:* After generating and reducing the scenarios of the regression residuals, the regression estimations and the optimally reduced scenarios of the regression uncertainty are provided for the two-stage stochastic linear programming model with the chance constraints. The three mathematical formulation alternatives for the regression-driven stochastic CC optimization proposed are a **JCC-approach**, an **integrated-CC**, and a **budget-constrained**. They will be detailed further in this section.

As mentioned, the scenario generation uses the distribution of the residuals obtained through the ridge regression commented in Section 4.1.1 and applies Monte Carlo Simulation to generate a large number of scenarios. In stochastic programming, the scenarios generated must represent as truthfully as possible the uncertainty, which usually is made considering a large number of scenarios. However, in most large-scale and real-life problems, it is only computationally feasible to optimize with a small number of scenarios. Several scenario generation and reduction methods exist to reduce the number of scenarios, but still keep them a reliable representation of the uncertainty (Dupačová et al. 2003, Heitsch and Römisch 2003). Problems with probabilistic constraints already require a significant computational effort; when combined with stochastic programming, the scenario sets ought to be as small as possible without losing their accuracy (Nikzad et al. 2019).

Therefore, after generating multiple scenarios with the Monte Carlo simulation, a Wasserstein-distance-based scenario reduction method is used to reduce the number of scenarios as much as possible while retaining the stochastic information of the data. To select the number of scenarios, the measures error of approximating (Kaut and Stein 2003, Pflug 2001, Luedtke and Ahmed 2008) and expected feasibility ratio (Hong et al. 2015) were analyzed.

Last, the scenarios are included in a two-stage stochastic MILP model, in which the probabilistic constraints were reformulated as linear constraints and the uncertainty parameter is estimated via regression. The scenario-based model is made by applying the stochastic reformulations of Nemirovski and Shapiro (2007) and Nikzad et al. (2019) on the regression-driven JCC model from Section 4.1.2.1. Its general representation is as follows:

Sets:

- $i \in \{1, 2, \ldots, N\}$: Variables dimensions.
- $j \in \{1, 2, \ldots, M\}$: Constraints dimensions.
- $\omega \in \{1, 2, \ldots, \Omega\}$: Scenarios for the uncertainty. Each scenario represents a realization of the uncertainty.

Parameters:

- c_i^1: First-stage costs.
- c_i^2 : Second-stage costs.
- φ_i : Penalization for infeasible solutions.
- ξ_i^ω: Regression estimations for the uncertainty parameter.
- ε_i^ω: Simulations of the regression error ε.
- π^ω : Probability of scenario ω's occurrence.
- α : Risk level of the probabilistic constraints.

Variables:

- X_{ij}: First-stage decision variables.
- Y_{ij}^ω: Second-stage decision variables.
- V_{ij}^ω: Slack variable between [0,1] that measures the relaxation of constraint (ij) feasibility in scenario ω (second-stage decision).
- Z_i^ω: Binary variables for not violating the joint chance-constraints of i in scenario ω (second-stage decision).

A stochastic model is usually divided into multi-stages decisions. In this study, we consider first-stage and second-stage decisions, both presented in the objective function (Equation (4-11)). The first-stage decisions (X_{ij}) are associated with the decisions made before the occurrence of the uncertainty, *i.e.*, before the value of the uncertain parameter $\tilde{\xi}_i$ is known. Alternatively, the second-stage costs are associated with the decisions made after the realization of the uncertainty (Y_{ij}^ω). Each scenario has a probability π^ω. We use binary variables that indicate when the JCCs are violated, as in Nemirovski and Shapiro (2007) and Nikzad et al. (2019). The slack variables between [0,1] was an adaption that we have made to allow some flexibility in the amount of feasibility tolerated and guarantee that the infeasibility was only associated with the regression error. The objective function and its constraints are presented as follows:

$$\text{Min } \sum_{i\in N}\sum_{j\in M} c_i^1 X_{ij} + \sum_{\omega\in\Omega}\pi^\omega \sum_{i\in J}\sum_{j\in M} c_i^2 Y_{ij}^\omega \tag{4-11}$$

$$\text{Subject to}$$

$$g_{ij}\left(X_{ij} + Y_{ij}^\omega\right) \leq \hat{\xi}_i + \varepsilon_i^\omega(1 - V_{ij}^\omega) \qquad \forall i, j, \omega \tag{4-12}$$

$$V_{ij}^\omega \geq Z_i^\omega \qquad \forall i, j, \omega \tag{4-13}$$

$$1 - V_{ij}^\omega \leq Z_i^\omega \qquad \forall i, j, \omega \tag{4-14}$$

$$\sum_{\omega\in\Omega} \pi^\omega Z_i^\omega \geq \alpha \qquad \forall i \tag{4-15}$$

$$X_{ij} \geq 0 \qquad \forall i, j \tag{4-16}$$

$$V_{ij}^\omega \in [1, 0] \qquad \forall i, j, \omega \tag{4-17}$$

$$Z_i^\omega \in \{1, 0\} \qquad \forall i, \omega \tag{4-18}$$

$$Y_{ij}^\omega \geq 0 \qquad \forall i, j, \omega. \tag{4-19}$$

The joint chance constraints are defined in constraints (4-12), (4-13), and (4-15). Constraint (4-12) has a slack variable V_{ij}^ω that allows some violations within a range (not greater than the value of ε_i^ω). If the slack variable is used for any (i, j) in constraint (4-12) (variable V_{ij}^ω lower than one), then constraints (4-13) and EQ:413b indicate that the joint constraint of i is infeasible in scenario ω, *i.e.*, Z_i^ω equals zero. Constraint (4-15) guarantees that the joint probabilistic constraint respects the confidence level α of being feasible for all i. The other constraints (4-16), (4-17), (4-18), and (4-19) define the variables' domains.

Two other regression-driven stochastic CC formulations are also proposed to be compared with the JCC approach. The first alternative was using regression-driven integrated chance constraints (integrated-CC). As mentioned earlier in Section 3.3.1, the integrated chance constraint is a measure of violation introduced by Haneveld (1986) in the expected magnitude of violation instead of its probability. Modifying the general integrated-CC formulation from Prékopa (2003) to the regression-driven stochastic approach and changing the variable Z_i^ω to measure the maximum violation would lead to the following constraints:

$$g_{ij}\left(X_{ij} + Y_{ij}^\omega\right) \leq \hat{\xi}_i + \varepsilon_i^\omega V_{ij}^\omega \qquad \forall i, j, \omega \tag{4-20}$$

$$\varepsilon_i^\omega(1 - V_{ij}^\omega) \leq Z_i^\omega \qquad \forall i, j, \omega \tag{4-21}$$

$$\sum_{\omega\in\Omega} \pi^\omega Z_i^\omega \leq \mu_\varepsilon \alpha + \frac{K\sigma_\varepsilon}{\sqrt{(|\Omega|)}} \qquad \forall i \tag{4-22}$$

$$X_{ij} \geq 0 \qquad \forall i, j \tag{4-23}$$

$$V_{ij}^\omega \in [1, 0] \qquad \forall i, j, \omega \tag{4-24}$$

$$Z_i^\omega \geq 0 \qquad \forall i, \omega \tag{4-25}$$

$$Y_{ij}^\omega \geq 0 \qquad \forall i, j, \omega. \tag{4-26}$$

The regression-driven integrated chance-constraints are represented by constraints (4-20), (4-21), and (4-22). Just as in the JCC, constraint (4-20) is the chance constraints with the slack variable V_{ij}^ω that allows some violations within a range (not greater than the value of ε_i^ω). The worst violation of the chance-constraint each scenario, *i.e.*, the largest product of ε_i^ω and V_{ij}^ω is calculated as Z_i^ω in constraint (4-21). The expected worst violation is calculated in constraint (4-22), which also guarantees that its values respect the upper bound considering K from the percentile point associated with an α confidence level of the standard Normal distribution. Last, constraints (4-23), (4-24), (4-25), and (4-26) are related to the variables' domains. Note that the variable Z_i^ω is continuous and non-negative in the regression-driven stochastic integrated-CC formulation.

The other scenario-based formulation is the budget-constrained formulation. This formulation focus on limiting the total amount of feasibility allowed in each scenario, regardless of their probability or expected value. This approach is subjected to the following constraints:

$$g_{ij}\left(X_{ij} + Y_{ij}^\omega\right) \leq \hat{\xi}_i + \varepsilon_i^\omega V_{ij}^\omega \qquad \forall i, j, \omega \tag{4-27}$$

$$\sum_{i\in N}\sum_{j\in M} \varepsilon_i^\omega(1 - V_{ij}^\omega) \leq M\alpha \qquad \forall \omega \tag{4-28}$$

$$X_{ij} \geq 0 \qquad \forall i, j \tag{4-29}$$

$$V_{ij}^\omega \in [1, 0] \qquad \forall i, j, \omega \tag{4-30}$$

$$Y_{ij}^\omega \geq 0 \qquad \forall i, j, \omega \tag{4-31}$$

Next, we present a case study regarding real-life-based instances of the workover rig scheduling problem and perform a data analysis focusing on the two first stages of the regression-driven methodologies, data treatment and predictive modeling (Chapter 5). Then, using the regression estimations, we implement the regression-driven optimization model in the WRSP and test it in several instances, analyzing its sensibility, simulating the solutions, and comparing it with the current company methodology (Chapter 6). After implementing the first regression-driven model in the WRSP, we implement and test the scenario-based methods, such as the regression-driven stochastic JCC, integrated-CC, and budget-constrained optimization models, simulating their first-stage solutions and comparing them by different metrics. Finally, these data-driven joint chance-constrained optimization methodologies are applied in the workover rig scheduling (Chapter 7).

5
The workover rig scheduling problem: assumptions and uncertainty analysis

As mentioned in the systematic literature review, the workover rig scheduling problem has received significant attention in the literature. However, very few problems in the SLR have been implemented or verified within the oil and gas industry. One of the main reasons for this gap between academia and industry is the lack of studies adapted to industry demands. The SLR also enabled us to detect trends and opportunities of approaches and methods. One of those was the data-driven optimization under uncertainty.

Aiming to address this problem, this thesis approaches a real-life case of the offshore workover rig scheduling problem in a major Brazilian oil company, proposing data-driven methodologies. In the previous chapter, the concepts and the state-of-art of data-driven optimization and optimization under uncertainty were reviewed. Remarkably, the data-driven chance-constrained optimization appeared to be the most suitable approach for this Thesis. Several regression-driven optimization methodologies were also proposed in Chapter 4. Before implementing these methodologies in the WRSP, the current section presents the case study, its instances, and its data sets.

First, we present the case study and its assumptions (Section 5.1). Then, we describe the instances and the datasets used for the mathematical models (Section 5.2). Last, the workover data is treated and predictive models are used to estimate the uncertainty affecting the problem (Section 5.3).

5.1
Case study

As mentioned earlier, the workover rig scheduling problem (WRSP) is a particular case of the rig scheduling problem for workover operations from a stand-alone planning perspective. In this study, we approach the offshore WRSP of Petrobras, a Brazilian oil company that operates the majority of the oil fields and needs to plan a large fleet of rigs on its offshore wells. As a result, this case study has some particularities. A large set of wells requires workover operations, and a fleet of rigs needs to be hired to serve them. The goal is to decide which wells will be served by which rig in the planning horizon, minimizing the costs associated with hiring the rigs and the oil production loss of the wells waiting for workover service.

The workover planning is performed separately from the others operations at a stand-alone planning level and a dedicated fleet of heterogeneous rigs is hired to execute them and no other operation. 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 fixed a cost when hired. Others resources besides rigs are not considered in this case study.

Each well has an oil production associated with it, regardless if it is an injector or producer well. Further details on the oil production of the wells are provided in the instance generation (Section 5.2). Every well requires only one maintenance or rework operation, existing only a single job for a well, and has a release date related to the date it starts needing the workover. There is a cost associated with the oil production loss of the wells waiting to be served, which extends until the end of the planning horizon if the well is not served. As in this case study, each well requires only one workover operation or task, *i.e.*, a single job scheduling problem, for which we use the terms *well*, *workover*, *operation*, *task*, and *job* interchangeably.

In this setting, the oilfields are located offshore. The 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. Wells only need to be scheduled, and the routing problem is negligible.

Last, the processing time for each workover operation varies for each class of rig. In the deterministic case, these processing times are known and constant for each rig and well. However, in the real-life process, there is much uncertainty associated with the workover operation. Most workover operations are unpredictable, being extremely complex to determine which wells will demand workover, when, and the duration of the workover. Figure 5.1 classifies the studied WRSP, labels in blue, according to the taxonomy proposed in Section 2.2.

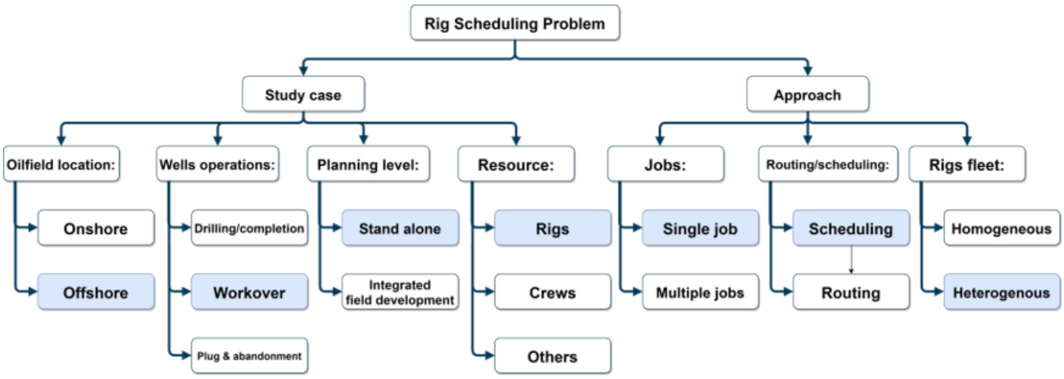


Figure 5.1: The studied workover rig scheduling problem according to the proposed taxonomy.

5.2 Instances and datasets

To test the proposed data-driven optimization methodology on the workover rig scheduling problem, data from a major Brazilian oil company was gathered and structured as follows in Figure 5.2.

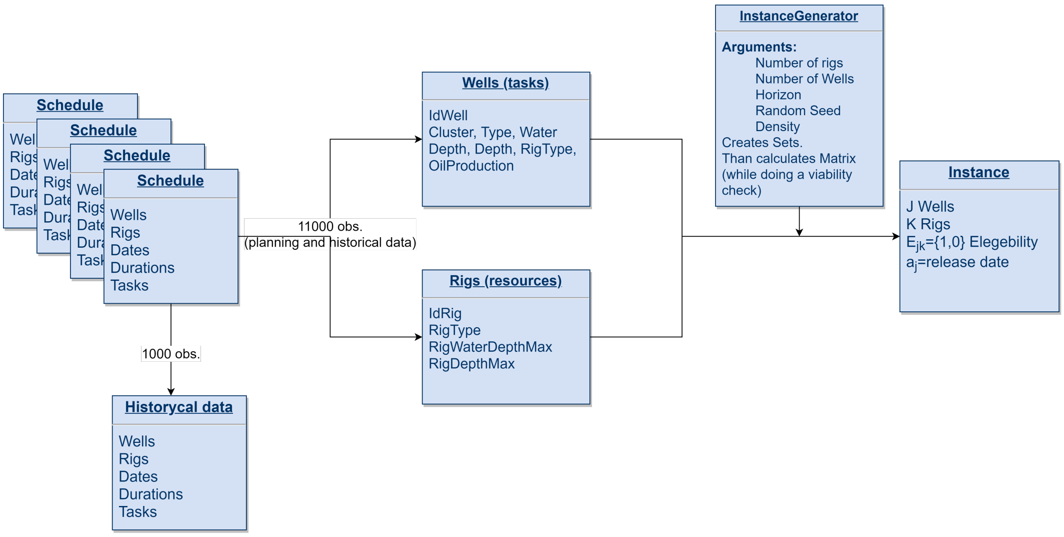


Figure 5.2: Data structure and instance generation frameworks.

Several historical workover rig schedules (1.000 records) were gathered with workover information regarding wells, rigs, dates, durations, and tasks. Information about the wells and rigs was extracted from the collected data, generating two large data sets:

- **Wells/tasks data**: The cluster in which the workover is grouped (obtained through the text mining and classification methods that were described in Section 4.1.1), the type of well (basically if it is injector, producer, or something else), the type of operation (workover, drilling, abandonment, etc.), the well's water depth, the well's depth, rig type that can serve the well (if exists a particular one), and the oil production of the well (further details on its calculation will be provided later in this section).

- **Rigs data**: The type of rig, the rig's maximum water depth, the rig's maximum depth, the rig's operation cost, and the rig's hiring costs. Details on calculating the rig's fees will be provided later in this section.

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 Equations (5-1) and (5-2):

$$p_i = Scale_i^{Operation} \cdot Scale_i^{Well} \cdot p_i^0 \tag{5-1}$$

$$p_i^0 \sim 10^3 \cdot \Gamma\left(\alpha = 2.3, \beta = 4.2\right), \tag{5-2}$$

where p_i is the oil production loss of the well, $Scale_i^{Operation}$ and $Scale_i^{Well}$ are parameters that make the oil production loss proportional with the operation type and well type (respectively), and p_i^0 are the random oil production generated using the Gamma distribution $\Gamma\left(\alpha = 2.3, \beta = 4.2\right)$, in which α and β are the shape and scale of the distribution (respectively). Fernández Pérez et al. (2018) suggested using the oil barrel price 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 5.1:

$Scale_i^{Operation}$		$Scale_i^{Well}$	
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

Table 5.1: Proportional scales of oil production according to the type of well and operation.

The rig hiring and operation costs were randomly selected obtained from Markit (2021) database, which has historical information on the rig average day rates according to the type of rig and 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, planning horizon, random seed, and density coefficient (represented by ρ). A 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 (ρ), it is a setting parameter between 0 and 1 that controls the release dates of the workovers. A small ρ tends to result in later 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 workovers. The algorithm selects random samples of the set to generate the 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 respect its maximum water depth and depth and its type. While constructing the eligibility matrix, the algorithm checks the feasibility of the instance, *i.e.*, if there is a rig for every well and if all rigs have a well to serve. In case of infeasibility, new samples are calculated until a feasible instance is found, outputting this instance to the data-driven models.

5.3 Workover uncertainty analysis

Considering that this study focuses on the short-term workover planning, where the demands are more predictable and known. However, their durations are still uncertain and subject to variations, being extremely challenging to predict the duration of the operations. Several different types of workovers exist and the processing times are affected by a handful of factors, such as well and rig properties. As a result, the duration can vary from days to more than a year, as shown in Figure 5.3.

Aiming to predict this disturbance in the workover duration and avoid infeasible schedules, we present in this section several data treatment and regression models to estimate the uncertainty. All techniques presented in this Section were implemented using the *R* programming language and the *RStudio* software.

5.3.1 Workover data treatment

In this study, the scheduler has a limited amount of available information when planning the workovers. Table 5.2 summarizes what is available according to the data group (well or rig attributes) and type (qualitative or quantitative data):

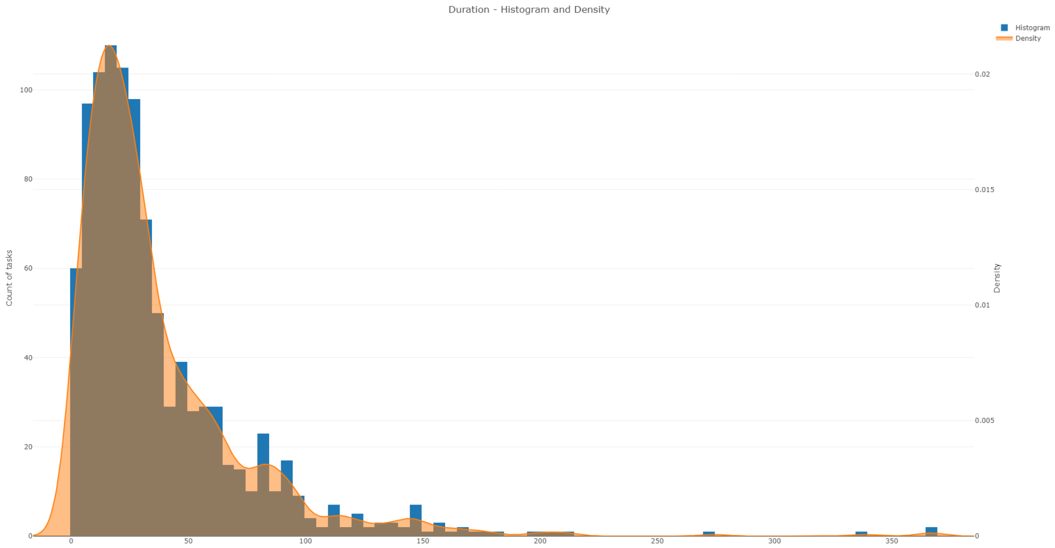


Figure 5.3: Histogram and density plot for the workovers durations.

Data	Group	Type	Description
Workover group	Well	Qualitative	Groups the workover in workover, light workover, and heavy workover.
Workover type	Well	Qualitative	Specifies the type of workover made.
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 in which the well is part of.
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.

Table 5.2: Description of the historical data gathered

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 problems 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. Figure 5.4 presents boxplots of the workover duration according to the type of workover.

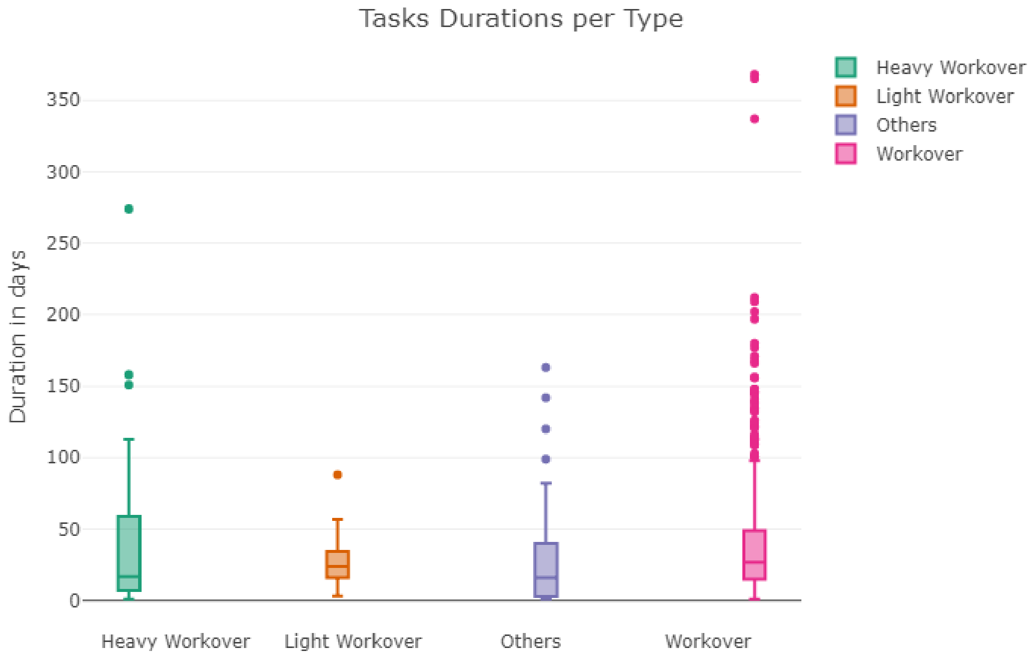


Figure 5.4: Box plots for the workover durations per each type of task.

Despite a large number of outliers in Figure 4.3, these observations are highly concentrated, especially the workover type (pink boxplot). This excess of outliers indicates that the groups are gathering operations with different duration behaviors.

Aiming to enhance the task grouping, the methodology presented in Figure 5.4 was used to obtain better clusters of tasks and to improve the qualitative data of 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. Figure 5.5 shows the most common fields in the original data of the task description.

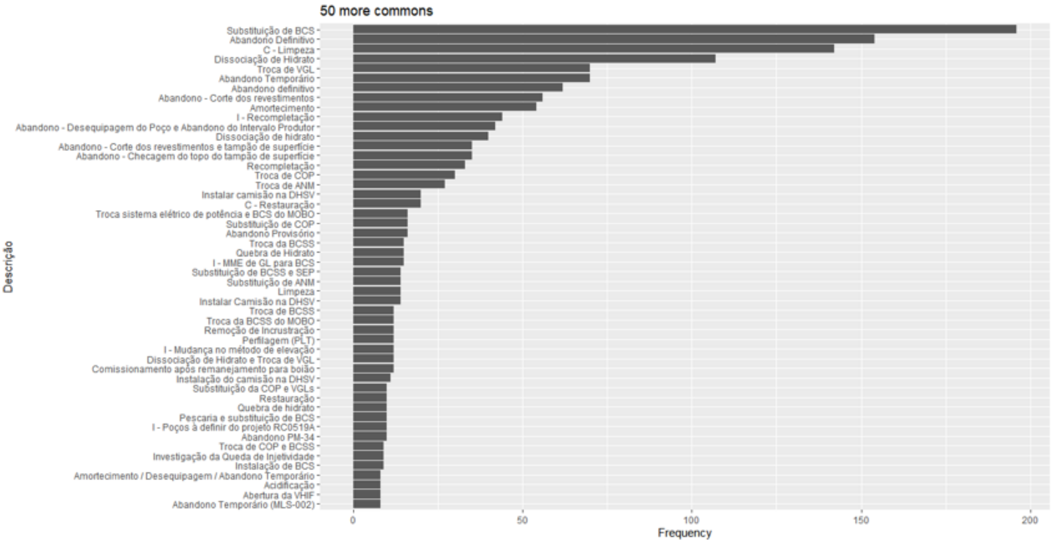


Figure 5.5: Bar-plot of the most common fields in the original task description.

After cleaning the text, the fields are simplified using the stemming technique (adapted for the Portuguese language). This technique reduces inflected or derived words to their respective stem-word, simplifying the text and making it easier to identify fields with the same meaning. Figure 5.6 shows the data simplified after this procedure:

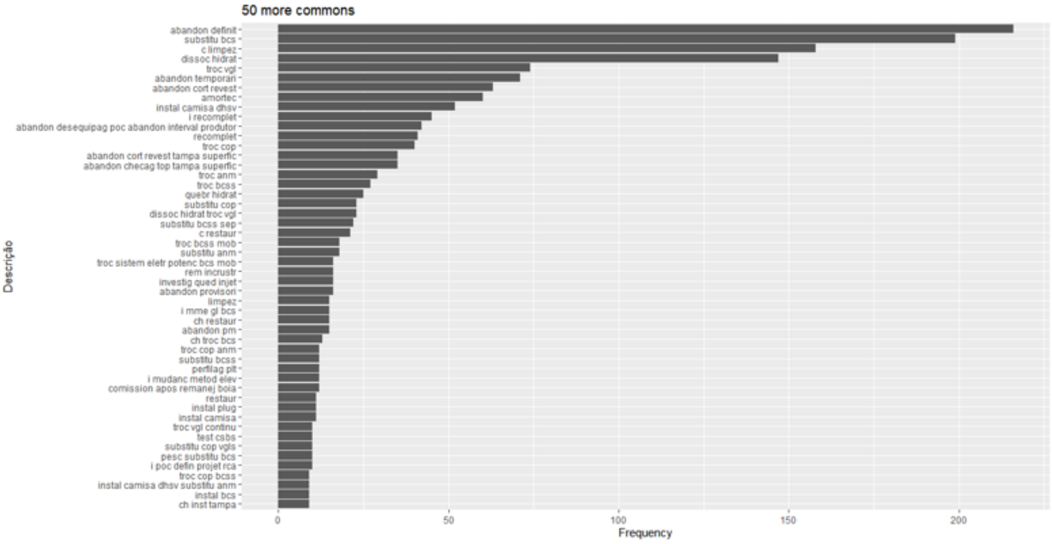


Figure 5.6: Bar-plot of the most common fields in the treated task description.

Comparing Figures 5.5 and 5.6, it is possible to observe that the data-cleaning process has successfully simplified the terms and made the count of most common descriptions more accurate. For instance, "abandono definitivo" was the second most common, but others were very similar in meaning and were not being counted with it. After cleaning and simplifying the data, the text "abandon definit" became the most common task description.

Word cloud plots were made to check if there is any pattern in the data. Figure 5.7 contains two word-clouds plots, (a) for one word alone (1-gram) and (b) for two words together (2-gram). 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, there are many sentences that still have similar meanings and, technically, could be considered as the same sentence. For instance, "substitu bcs" (replacement of ESP) and "bcs substitu" (ESP replacement) share the same meaning. This also happens with "abandon definit" (abandonment definitive) and "definit abandon" (definitive abandonment) and other sentences.

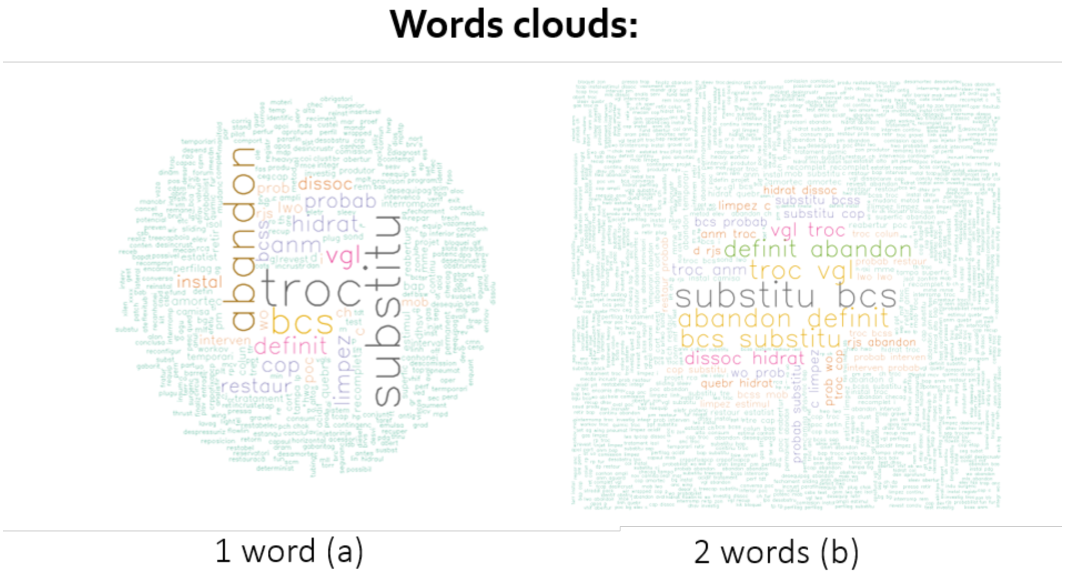


Figure 5.7: Word clouds for one word (a) and two words (b) using the simplified task description.

String similarity and distance tools can be used to measure how close these sentences are, and combined with clustering methods, such as h-cluster and k-means, a classification method for the task description can be created. The string similarity measure in Equation (4-1) was used as the distance measure of a k-means algorithm (Likas et al. 2003b), aiming to group the textual descriptions according to their similarities. Figure 5.8 shows a plot with the silhouette analysis that was used to determine the cluster size.

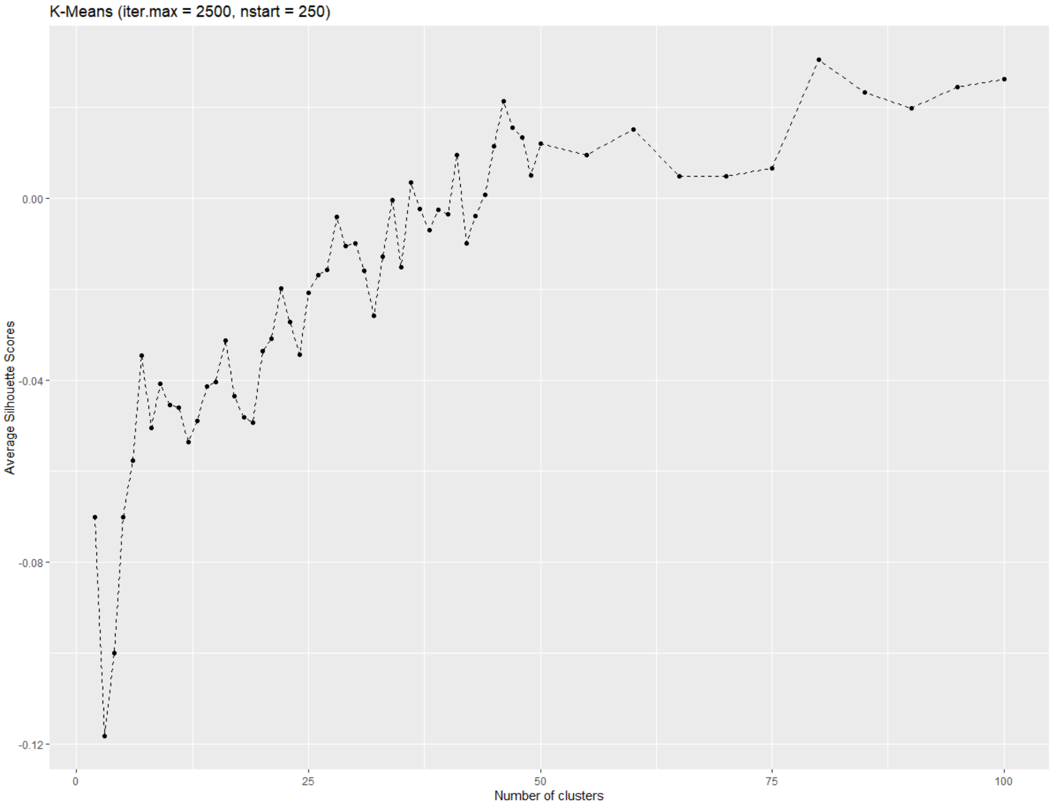


Figure 5.8: Plot with the average silhouette scores for different numbers of clusters (k) using k-means.

Our ultimate goal is to achieve extract information from the data through new labels and regression models, it is important to choose a number of clusters that it is not excessively big, disabling the training and testing of the regression model in the data. As a result, the number of clusters does not need to have an optimal silhouette, just enough to obtain better labels than the current ones.

Therefore, two strategies are possible for clustering and classifying the workover tasks. The first is to classify into major groups of tasks (a low number of clusters). The second would be to select smaller groups but not too small (a medium or a large number of clusters). Analyzing Figure 5.8, there are two particular peaks of the average silhouette score that fit these strategies: $k = 7$ (a smaller number of clusters) and $k = 45$ (a bigger number of clusters). Table 5.3 summarizes some of the final clusters for k (clusters) equal to 45

Cluster	Original data (in Portuguese)	Number of obs.
A	*Substituição de BCS*	100
A	*Substituição de ANM*	3
A	*Substituição de BCSS + SEP*	2
A	*Substituição da COP e VGLs*	2
A	*Substituição de BCSS e SEP*	2

Table 5.3: Examples of clusters using k-means with k=45.

Cluster	Original data (in Portuguese)	Number of obs.
A	*Substituição da BCSS*	2
B	*Abandono Definitivo*	32
B	*Abandono Temporário*	14
B	*Abandono - Corte dos revestimentos*	9
B	*Abandono - Checagem do topo do tampão de superfície*	5
B	*Abandono - Corte dos revestimentos e tampão de superfície*	5
B	*Abandono Permanente*	2
B	*Abandono Definitivo (Interrompido)*	1
B	*Abandono Definitivo (não finalizado)*	1
B	*Abandono - Concluir recuperação do revestimento de 30"*	1
B	*Abandono Temporário (MLS-002)*	1
C	*Dissociação de Hidrato*	37
C	*Dissociação de Hidrato e Troca de VGL*	4
C	*Dissociação de Hidrato + Troca de VGL*	3
C	*Dissociação de hidrato (Retorno)*	1
C	*Dissociação de Hidrato e Abandono Temporário*	1
C	*Dissociação de hidrato nas LGL e LPO*	1
C	*Dissociação de Hidrato + Desincrustração*	1
D	*C - Restauração*	7
D	*Restauração*	6
D	*Teste de estanqueidade*	2
D	*BG-16 - Restauração*	2
D	*CH-32 - Restauração*	2
D	*CH-27 - Restauração*	1

Table 5.3: Examples of clusters using k-means with k=45.

The results in Table 5.3 indicate that using text mining and clustering algorithms to classify the workover operations according to their descriptions is possible. 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 potent tools to group the critical information and obtain new data classifying the workovers. This new classification will be used as an input to the duration prediction in Section 5.3.2.

Though, some improvements might still be possible, such as:

- **Text cleaning:** Replacing specific words of the problem or the Portuguese language that are semantically similar. For instance: "troc" (Portuguese stem word for "change") and "substit" (Portuguese stem word for "replace") share similar meanings and could one could be replaced by the other.
- **String distance and similarity:** testing the combination of other string distances with the methodology and comparing the final results.
- **Classification:** Others clustering and classification algorithms could be tested as well. A comparison between the k-means and the hieraquical clustering (h-cluster) is presented in Appendix E using heatmaps for the workover data. As shown in Appendix E, the k-means resulted in better groups. However, others classification algorithms and variations could still be tested.

The following section presents the regression models used to model the uncertainty in the workovers' durations after the treatment of the workover data in this Section.

5.3.2 Regression models for the workover duration

Statistical techniques and big data play an important role in the oil and gas upstream. There have been several successful cases using statistics to predict operations 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.

In this section, we apply the workover data treated in Section 5.3.1 in some parametric regression models to predict duration's uncertainty, as explained earlier in Figure 4.1. Two types of regressions are tested and evaluated: generalized linear models and ridge regression models.

To test and obtain a better fitting of the regression, some transformations of the duration of workover i in rig k (d_i^k) were also tested with the models. 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 in the data. Last, different settings on the regression modes were made. For instance, the GLMs were tested using Gaussian and Gamma distributions, and the ridge regression (RR) models were tested using Gaussian and Poisson distributions. These different settings are described below:

- GLM for duration (d_i^k) and Gaussian distribution: GLMdGaus.
- GLM for duration and Gamma distribution: GLMdGam.

- GLM for log of the duration ($log(d_i^k)$) and Gaussian distribution: GLMlogGaus.
- GLM for log of the duration and Gamma distribution: GLMlogGam.
- GLM for normalized Duration ($\frac{d_i^k - \min(d_i^k)}{\max(d_i^k) - min(d_i^k)}$) and Gaussian distribution: GLMnormGaus.
- GLM for normalized Duration and Gamma distribution: GLMnormGam.
- RR for duration and Gaussian distribution: RRdGaus.
- RR for duration and Poisson distribution: RRdPois.
- RR for log of the duration and Gaussian distribution: RRlogGaus.
- RR for log of the duration and Poisson distribution: RRlogPois.
- RR for normalized duration and Gaussian distribution: RRnormGaus.
- RR for normalized duration and Poisson distribution: RRnormPois.

Using the testing procedure described on Section 4.1.1, all combinations of response variables to predict were tested exhaustively for each of these regressions mentioned above. The best results for each regression model and setting are presented in Table 5.4.

Method	Distribution	Dependent Variable	Independent Variable	R^2	RMSE	pValue
GLM	Gaussian	d_i^k	WellWaterDepth, WellDepth, Subpool, Group, Basin, RigWaterDepth, RigDepth, $Clusters^{45}$, RigType	0.134	7.688	0.401
GLM	Gaussian	$log(d_i^k)$	WellWaterDepth, WellDepth, Subpool, Group, Basin, RigWaterDepth, RigDepth, $Clusters^{45}$, RigType	0.134	0.538	0.145
GLM	Gaussian	$\frac{d_i^k - \min(d_i^k)}{\max(d_i^k) - min(d_i^k)}$	WellWaterDepth, WellDepth, Subpool, Group, Basin, RigWaterDepth, RigDepth, $Clusters^{45}$, RigType	0.134	0.078	0.401
GLM	Gamma	d_i^k	WellWaterDepth, WellDepth, Subpool, Group, Basin, RigWaterDepth, RigDepth, $Clusters^{45}$, RigType	0.134	7.688	0.676
GLM	Gamma	$log(d_i^k)$	WellWaterDepth, WellDepth, Subpool, Group, Basin, RigWaterDepth, RigDepth, $Clusters^{45}$, RigType	0.134	0.540	0.169
GLM	Gamma	$\frac{d_i^k - \min(d_i^k)}{\max(d_i^k) - min(d_i^k)}$	WellWaterDepth, WellDepth, Subpool, Group, Basin, RigWaterDepth, RigDepth, $Clusters^{45}$, RigType	0.134	0.079	0.674
RR	Gaussian	d_i^k	Basin, RigType	0.241	7.898	0.699

Table 5.4: Best results for the regressions models

Method	Distribution	Dependent Variable	Independent Variable	R^2	RMSE	pValue
RR	Gaussian	$log(d_i^k)$	WellDepth, Subpool, Basin, $Clusters^{45}$, RigType	0.400	0.574	0.086
RR	Gaussian	$\frac{d_i^k - \min(d_i^k)}{\max(d_i^k) - min(d_i^k)}$	Group, Basin, $Clusters^{45}$, RigType	0.122	0.087	0.214
RR	Poisson	d_i^k	Basin, RigType	0.241	7.897	0.567
RR	Poisson	$log(d_i^k)$	WellWaterDepth, Basin, RigType	0.401	0.574	0.067
RR	Poisson	$\frac{d_i^k - \min(d_i^k)}{\max(d_i^k) - min(d_i^k)}$	Group, Basin, $Clusters^{45}$, RigType	0.120	0.087	0.219

Table 5.4: Best results for the regressions models

Analyzing Table 5.4, we can observe that all best 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 small clusters resulting from the text mining and classification in Section 5.3.1 ($Clusters^{45}$) were also a common attribute in most of the regression models, which indicates that the techniques were successful in revealing the underneath the 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 dummy variables.

The best-fitted models used ridge regression and a logarithmic duration for the workover ($log(d_i^k)$). The Gaussian distribution has a good R^2 (slightly lower than using the Poisson distribution) and a better p-value for a normal distribution for the errors, which suggests that it would be easier to fit distributions for them. Therefore, we have chosen to work with the log of the duration as variable dependent and to use the following Equation (5-3) obtained through the ridge regression model:

$$log(d_i^k) \sim (Intercept) + \alpha WellDepth_i + \beta Subpool_i + \gamma Basin_i + \delta Cluster_i^{45} + \phi RigType^k + \varepsilon, \tag{5-3}$$

where:

- d_i^k: Duration of the workover from well i by rig k.
- $WellDepth_i$: Depth of well i.
- $Subpool_i$: Subpool responsible for well i.
- $Basin_i$: Exploratory Basin where well i is located.
- $Cluster_i^{45}$: Cluster for the descriptions of the operation executed in well i (type of workover), obtained using k-means for $k = 45$.
- $RigType^k$: Type of the rig k.
- ε: Residuals or errors of the regression, its uncertainty.

The final coefficients for this regression are presented in Appendix F. Using this regression, Equation (5-3) can be rewritten and simplified to the following linear regression:

$$\begin{aligned} log(d_i^k) \sim &\quad (Intercept) + \alpha WellDepth_i + \psi Subpool_i + \gamma Basin_i + \\ &\quad \delta Cluster_i^{45} + \phi RigType^k + \varepsilon \quad (5\text{-}4) \\ log(d_i^k) \sim &\quad Intercept + WellEffect_i + RigEffect^k + \varepsilon \quad (5\text{-}5) \\ d_i^k \sim &\quad e^{Intercept + WellEffect_i + RigEffect^k} + \varepsilon \quad (5\text{-}6) \end{aligned}$$

$$d_i^k \sim \qquad \tilde{d_i^k} = \hat{d_i^k} + \varepsilon, \quad (5\text{-}7)$$

where d_i^k is the real duration of workover i in rig k, $WellEffect_i = \alpha WellDepth_i + \psi Subpool_i + \gamma Basin_i + \delta Cluster_i^{45}$, $RigEffect^k = \phi RigType^k$, $\tilde{d_i^k}$ is its approximation, $\hat{d_i^k}$ is its prediction from the regression ($\hat{d_i^k} = e^{Intercept + \alpha WellData_i + \beta RigData^k}$), and the distribution of the last ε (after the exponential operation) can be estimated using the residuals from the regression as a Normal distribution $N \sim (\mu = 2.522719, \sigma = 8.261636)$.

To summarize, this section describes how text mining and clustering algorithms can be used to clean, extract, and reveal data from the workover operations and applies the data in regression models, obtaining estimations of the duration of the workover operation for each well and rig and the uncertainty that affects it. Some improvements could still be possible. For instance, other regression methods and calibration algorithms, such as neural networks, or stacked ensembles, might achieve better results and should be tested in the future.

The following sections use the outputs of this predictive algorithm in regression-driven optimization models adapted for WRSP, comparing each proposed method.

6
A regression-driven optimization model for the workover rig scheduling problem

This chapter applies the regression-driven optimization formulation from Figure 4.1 in the workover rig scheduling problem presented in Section 5.1. For that, the regression estimations from the previous chapter are used with the instances generated in Section 5.2 as inputs for optimization models proposed in this Section. While approaching the regression-driven WRSP, two formulations are created and a final formulation for the regression-driven WRSP is proposed and compared with the studied company's current methodology.

Sections 6.1 and 6.2 present the first and final regression-driven formulations, respectively. The computational experiments comparing both proposed formulations and the current company's method are described in Section 6.3, which also includes a sensitivity analysis that studies the impact of regression error in the solution quality (Section 6.3.1).

6.1
The workover rig scheduling problem

As mentioned in the literature review in Section 2.4, 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 and represent 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. Pérez et al. (2016) adapted this formulation by removing the routing elements and reducing the variables indices, obtaining a time-index model with low dimensions. More recently, Carrilho et al. (2018) proposed a bucket-index model for the DRSP and Monemi et al. (2015) proposed an Arc-time-index formulation for the WRSP. The problem with these formulations is that they are unsuitable for chance-constrained models with uncertainty on the duration time. As the variables are indexed by a time-related set, any parameter with time uncertainty will lead to uncertainty on the variables indices, precluding the development of linear formulations. Therefore, these were deemed unsuitable as a formulation for the non-stochastic WRSP.

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 loss. As a result, this formulation was quickly adapted for this WRSP study, removing the time-index elements and modifying it to a routing formulation with the assumptions discussed in Section 5.1, such as release dates for the operations, rig hiring costs, and the selection of which wells to serve.

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 planning 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 and its sets, parameters, variables, objective function, and constraints are presented next.

Sets:

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

Parameters:

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

Variables:

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

The objective function (6-1) 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.

$$\text{Min} \quad \sum_{i \in J | i \neq 0} l_i \left[S_j + \sum_{j \in J} \sum_{k \in K} (d_i^k - a_i) X_{ij}^k + (H - a_i)(1 - \sum_{j \in J} \sum_{k \in K} X_{ij}^k) \right] + \sum_{k \in K} c^k Z^k \tag{6-1}$$

$$\begin{aligned}
\text{Subject to} \quad & \sum_{j \in J} X_{ji}^k = \sum_{j \in J} X_{ij}^k & \forall i \in J, k \in K \quad & \text{(6-2)} \\
& \sum_{k \in K} \sum_{i \in J} X_{ij}^k \leq 1 & \forall j \in J | j \neq 0 \quad & \text{(6-3)} \\
& \sum_{k \in K} \sum_{j \in J} X_{ij}^k \leq 1 & \forall i \in J | i \neq 0 \quad & \text{(6-4)} \\
& S_j - d_i^k \geq S_i - M(1 - X_{ij}^k) & \forall i \in J, j \in J, k \in K | i \neq 0 \quad & \text{(6-5)} \\
& S_i \geq a_i \sum_{k \in K} \sum_{j \in J} X_{ij}^k & \forall i \in J | i \neq 0 \quad & \text{(6-6)} \\
& \sum_{j \in J} X_{ij}^k \leq Z^k & \forall i \in J, k \in K \quad & \text{(6-7)} \\
& X_{ij}^k \in \{1, 0\} & \forall i \in J, j \in J, k \in K | e_i^k, i \neq j \quad & \text{(6-8)} \\
& S_i \in \mathbb{Z}^+ & \forall i \in J | i \neq 0 \quad & \text{(6-9)} \\
& Z^k \in \{1, 0\} & \forall k \in K. \quad & \text{(6-10)}
\end{aligned}$$

Constraints (6-2), (6-3), and (6-4) 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 (6-5) calculate each task j starting time (S_j) according to the previous service of the rig ($S_i + d_i^k$). Constraints (6-6) guarantee that the task i starting time (S_i) respects its release date (a_i). Constraints (6-7) 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 (6-8), (6-9), and (6-10) 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 equations (6-1) and (6-5). Next, we show how we have reformulated the model (6-1)–(6-10) to achieve better computational performance.

6.2
Reformulated workover rig scheduling problem

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 below:

- $X1_i^k$: If a rig k arrives at well j.
- $X2_i^k$: If a rig k leaves well i.
- W_i: If any rig serves (enters and leaves) well i.

The use of the auxiliary variables 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 Equation (6-11), it minimizes the total costs associated with the oil production losses and the fleet size cost.

$$\text{Min} \quad \sum_{i \in J | i \neq 0} l_i \left[S_i + \sum_{k \in K} (d_i^k - a_i) X1_i^k + (H - a_i)(1 - W_i) \right] + \sum_{k \in K} c^k Z^k \tag{6-11}$$

$$\begin{aligned}
\text{Subject to: } X1_i^k &= X2_i^k && \forall i \in J, k \in K && (6\text{-}12)\\
X1_i^k &= \sum_{j \in J} X_{ji}^k && \forall i \in J, k \in K && (6\text{-}13)\\
X2_i^k &= \sum_{j \in J} X_{ij}^k && \forall i \in J, k \in K && (6\text{-}14)\\
W_i &= \sum_{k \in K} X1_i^k && \forall i \in J | i \neq 0 && (6\text{-}15)\\
W_i &= \sum_{k \in K} X2_i^k && \forall i \in J | i \neq 0 && (6\text{-}16)\\
S_i - d_j^k &\geq S_j - M(1 - X_{ij}^k) && \forall i \in J, j \in J, k \in K | i \neq j && (6\text{-}17)\\
S_i &\geq a_i W_i && \forall i \in J | i \neq 0 && (6\text{-}18)\\
X1_i^k &\leq Z^k && \forall i \in J, k \in K && (6\text{-}19)
\end{aligned}$$

$$X_{ij}^k \in \{1,0\} \qquad \forall i \in J, j \in J, k \in K | e_i^k, i \neq j \tag{6-20}$$

$$X1_i^k \in \{1,0\} \qquad \forall i \in J, k \in K \tag{6-21}$$

$$X2_i^k \in \{1,0\} \qquad \forall i \in J, k \in K \tag{6-22}$$

$$W_i \in \{1,0\} \qquad \forall i \in J | i \neq 0 \tag{6-23}$$

$$S_i \in \mathbb{Z}^+ \qquad \forall i \in J \tag{6-24}$$

$$Z^k \in \{1,0\} \qquad \forall k \in K \tag{6-25}$$

New constraints were added to define the auxiliary variables and simplify the equations. Constraints (6-12) are flow balance rules. The new auxiliary variables ($X1_i^k$, $X2_i^k$, and W_i) are defined in constraints (6-13), (6-14), (6-15), and (6-16), and they guarantee that a well i can only be served once. Constraints (6-17) calculate the starting time (S_i) of each task i according to the previous service of the rig ($S_i + d_i^k$). Constraints (6-18) guarantee that the task i starting time (S_i) satisfies its release date (a_i). Constraints (6-19) connect variables Z^k and $X1_i^k$, forcing the model to hire a rig (Z^k) to execute a task i with this rig k ($X1_i^k$). The other constraints (6-20) to (6-25) state the domains of the variables.

6.3 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 instances were created based on these data. A detailed description of the instance generator was provided Section 5.2. 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 an 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 6.1 presents a solution comparison between both models, the original model (I) and the reformulated model (II), for different instances with a planning 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 seconds was also enforced to solve the models.

Jobs	Rigs	Instace	UB (avg. in M.)		LB (avg. in M.)		GAP (avg.)		Time (avg. in sec.)	
		Density	I	II	I	II	I	II	I	II
15	2	0.1	457.0	457.0	427.0	444.0	5%	2%	1204	1200
		0.5	484.9	484.2	419.5	441.5	10%	7%	1812	1231
		0.9	465.7	465.7	388.1	441.1	15%	4%	2400	1841
	3	0.1	431.7	431.7	431.7	431.7	0%	0%	11	3
		0.5	446.3	446.3	446.3	446.3	0%	0%	8	1
		0.9	459.5	459.5	459.5	459.5	0%	0%	7	1
25	5	0.1	697.0	697.0	615.8	675.6	11%	3%	2462	2406
		0.5	737.6	737.6	612.1	672.6	16%	8%	2512	2405
		0.9	715.2	704.5	657.0	704.5	7%	0%	1875	243
	10	0.1	649.2	649.2	643.1	649.2	1%	0%	1861	526
		0.5	684.0	685.0	684.0	642.8	0%	5%	1444	1860
		0.9	664.8	664.8	664.8	664.8	0%	0%	575	714
	15	0.1	609.3	609.3	603.6	609.3	1%	0%	1852	1746
		0.5	639.2	639.2	605.1	599.2	5%	6%	1829	1820
		0.9	655.5	655.2	610.6	591.6	6%	8%	1839	1825
50	5	0.1	1334.4	1330.9	840.4	941.6	37%	29%	3600	3600
		0.5	1434.4	1427.7	836.9	935.6	42%	35%	3600	3600
		0.9	1434.2	1385.0	824.7	939.3	42%	32%	3600	3600
	10	0.1	1133.2	1153.5	861.8	898.4	24%	22%	3600	3600
		0.5	1165.4	1142.4	871.3	904.1	25%	21%	3600	3600
		0.9	1124.1	1126.0	885.7	899.7	21%	20%	3600	3600
	15	0.1	1318.2	1212.9	759.2	814.6	42%	32%	3600	3600
		0.5	1300.9	1221.8	759.1	846.5	39%	29%	3600	3600
		0.9	1212.5	1221.4	760.1	864.7	36%	27%	3600	3603
75	5	0.1	1838.9	1794.5	1109.9	1183.5	38%	33%	3600	3600
		0.5	2003.3	1879.8	1105.2	1161.2	42%	36%	3600	3600
		0.9	1976.0	1852.7	1105.5	1153.3	40%	35%	3600	3600
	10	0.1	1726.3	1633.4	951.4	1055.7	45%	35%	3600	3600
		0.5	1741.0	1620.0	950.4	1049.4	45%	35%	3600	3600
		0.9	1543.2	1616.7	957.6	1088.4	38%	32%	3600	3600
	15	0.1	1630.0	1635.2	966.7	1054.4	41%	36%	3600	3600
		0.5	1783.3	1715.1	966.8	1042.4	46%	39%	3600	3600
		0.9	1620.7	1574.9	966.9	1040.5	40%	34%	3600	3600

Table 6.1: Comparison between models (I) and (II).

The results in Table 6.1 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 reached the 3600-seconds 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 enables the model to obtain better solutions.

Another important analysis is to compare the non-stochastic optimization model with the current approach of the company. As mentioned in Section 5.1, 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 conservative 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 and how much the solution needs to change to adjust itself to the actual workover durations, two comparisons were made. First, the schedule was obtained from the optimization model, using the conservative duration and the regression estimation. These are then compared with the schedule obtained knowing the actual duration of the well i, that is, the schedule with perfect information. The comparison is performed in terms of their relative difference between their objective function values, where the schedule with perfect information is the reference. This comparison analyzes the robustness of the models' solutions and can be seen in the green box plots in Figure 6.1.

The second comparison considers that there is an option for rescheduling considering the actual workover duration for each well i ($\tilde{d}_i$), but considering the rig fleet and list of served wells obtained when performing the schedule using estimated workover durations ($\bar{d}_i$ or $\hat{d}_i^k$). 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 focuses on the solution flexibility and is presented as the orange bars in Figure 6.1. The vertical axis in Figure 6.1 is the percentage of deviation of each comparison. The comparisons made are specified according to the box-plot color (green refers to the models robustness and orange represents the solution flexibility), and the rows are, respectively, the current method of the studied company (average or conservative duration) and the proposed data-driven methodology (regression duration).

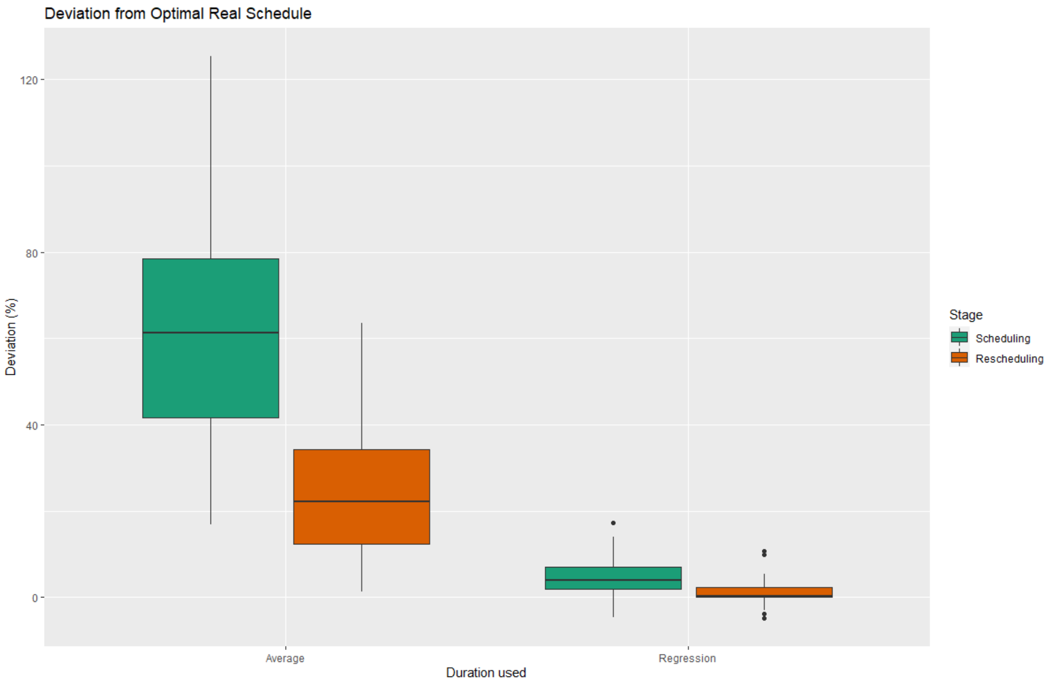


Figure 6.1: Box-plot comparing the objective function difference to the schedule with perfect information according to the duration estimation and the stage (scheduling or rescheduling).

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 conservative duration.

6.3.1 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 subjected to 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.*, $\tilde{d_i^k} = \hat{d_i^k} + \varepsilon$, 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 $N \sim (\mu = 2.522719, \sigma = 8.261636)$, as mentioned in Section 5.3.1. 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. Five hundred simulations of the regression error (ε) are made by sampling from the Normal distribution $N \sim (\mu = 2.522719, \sigma = 8.261636)$, 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. Figure 6.2 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. 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.

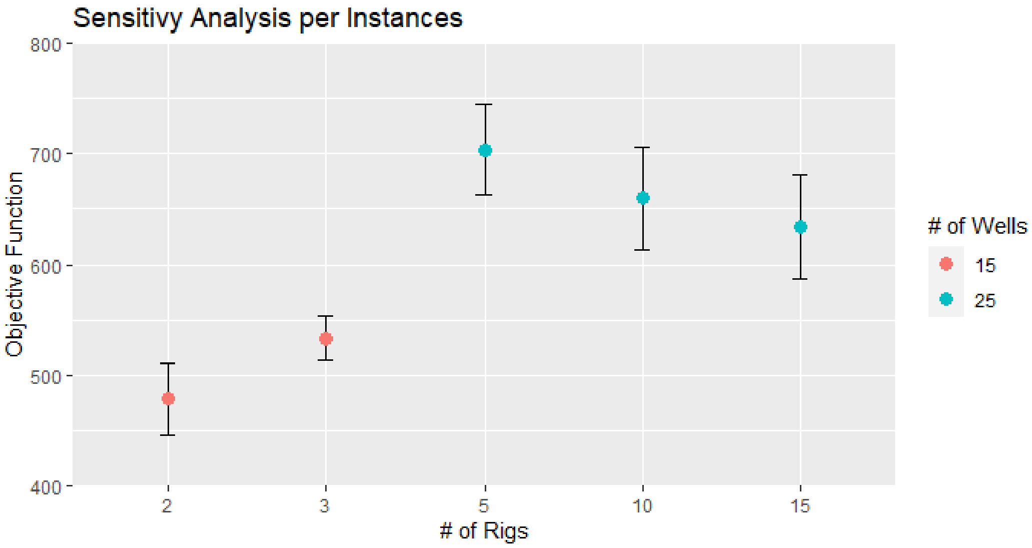


Figure 6.2: Confidence intervals of the objective function due to the regression error.

Analyzing Figure 6.2, we can observe that the objective function and its variability is highly influenced by the instance size. The larger the number of wells needing intervention, the larger are 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. Regardless of the instance, the regression residuals do not disturb more than 10% of the object function value.

Despite generating small disturbance, the duration uncertainty can not be diminished as, in some cases, it can lead to infeasible or undesirable solutions. For instance, one of the assumptions of this sensitivity analysis was that a well expected to be served could be easily deselected if needed. However, deselecting a well in the short-term should always be avoided as this decision has a tremendous impact on several plannings and agreements made by the company, possibly resulting in large expenses of capital and lost of credibility with suppliers and clients.

Therefore, it is crucial to guarantee a hedge for these potential losses and minimize their risk to an acceptable level. For that, a better approach with a complete analysis of the data and considering the uncertainty in the workover operations is crucial. The regression-driven chance-constrained optimization methodologies proposed in Section 4.1.2.1 could solve this problem.

The regression-driven optimization methodology proposed in this chapter and its application in the WRSP culminated in the publishing of an article "A data-driven optimization model for the workover rig scheduling problem: Case study in an oil company" in the journal Computers & Chemical Engineering (Santos et al. 2023).

Next, we apply the regression-driven chance-constrained optimization methodologies on the studied WRSP, aiming to deal with the solution feasibility uncertainty.

7
Regression-driven optimization models for the workover rig scheduling problem

The SLR allowed us to identify opportunities for the WRSP, including the application of joint chance-constrained formulations and data-driven optimization models. Another insight from SLR was related to the gap between academic studies and industry demands. Very few studies were actually implemented in the industry, despite the importance of the rig scheduling problems. Aiming to address this gap and exploit the aforementioned opportunities, a regression-driven optimization model was formulated for a real case study of the WRSP in Section 5. A sensitivity analysis was made and the importance of data-driven joint chance-constrained models was once again reinforced.

However, most data-driven JCC models in the literature are non-linear, as mentioned earlier in Chapter 3. Thus, a new approach based on stochastic programming was proposed as a linear alternative. This new methodology of regression-driven stochastic joint chance-constrained optimization (presented in Section 4.1.2.2) is applied to the WRSP in this section. In addition, we also compare this proposed regression-driven stochastic JCC model with two other regression-driven stochastic chance-constraints variations: integrated-CC and budget-constrained. As the regression-driven models from Chapter 6, the stochastic models proposed and tested in this section use the data from the regression algorithms from Section 5.3.1. The main difference is that the regression uncertainty is also considered in these models.

7.1
Mathematical modeling

As mentioned in Section 5.1, the workover rig scheduling problem setting is surrounded by risks. One of its main uncertainties is the duration of the operations. In Section 5.3, the duration was analyzed and treated using text mining, clustering, and regression algorithms. In this section, the outputs of these data processes are used in the optimization under uncertainty in a data-driven optimization approach with three chance-constraints variations, in which the mathematical models are based on the reformulated regression-driven mathematical model proposed in Section 6.2 and the scenario-based data-driven methodology from Section 4.1.2.2. Sections 7.1.2 and 7.1.3 present the regression joint CC workover rig scheduling models.

7.1.1 The chance-constrained workover rig scheduling problem

In the WRSP, the uncertainty exists in the duration of the tasks. When fleet sizing or scheduling for the short and mid-terms, the ideal is to have a solution that can accommodate minor disruptions in the planning. Therefore, it might be helpful to allow some tasks to overlap with each other (when a rig has to serve two or more wells simultaneously) within some tolerance, as the main goal is to generate a solution that will not need major reschedules. Another particularity of this short-term case study is that after sequencing a rig, several contracts and commitments are taken. The ideal is to keep the actual schedule as similar as possible to the planned schedule, respecting the planned fleet sizing, rig allocation, and sequencing.

This problem can be formulated as a chance-constrained model, in which constraint (6-17) from the regression-driven WRSP (Section 6.2) becomes a probabilistic constraint. The original constraint is for each well i, well j, and rig k. However, the goal is to guarantee that the sequence of the schedule is feasible and every well's workover does not exceed the start of any other workover or the planning horizon with a certain probability, *i.e.*, the new chance constraint should guarantee that the schedule is feasible with a confidence level α for every sequencing of wells i and j and rig k associated with the constraint. As a result, we have chosen to model the WRSP as a joint chance-constrained programming problem, represented by the constraint (7-7). Using the notation and variables from Section 6.2, the joint chance-constrained workover rig scheduling problem is given as follows:

$$\text{Min} \quad \mathbb{E}\left(\sum_{i\in J|i\neq 0} l_i \left[S_i + \sum_{k\in K}(\tilde{d}_i^k - a_i)X1_i^k\right]\right) + \sum_{i\in J|i\neq 0}(H-a_i)l_i(1-W_i) + \sum_{k\in K} c^k Z^k \tag{7-1}$$

Subject to

$$X1_i^k = \sum_{j\in J} X_{ji}^k \qquad \forall i,k \tag{7-2}$$

$$X2_i^k = \sum_{j\in J} X_{ij}^k \qquad \forall i,k \tag{7-3}$$

$$W_i = \sum_{k\in K} X1_i^k \qquad \forall i|i\neq 0 \tag{7-4}$$

$$W_i = \sum_{k \in K} X2_i^k \qquad \forall i | i \neq 0 \qquad (7\text{-}5)$$

$$X1_i^k = X2_i^k \qquad \forall i, k \qquad (7\text{-}6)$$

$$\mathbb{P}\left[S_j + H(1 - X_{ij}^k) \geq S_i + \tilde{d}_i^k, \forall i, j, k | i \neq 0, i \neq j\right] \geq \alpha \qquad (7\text{-}7)$$

$$S_i \geq a_i * W_i \qquad \forall i | i \neq 0 \qquad (7\text{-}8)$$

$$X_{ij}^k \leq Z^k \qquad \forall i, j, k | i \neq j \qquad (7\text{-}9)$$

$$X_{ij}^k \in \{1, 0\} \qquad \forall i, j, k | i \neq j \qquad (7\text{-}10)$$

$$X1_i^k \in \{1, 0\} \qquad \forall i, k \qquad (7\text{-}11)$$

$$X2_i^k \in \{1, 0\} \qquad \forall i, k \qquad (7\text{-}12)$$

$$S_i \geq 0 \vee S_i \in \mathbb{Z}^+ \qquad \forall i | i \neq 0 \qquad (7\text{-}13)$$

$$W_i \in \{1, 0\} \qquad \forall i \qquad (7\text{-}14)$$

$$Z^k \in \{1, 0\} \qquad \forall k \qquad (7\text{-}15)$$

The objective function (7-1) minimizes the expected costs. The first cost term represents the expected oil production loss associated with the tardiness of the task, which is the expected time to execute the task after its release. The second term refers to the cost of not serving a well, which can be explained as the production loss from the entire time horizon starting from its release date. The last term of the objective function is related to the fleet size cost.

Auxiliary variables are defined in constraints (7-2), (7-3), (7-4), and (7-5), and they guarantee that a well can only be served once. Constraints (7-6) are flow balance rules. Note that constraint (7-6) is already presented in constraints (7-4) and (7-5), but we insert it again in the model as valid equalities, potentially improving its linear relaxation. The joint chance constraint (7-7) calculates the starting time of each task according to the rig's previous service and determines that each well has a confidence level $1 - \alpha$ of being feasible. This confidence level refers to the probability of this constraint for a well i be simultaneously satisfied for all other wells j and rigs k, *i.e.*, the well i has a probability $1 - \alpha$ of respecting, in any rig k, the start of any other well. Constraints (7-8) guarantee that the task starting time respects its release date.

Constraints (7-9) connect variables Z^k and X_{ij}^k, forcing the model to hire a rig in order to execute a task with this rig. The last constraints (7-10), (7-11), (7-12), (7-13), (7-14), and (7-15) are related to the variables' domains.

Thanks to the text mining, classification algorithms, and regression models that were used to estimate the duration of a workover in a specific well and rig, it is possible to adapt this generalized chance-constrained WRSP model to a regression-driven model, where the uncertain duration $\tilde{d}_i^k$ is replaced by the estimated duration and regression error ($\hat{d}_i^k + \varepsilon$). These estimations ($\hat{d}_i^k$) represent the effects of the well and rig properties on the workover duration, and the distribution of the regression residuals can be used to represent the uncertainty unrelated to the well and the rig (ε), thus reducing infeasibility risks. As a result, we can transform the objective function (7-1) and constraint (7-7), respectively, into:

$$\text{Min} \quad \mathbb{E}\left(\sum_{i \in J | i \neq 0} l_i \left[S_i + \sum_{k \in K} (\hat{d}_i^k + \varepsilon - a_i) X1_i^k \right] \right) + \sum_{i \in J | i \neq 0} (H - a_i) l_i (1 - W_i) + \sum_{k \in K} c^k Z^k \tag{7-16}$$

$$\mathbb{P}\left[S_i - S_j - M(1 - X_{ij}^k) \leq -\hat{d}_i^k - \varepsilon, \forall i, j, k | i \neq 0, i \neq j \right] \geq \alpha \tag{7-17}$$

where $\varepsilon \sim N(\mu, \sigma)$, the residuals error follow a normal distribution, and $\hat{d}_i^k$ are the regression estimations of the duration of a workover on well i and rig k.

Nevertheless, the regression-driven joint chance constraint represented in (7-17) cannot yet be used in any optimization solver due to the probability function. The following sections use the methodologies of Section 4.1.2.1 to explicitly represent the JCC with approximations and reformulations, enabling its application. Another possibility would be to formulate the JCC with Kernel smoothing methods, as made by Calfa et al. (2015). This alternative is outside of the thesis scope but is presented in the Appendix G.

7.1.2
Regression-driven JCC deterministic-equivalent WRSP model

Aiming to propose a deterministic-equivalent for the regression-driven JCC-WRSP model presented in the previous section, this section applies the methodology proposed in Section 4.1.2.1 to the JCC-WRSP. This methodology is based on Biswal et al. (2005) and Sahoo and Biswal (2005) and exploits properties of normal and log-normal distributions to reformulate the joint constraints when the uncertainty in the RHS follows normal and log-normal distributions. Following the steps of the reformulations from (3-14)-(3-17), we can modify constraint (7-17) for the case in which the residuals are independent random variables and follow a Normal distribution $N(\mu, \sigma)$.

$$\mathbb{P}\left[S_i - S_j - M(1 - X_{ij}^k) + \hat{d}_i^{\,k} \leq -\varepsilon, \forall i, j, k | i \neq 0, i \neq j \right] \geq \alpha \tag{7-18}$$

$$\prod_{i \in J, j \in J, k \in K | i \neq 0, i \neq j} \mathbb{P}\left[S_i - S_j - H(1 - X_{ij}^k) + \hat{d}_i^k \leq -\varepsilon \right] \geq \alpha \tag{7-19}$$

$$\prod_{i\in J, j\in J, k\in K | i\neq 0, i\neq j} \mathbb{P}\left[-\varepsilon \geq S_i - S_j - H(1 - X_{ij}^k) + \hat{d}_i^k\right] \geq \alpha \tag{7-20}$$

$$\prod_{i\in J, j\in J, k\in K | i\neq 0, i\neq j} \mathbb{P}\left[\varepsilon^{'} \geq g_{ij}^k(x)\right] \geq \alpha \tag{7-21}$$

$$\prod_{i\in J, j\in J, k\in K | i\neq 0, i\neq j} \mathbb{P}\left[\frac{\varepsilon^{'} - \mu^{'}}{\sigma} \geq (\frac{g_{ij}^k(x) - \mu^{'}}{\sigma}\right] \geq \alpha \tag{7-22}$$

$$\prod_{i\in J, j\in J, k\in K | i\neq 0, i\neq j} \left[1 - \Phi\left(\frac{g_{ij}^k(x) - \mu^{'}}{\sigma}\right)\right] \geq \alpha, \tag{7-23}$$

where $\varepsilon^{'} = -\varepsilon$, $\mu^{'} = -\mu$, $\varepsilon^{'} \sim N(\mu^{'}, \sigma)$, $g_{ij}^k(x) = S_i - S_j - H(1 - X_{ij}^k) + \hat{d}_i^k$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, which is equal to $\Phi(x) = \mathbb{P}(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\inf}^{\inf} e^{-\frac{u^s}{2}} du$. As mentioned in Section 4.1.2.1, if we considered $\mu = 0$, then Constraint (7-23) could be simplified into:

$$\prod_{i\in J, j\in J, k\in K | i\neq 0, i\neq j} \left[1 - \Phi\left(\frac{S_i - S_j - H(1 - X_{ij}^k) + \hat{d}_i^k}{\sigma}\right)\right] \geq \alpha \tag{7-24}$$

The problem is that the regression-driven joint chance-constrained deterministic equivalent reformulation for the workover rig scheduling problem is still a MINLP model. A possible solution for this is to use the regression-based formulation of the JCC-WRSP and adapt it to a scenario-based model, as proposed in the regression-driven stochastic JCC formulations from Section 4.1.2.2. Their application in the WRSP will be discussed in the next section.

7.1.3
Regression-driven stochastic joint chance-constrained WRSP model

Aiming to develop a linear programming alternative for the joint chance-constrained workover rig scheduling problem, we apply the methodology from Section 4.1.2.2 on the workover rig scheduling problem, reformulating the model from Section 7.1.2 to a stochastic optimization model based on Luedtke and Ahmed (2008), Nemirovski and Shapiro (2007) and Nikzad et al. (2019). For this, sets, parameters, and variables were required to be modified or added as follows:

New Set:

- $\omega \in \{1, 2, \ldots, \Omega\}$: Scenarios for the uncertainty. Each scenario represents a realization of the uncertainty.

Parameters modified or added:

- π^{ω}: Probability (weight of the Scenario, initially equals to $1/|\Omega|$).
- $\hat{d}_i^k$: Duration estimated by the ridge regression on the workover of rig k in well i.
- $\varepsilon_i^{k\omega}$: Duration disturbance associated with regression estimation error for the workover duration of well i in rig k and scenario ω.

Variables modified or added:

- S_i^{ω}: Non-negative integer variable that indicates the starting time of task i at scenario ω (second-stage decision of scheduling).
- $V_{ij}^{k\omega}$: Slack variable between [0,1] that measures the relaxation of constraint (ijk) feasibility in scenario ω (second-stage decision related with the JCC).
- Y^{ω}: Binary variables indicating the respect of the joint chance constraints of the schedule in scenario ω, *i.e.*, if the well i is feasible or not in scenario ω (second-stage decision related to the JCC).
- X_{ij}^k, $X1_i^k$, $X2_i^k$, W_i and Z^k: These variables are first-stage decisions (fleet sizing, well service selection, rig allocation, and sequencing) and, therefore, no modification in them is needed.

The objective function of the model is presented in Equation (7-25). Its goal is to minimize the expected costs. In this study case, the first-stage decisions are the fleet size (the decision of which rigs will be hired), the selection of the wells (which wells will be served), rig allocation (which rig will serve each well), and rig sequencing (the sequence/order of the wells that each rig serve). Meanwhile, the second-stage decisions are the scheduling (when each wells selected will be served in that scenario) and the infeasibilities (if the schedule will be feasible in that scenario and how much will that infeasibility be). As a result, the objective function is divided into first-stage costs and second-stage costs. The first-stages costs are associated with the decisions made before the occurrence of the uncertainty, *i.e.*, before the duration of workovers is known. The first-term of these costs is the oil production loss for not serving a well, which is equal to oil production loss for the entire time horizon observing the release date. The second term is the fleet size cost. Alternatively, the second-stage costs are associated with the decisions made after the realization of the uncertainty. Each scenario has a probability π^{ω} and the second-stage expected costs are equal to the expected oil production loss associated with the time to execute the task after its release. The objective function and its constraints are presented as follows:

$$
\begin{aligned}
\text{Min} \sum_{i\in J|i\neq 0} (H-a_i)l_i(1-W_i) + \sum_{k\in K} c^k Z^k \\
+ \sum_{\omega\in\Omega} \pi^{\omega} \sum_{i\in J|i\neq 0} l_i \left[S_i^{\omega} + \sum_{k\in K} (\hat{d}_i^k + \varepsilon_i^{k\omega} - a_i) X1_i^k \right]
\end{aligned}
\tag{7-25}
$$

Subject to

$$X1_i^k = \sum_{j\in J} X_{ji}^k \qquad \forall i,k|e_i^k = 1 \tag{7-26}$$

$$X2_i^k = \sum_{j\in J} X_{ij}^k \qquad \forall i,k|e_i^k = 1 \tag{7-27}$$

$$W_i = \sum_{k\in K} X_i^k \qquad \forall i \tag{7-28}$$

$$W_i = \sum_{k\in K} X_i^k \qquad \forall i \tag{7-29}$$

$$X1_i^k = X2_i^k \qquad \forall i,k|e_i^k \tag{7-30}$$

$$S_j^{\omega} + H(1 - X_{ij}^k) \geq S_i^{\omega} + \hat{d}_i^k + \varepsilon_i^{k\omega} V_{ij}^{k\omega} \qquad \forall i,j,k,\omega|e_i^k, 0 \neq i \neq j \tag{7-31}$$

$$1 - V_{ij}^{k\omega} \leq X_{ij}^k \qquad \forall i,j,k,\omega|e_i^k, 0 \neq i \neq j \tag{7-32}$$

$$V_{ij}^{k\omega} \leq Y^{\omega} \qquad \forall i,j,k,\omega|e_i^k, 0 \neq i \neq j \tag{7-33}$$

$$1 - V_{ij}^{k\omega} \geq Y^{\omega} \qquad \forall i,j,k,\omega|e_i^k, 0 \neq i \neq j \tag{7-34}$$

$$\sum_{\omega\in\Omega} \pi^{\omega} Y^{\omega} \geq \alpha \tag{7-35}$$

$$S_i^{\omega} \geq a_i * W_i \qquad \forall i,\omega|i \neq 0 \tag{7-36}$$

$$X_{ij}^k \leq Z^k \qquad \forall i,j,k|e_i^k = 1, 0 \neq i \neq j \tag{7-37}$$

$$X_{ij}^k \in \{1,0\} \qquad \forall i,j,k|e_i^k = 1, 0 \neq i \neq j \tag{7-38}$$

$$X1_i^k \in \{1,0\} \qquad \forall i,k|e_i^k = 1 \tag{7-39}$$

$$X2_i^k \in \{1,0\} \qquad \forall i,k|e_i^k = 1 \tag{7-40}$$

$$V_{ij}^{k\omega} \in [1,0] \qquad \forall i,j,k,\omega|e_i^k = 1, 0 \neq i \neq j \tag{7-41}$$

$$Y^{\omega} \in \{1,0\} \qquad \forall \omega \tag{7-42}$$

$$S_i^{\omega} \in \mathbb{Z}^+ \qquad \forall i,\omega|i \neq 0 \tag{7-43}$$

$$W_i \in \{1,0\} \qquad \forall i \tag{7-44}$$

$$Z^k \in \{1,0\} \qquad \forall k \tag{7-45}$$

Auxiliary variables are defined in constraints (7-26), (7-27), (7-28), and (7-29). Constraints (7-28) and (7-29) also define when a well is served. Constraints (7-30) are flow balance rules from vehicle routing models.

The joint chance constraints are defined in constraints (7-31), (7-32), (7-33), EQ:733b, and (7-35). Constraint (7-31) calculate the starting time of each task according to the rig's previous service, where the slack variable $V_{ij}^{k\omega}$ allows some violation within a range. Constraints (7-32) allow that slack variable $V_{ij}^{k\omega}$ can only be applied to reduce $\varepsilon_i^{k\omega}$ when the probabilistic constraint is active, *i.e.*, when X_{ij}^k equals to one. If the slack variable is used, then constraints (7-33) indicate that the sequencing is infeasible in scenario ω, *i.e.*, Y^ω equals to one. Last, constraint (7-35) guarantees that the joint probabilistic constraint respects the confidence level α of the schedule being feasible.

Constraints (7-36) guarantee that the task starting time respects its release date. Constraints (7-37) connect variables Z^k and X_{ij}^k, forcing the model to hire a rig in order to execute a task with this rig. The others constraints (7-38), (7-39), (7-40), (7-41), (7-42), (7-43), (7-44), and (7-45) are related with the variables domains.

Next, other chance-constraints variations are presented for the WRSP.

7.1.3.1 Regression-driven stochastic integrated chance-constrained WRSP model

Instead of guaranteeing the probability of infeasibility under a certain level, one might desire to guarantee that the expected magnitude of the infeasibility is never greater or equal to a value. This is the case of the integrated chance-constraint, introduced by Haneveld (1986) and Prékopa (2003). In Section 4.1.2.2, a methodology for regression-driven stochastic integrated chance-constrained optimization was proposed as an alternative for the JCCs. In this section, we apply this alternative approach in our WRSP.

For that, variable Y^ω must be converted to a non-negative interval and the regression-driven joint chance constraints (7-31), (7-32), (7-33), and (7-35) must be replaced by the following constraints:

$$S_j^\omega + H(1 - X_{ij}^k) \geq S_i^\omega + \hat{d}_i^k + \varepsilon_i^{k\omega} V_{ij}^{k\omega} \qquad \forall i, j, k, \omega | e_i^k, 0 \neq i \neq j \qquad (7\text{-}46)$$

$$1 - V_{ij}^{k\omega} \leq X_{ij}^k \qquad \forall i, j, k, \omega | e_i^k, 0 \neq i \neq j \qquad (7\text{-}47)$$

$$\varepsilon_i^{k\omega} \cdot (1 - V_{ij}^{k\omega}) \leq Y^\omega \qquad \forall i, j, k, \omega | e_i^k, 0 \neq i \neq j \qquad (7\text{-}48)$$

$$\sum_{\omega \in \Omega} \pi^\omega Y^\omega \leq \mu_\varepsilon \alpha + \frac{K\sigma_\varepsilon}{\sqrt{(|\Omega)|}} \qquad (7\text{-}49)$$

$$V_{ij}^{k\omega} \in [1, 0] \qquad \forall i, j, k, \omega | e_i^k, 0 \neq i \neq j \qquad (7\text{-}50)$$

$$Y^\omega \geq 0 \qquad \forall \omega \qquad (7\text{-}51)$$

$$S_i^\omega \in \mathbb{Z}^+ \qquad \forall i, \omega | i \neq 0 \qquad (7\text{-}52)$$

$$X_{ij}^k \in \{1, 0\} \qquad \forall i, j, k | e_i^k, 0 \neq i \neq j \qquad (7\text{-}53)$$

The regression-driven integrated chance-constraints are represented by constraints (7-46), (7-47), and (7-48). Just as in the JCC, constraints (7-46) are the chance constraints with the slack variable $V_{ij}^{k\omega}$ that allows some violations within a range (not greater than the value of ε_i^{ω}). Constraints (7-47) force that the slack variable is only used when the chance-constraint is active. The worst violation of the chance-constraint each scenario, *i.e.*, the largest product of ε_i^{ω} and $V_{ij}^{k\omega}$ is calculated as Y^{ω} in constraint (7-48). The expected worst violation is calculated in constraint (7-49), which also guarantees that its values respect the upper bound considering K from the percentile point associated with an α confidence level of the standard Normal distribution. Last, constraints (7-50), (7-51), (7-52), and (7-53) are related to the variables' domains.

Next, another chance-constraint variation is also applied in the WRSP.

7.1.3.2 Regression-driven stochastic budget-constrained WRSP model

The other scenario-based formulation proposed in Section 4.1.2.2 is the budget-constrained, which focuses on limiting the total amount of feasibility allowed in each scenario, regardless of their probability or expected value. This approach can be applied in the WRSP by replacing the regression-driven joint chance constraints (7-31), (7-32), (7-33), and (7-35) with the following set of constraints:

$$
\begin{aligned}
& S_j^{\omega} + H(1 - X_{ij}^k) \geq S_i^{\omega} + \hat{d}_i^k + \varepsilon_i^{k\omega} V_{ij}^{k\omega} && \forall i,j,k,\omega | e_i^k, 0 \neq i \neq j && \text{(7-54)}\\
& 1 - V_{ij}^{k\omega} \leq X_{ij}^k && \forall i,j,k,\omega | e_i^k, 0 \neq i \neq j && \text{(7-55)}\\
& \sum_{i \in J} \sum_{j \in J} \sum_{k \in K} \varepsilon_i^{k\omega}(1 - V_{ij}^{\omega}) \leq H\alpha && \forall \omega && \text{(7-56)}\\
& V_{ij}^{k\omega} \in [1,0] && \forall i,j,k,\omega | e_i^k, 0 \neq i \neq j && \text{(7-57)}\\
& S_i^{\omega} \in \mathbb{Z}^+ && \forall i,\omega | i \neq 0 && \text{(7-58)}\\
& X_{ij}^k \in \{1,0\} && \forall i,j,k | e_i^k, 0 \neq i \neq j && \text{(7-59)}
\end{aligned}
$$

The regression-driven budget-constraints are represented by constraints (7-54), (7-55), and (7-56). Just as in the other models, constraint (7-46) is the chance constraints with the slack variable $V_{ij}^{k\omega}$ that allows some violations within a range (not greater than the value of ε_i^{ω}). Constraint (7-47) forces that the slack variable is only used when the chance-constraint is active. The total violation of the chance-constraints in each scenario, *i.e.*, the sum of the product of ε_i^{ω} and $1 - V_{ij}^{k\omega}$ is calculated in constraint (7-56), which assures that this sum respects the bound $H \cdot \alpha$. Last, constraints (7-57), (7-58), and (7-59) are related to the variables domains.

7.2 Computational experiments

The models proposed in this chapter were implemented using the Julia programming language and Gurobi solver and executed in a computer with *Intel(R) Core(TM) i7-5960X CPU 3.00 GHz* and RAM memory of 64 Gigabytes. These computer experiments were made using the instances discussed in Section 5.2 and are presented in this section. First, Section 7.2.1 details the scenario generation and reduction. Section 7.2.2 first illustrates an example of the proposed regression-driven stochastic JCC-WRSP model, tests it in several instances, and analyzes the solution behavior. Then, Section 7.2.3 compares, under the offshore WRSP, this regression-driven stochastic JCC model with other chance constraint variations. Last, Section 7.2.4 compares the stochastic models against the non-stochastic approach from Chapter 6.

7.2.1 Scenario generation and reduction

Before testing a scenario-based stochastic model, it is important to generate a scenario population that faithfully represents the uncertainty and, in most cases, this scenario population needs to be reduced to a smaller, yet statistically significant, sample. This section is dedicated to presenting the results of this scenario generation and reduction. We apply the methodology from Figure 4.5 to generate a scenario sample for regression error, which follows the Normal distribution estimated in Section 5.3.1 ($N \sim (\mu = 2.522719, \sigma = 8.261636)$).

First, we use Monte Carlo simulation (MCS) to generate several samples of the regression uncertainty. The results of these MCS are shown in Figure 7.1. Each line represents the sample mean (y-axis) of a MCS run (colors) as its size grows (x-axis). Each run stops when the absolute difference between the sample mean and the actual distribution's mean ($\mu = 2.522719$) is lower than $\frac{10^{-3}}{\mu}$.

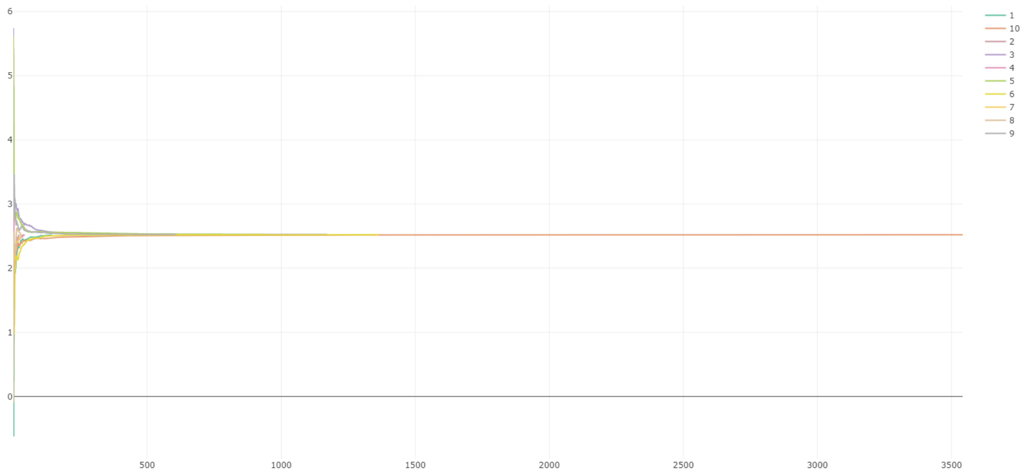


Figure 7.1: Monte Carlo simulation for the regression uncertainty.

Analyzing Figure 7.1, we can observe that, for most runs, the difference between the sample mean and the actual distribution's mean is lower than 0.1% after two thousand iterations, indicating that the sample mean converges to the distribution mean after this number of iterations. In other words, a set of 2,000 scenarios ($S = 2000$) generated sampling the Normal distribution, $N \sim (\mu = 2.522719, \sigma = 8.261636)$, can be used to accurately represent the regression error ε.

However, this number of scenarios is extremely large and culminates in an enormous and practically infeasible computational effort to solve an optimization model of this size.

Aiming to determine a smaller set of scenarios that can accurately represent these 2,000 scenarios, the Wasserstein distance-based method described in Section 3.3.3 was used iteratively for $S^* = \{5, 10, 15, \dots, 40, 45, 50\}$, keeping the scenarios subsets contained in each other, *i.e.*, the S^* scenarios obtained for 5 must be in the S^* scenarios for 10 and so recursively for 10 and 15, 15 and 20, and so forth, where the last S^* scenarios of 50 are contained in the 2,000 initial scenarios (S) obtained in the MCS. To obtain a number of scenario suitable for any confidence level, six different confidence levels were also tested ($\alpha = \{0.70, 0.75, \dots, 0.90, 0.95\}$) and the two metrics presented in Chapter 3, *Error of Approximation* and *Feasibility Frequency*, were used. Their results are presented in Figure 7.2, where the charts in the first row is for the *Error of Approximation (%)* metric, the second chart row is related to the *Feasibility Frequency*, and each column of the charts is for a joint chance constraint confidence level α. All charts have the same shared x-axis that represents the scenario size. The results are average values taken of the metrics calculated separately for different instances.

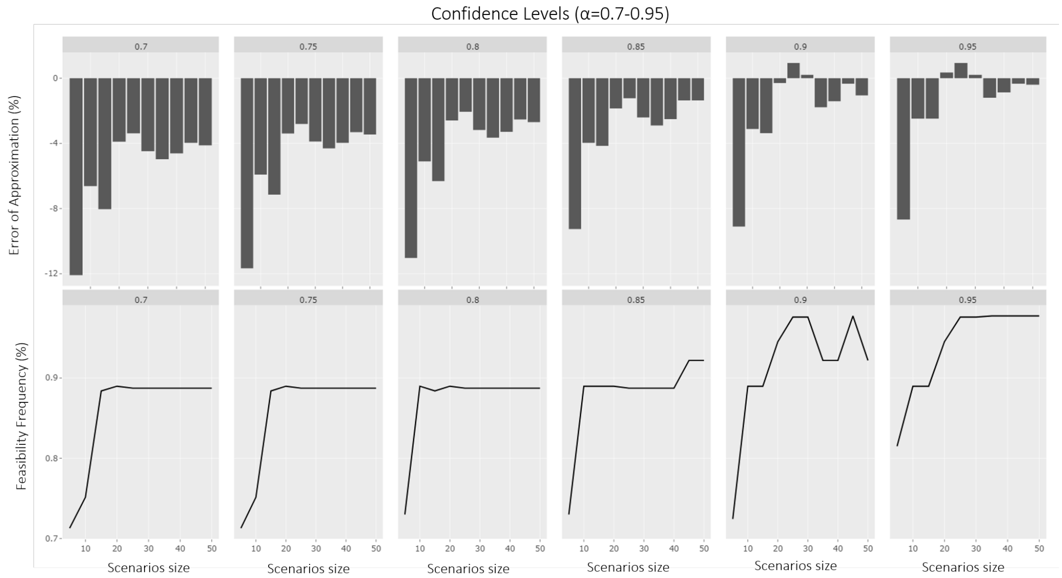


Figure 7.2: Scenario reduction metrics for different scenario sizes and confidence levels.

We can observe in Figure 7.2 that the performance of both metrics is related not only to the number of scenarios but also to the confidence level considered. A very small number of scenarios, such as 5, generates a bad representation of the uncertainty with large *Error of Approximation* (deviations of more than 8% between the predicted objective function and the actual objective function) and feasibility frequencies lower than their expected confidence levels. As the number of scenarios enlarges, the feasibility frequency also rises and approximates itself to its respective confidence level. A sample size with at least 20 scenarios is enough to respect the confidence level regardless if it is low (such as 0.7 and 0.75) or high (greater or equal to 0.9). Regarding the Error of Approximation, its values stays in an acceptable difference (lower than 5%) from 20 scenarios and so on.

However, the performance difference between these two scenario sizes (20 and 25) is very small and both levels respect the confidence levels. As we aspire to reduce the computational effort required for the models as much as possible without affecting the model performance, we have selected the scenario reduction size to 20, which is enough to respect the confidence levels considered and achieve a solution with low deviations from the distribution reality. This setting of generating 2,000 initial scenarios with the MCS and reducing to 20 scenarios with the Wasserstein distance-based algorithm will be used in the subsequent experiments for the regression-driven stochastic JCC-WRSP.

7.2.2 Stochastic JJC-WRSP numerical results

This section is dedicated to the results of the regression-driven stochastic JCC-WRSP model presented in Section 7.1.3. However, prior to presenting its results, it is important to illustrate the joint probabilistic constraints using the scenario-based approach with the slack variable $V_{ij}^{k\omega}$ and the binary variable Y^{ω}. Figure 7.3 shows examples of schedules with different values of $V_{ij}^{k\omega}$ and Y^{ω}.

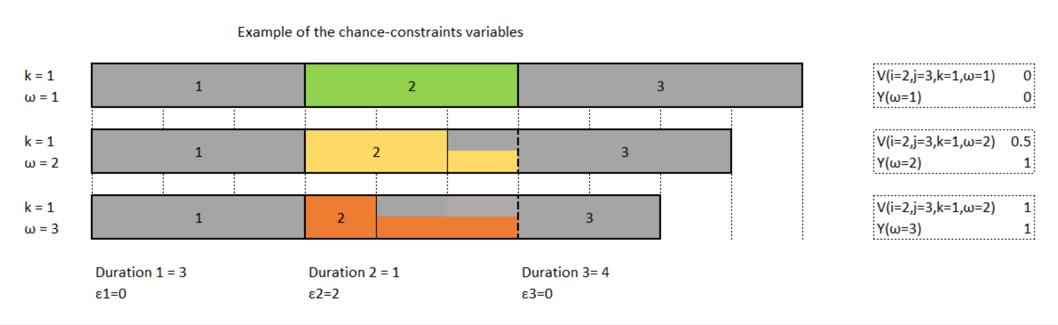


Figure 7.3: Examples of the scenario-based JCC reformulation for different values of the chance-constrained variables.

When no violation occurs in a scenario, both variables $V_{ij}^{k\omega}$ and Y^{ω} are equal to zero in that scenario. This is the case of the first row of Figure 7.3, when no task overlaps with the duration of the others. However, this is not always the case. In the second row, Task 2 overlaps with the start of task 3 with a violation of $\varepsilon_i^{k\omega} \times V_{ij}^{k\omega} = 2 \times 0.5 = 1$ day, making this scenario infeasible and Y^{ω} equal to one. Another case of an infeasible schedule is shown in the third row when both variables $V_{ij}^{k\omega}$ and Y^{ω} are equal to one: Well 2 overlaps the start of workover 3 in 2 days ($\varepsilon_i^{k\omega} \times V_{ij}^{k\omega} = 2 \times 1$).

Computational experiments were performed with instances considering 20 scenarios, a planning horizon of 360 days and a confidence level of 0.90. The size of the resulting models are presented on Table 7.1, which shows the size of the stochastic model and its solutions for different instances.

Setting		# Variables	# Integer Variables	# Binary Variables	# Constraints	# Non-zeros
Wells	Rigs					
25	10	71466	41526	40746	142375	407098
	20	169876	95896	95116	339475	981162
50	10	333541	179611	178081	665750	1950797
	20	578291	310661	309131	1155530	3390953

Table 7.1: Number of variables/constraints and size of the scenario-based JCC-WRSP via-regression model (before Gurobi presolve).

We can observe that all models have a large size, even though there are only 20 scenarios. Table 7.2 presents the solutions and computational time for each instance. Seven metrics are used to evaluate the model: average GAP in percentage; average solving time in seconds, the models were forced to have an execution time limit of one hour (3600 seconds); average objective function (OF) in millions (M); average expected value solution (EEV) in millions (M), which is the expected objective function fixing the first stage solution, simulating 500 scenarios, and solving the second-stage optimally; the approximation error (EOA) in millions (M) and in percentage (%), which is the difference between the estimated objective function and the EEV; feasibility frequency in percentage (%) of the second-stage simulating 500 scenarios.

Wells	Rigs	Confidence	GAP	Time	OF	EEV	EOA	Feasibility	EOA
		Level	(%)	(s)	(M)	(M)	(M)	Frequency (%)	(%)
10	2	0.7	0%	11	166.0	169.3	-3.2	84.2%	-1.9%
		0.8	0%	23	183.8	179.5	4.3	98.0%	2.3%
		0.9	0%	9	178.1	182.3	-4.2	98.7%	-2.4%
	3	0.7	0%	13	130.3	133.0	-2.7	91.0%	-2.1%
		0.8	0%	16	138.4	141.0	-2.5	91.5%	-1.8%
		0.9	0%	19	132.9	136.4	-3.5	99.5%	-2.6%
	5	0.7	0%	1	71.2	74.5	-3.3	98.9%	-4.7%
		0.8	0%	2	70.9	75.6	-4.8	96.8%	-6.7%
		0.9	0%	2	72.0	74.4	-2.4	96.8%	-3.3%
15	2	0.7	27%	3600	317.8	316.3	1.6	89.9%	0.5%
		0.8	21%	3600	312.4	318.1	-5.7	96.2%	-1.8%
		0.9	0%	712	309.3	310.1	-0.8	88.7%	-0.3%
	3	0.7	0%	481	224.0	219.8	4.2	96.3%	1.9%
		0.8	0%	105	227.2	231.9	-4.8	95.3%	-2.1%
		0.9	0%	174	233.3	228.5	4.8	98.0%	2.0%
	5	0.7	0%	67	118.7	121.6	-2.9	85.1%	-2.4%
		0.8	0%	43	124.0	123.1	1.0	95.6%	0.8%
		0.9	0%	34	122.1	122.1	0.0	94.1%	0.0%
25	2	0.7	39%	3600	626.5	627.9	-1.3	92.6%	-0.2%
		0.8	29%	3600	627.0	614.2	12.8	91.0%	2.0%
		0.9	16%	3600	616.2	610.1	6.0	88.9%	1.0%
	3	0.7	30%	3600	441.6	437.8	3.8	81.5%	0.9%
		0.8	30%	3600	447.1	444.1	3.0	90.3%	0.7%
		0.9	13%	3600	447.6	436.9	10.7	91.3%	2.4%
	5	0.7	8%	1938	264.8	264.5	0.3	84.8%	0.1%
		0.8	0%	596	272.9	272.4	0.6	94.3%	0.2%
		0.9	0%	257	268.5	272.4	-3.8	93.7%	-1.4%
50	2	0.7	60%	3600	1345.6	1344.6	0.9	89.6%	0.1%
		0.8	53%	3600	1346.2	1339.7	6.4	89.7%	0.5%
		0.9	53%	3600	1348.8	1342.0	6.8	84.2%	0.5%
	3	0.7	53%	3600	1173.8	1176.8	-3.0	86.8%	-0.3%
		0.8	50%	3600	1148.6	1163.4	-14.8	86.7%	-1.3%
		0.9	41%	3600	1147.5	1147.8	-0.2	78.6%	0.0%
	5	0.7	36%	3601	718.2	726.9	-8.8	74.6%	-1.2%
		0.8	30%	3600	718.8	725.0	-6.2	83.0%	-0.9%

Table 7.2 continued from previous page

Wells	Rigs	Confidence Level	GAP (%)	Time (s)	OF (M)	EEV (M)	EOA (M)	Feasibility Frequency (%)	EOA (%)
		0.9	27%	3600	713.7	718.0	-4.3	80.8%	-0.6%

Table 7.2: Computational results for the regression-driven stochastic JCC-WRSP.

Analyzing Table 7.2, the computational requirement seems to be related not only to the size of the model (number of wells and rigs) but also to the complexity of the instance. Apparently, the complexity of the instances varies according to the ratio of wells per rig (a low ratio is more straightforward to solve than a higher ratio) and the density of the release dates (low densities imply in concentrated release date on the end of the planning horizon and seems to be harder to solve than a high-density number).

The results from Table 7.2 also indicate that the model can generate solutions for instances with 20 scenarios in a feasible time. In the instances with at least 25 wells, the model had difficulty closing the gap but still was able to obtain optimal solutions in some cases. This number of scenarios is usually enough to obtain, through scenario reduction algorithms, a set of scenarios that accurately represents the uncertainty. Another instance behavior is that instances with a smaller number of rigs require more computational efforts to be solved than other instances with the same number of wells but more rigs available. Last, instances with a lower confidence level also seem to have more difficulty closing the gap.

As to the solution quality of the model, we can observe that the confidence level is respected in most instances, especially those solved close to optimally, and that instances with a larger confidence level usually have a higher feasibility frequency. We can also observe very low EOA (in percentage), which indicates that the estimated costs are also really close to the actual solution costs.

Nonetheless, there are other chance-constraints alternatives that can be tested in this problem. Next, section 7.2.3 compares these different data-driven stochastic formulations.

7.2.3 Comparison between stochastic formulations

This section compares the chance-constraints variations for the data-driven workover rig scheduling problem: the JCC (Section 7.1.3), the integrated chance-constrained (Section 7.1.3.1), and the budget-constrained (Section 7.1.3.2). These models were tested with instances with 20 scenarios, a planning horizon of 360 days, a release date density of 0.50, and two random seeds. Averages results (fifth to last columns) comparing the performance of each model (fourth column) according to the number of wells (first column), the number of rigs (second column), and the confidence level (third column) are presented on Table 7.3.

Wells	Rigs	Confidence Level	Chance Constraints	GAP (%)	Time (s)	OF (M)	EEV (M)	EOA (M)	Feasibility Frequency (%)	EOA (%)
15	2	0.7	JCC	25%	3600	297.4	292.1	5.3	92.6%	1.8%
			Integrated	12%	3600	273.5	258.9	14.6	59.7%	5.3%
			Budget	0%	366	235.0	237.2	-2.2	26.5%	-0.9%
		0.8	JCC	23%	3600	291.8	281.3	10.5	75.5%	3.6%
			Integrated	17%	3600	271.2	268.2	3.0	67.8%	1.1%
			Budget	0%	1151	232.8	258.9	-26.2	59.7%	-11.2%
		0.9	JCC	15%	3600	291.3	293.0	-1.7	88.5%	-0.6%
			Integrated	0%	3393	295.3	293.0	2.2	88.5%	0.8%
			Budget	17%	3600	257.2	269.4	-12.2	68.8%	-4.7%
	3	0.7	JCC	0%	22	234.9	234.8	0.1	94.4%	0.0%
			Integrated	0%	203	228.0	222.1	5.9	47.4%	2.6%
			Budget	0%	39	202.5	198.4	4.1	5.5%	2.0%
		0.8	JCC	0%	9	225.2	235.7	-10.5	94.7%	-4.7%
			Integrated	0%	166	224.1	227.3	-3.2	64.0%	-1.4%
			Budget	0%	102	200.2	219.4	-19.2	36.0%	-9.6%
		0.9	JCC	0%	11	230.2	233.9	-3.7	93.4%	-1.6%
			Integrat	0%	44	232.2	235.1	-2.9	94.3%	-1.3%
			Budget	0%	130	208.5	212.2	-3.6	16.5%	-1.7%
25	2	0.7	JCC	34%	3604	602.7	625.4	-22.7	87.1%	-3.8%
			Integrated	36%	3600	567.8	562.7	5.2	39.3%	0.9%
			Budget	33%	3601	513.4	532.8	-19.4	13.7%	-3.8%
		0.8	JCC	29%	3600	629.7	635.1	-5.4	98.9%	-0.9%
			Integrated	38%	3600	602.8	601.7	1.2	66.7%	0.2%
			Budget	36%	3601	555.7	572.0	-16.3	36.9%	-2.9%
		0.9	JCC	28%	3600	641.6	632.9	8.6	88.5%	1.3%
			Integrated	32%	3602	618.8	613.9	4.9	89.2%	0.8%
			Budget	35%	3600	559.4	570.8	-11.4	37.8%	-2.0%
	3	0.7	JCC	32%	3600	408.3	423.6	-15.3	88.3%	-3.7%
			Integrated	27%	3600	385.6	402.9	-17.2	62.1%	-4.5%
			Budget	17%	3601	335.4	343.1	-7.7	7.3%	-2.3%
		0.8	JCC	24%	3600	406.7	409.3	-2.6	87.7%	-0.6%
			Integrated	25%	3600	392.4	396.0	-3.7	63.7%	-0.9%
			Budget	16%	3600	344.4	349.2	-4.8	7.6%	-1.4%
		0.9	JCC	12%	3600	400.0	410.3	-10.4	89.9%	-2.6%
			Integrated	17%	3600	396.1	407.8	-11.6	88.4%	-2.9%
			Budget	19%	3601	344.2	363.2	-19.0	20.3%	-5.5%

Table 7.3: Results comparing the regression-driven chance-constraints variations.

Analyzing Table 7.3, we can observe that despite being slightly more costly in terms of computational effort, the data-driven JCC model tends to obtain solutions with lower approximation error and greater feasibility frequency than the other stochastic formulations, being superior in 9 out of 12 instances and having similar results in the remaining. The integrated chance-constrained model achieves inferior results to the JCC in 8 of 12 instances, violates the confidence level set in 11 instances, and requires computational efforts similar to the JCC. On the other hand, the budget-constrained model is usually easier to be solved but often has a larger EOA. It is also much more challenging to set a good confidence level, as it does not consider any probability measure in its modeling, which leads to extremely low feasibility frequency in comparison with the other formulations. Aiming to support this comparison, Figure 7.4 separates the average EOA in percentage (y-axis) and the feasibility frequency in percentage (x-axis) for each model according to the confidence level (facet charts):

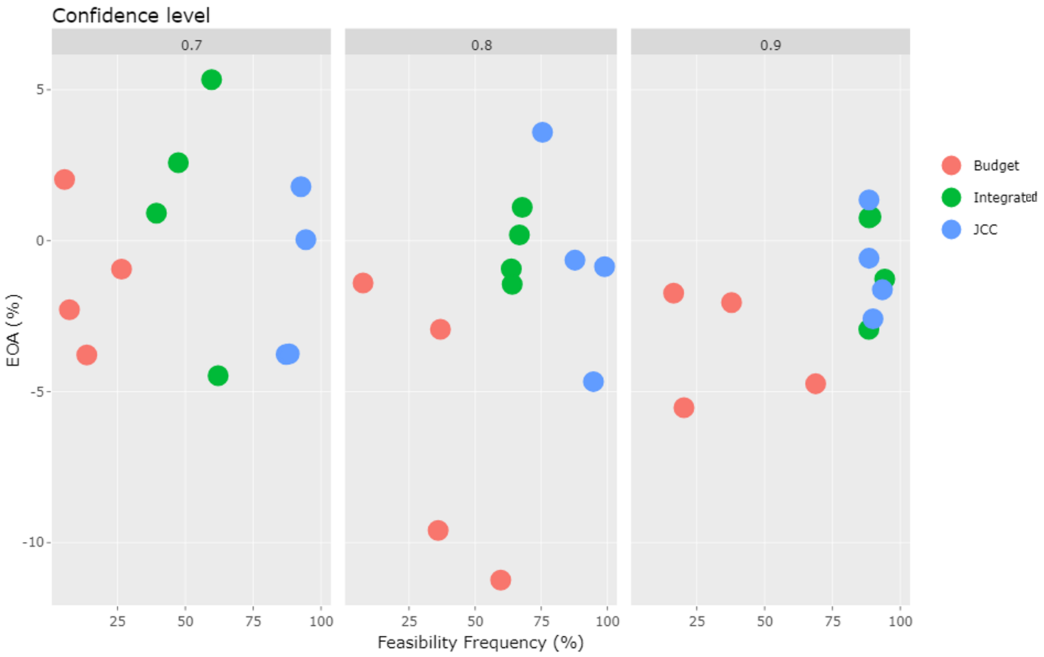


Figure 7.4: Performance comparison of the difference regression-driven chance-constrained approaches.

We can observe in Figure 7.4 that the JCC (blue labels) has much less variability than the others models in terms of EOA and feasibility frequency. The least consistent results were for the budget-constrained model (red labels), with the average feasibility frequency varying from 10% to 70% and the average EOA between -12% and 3%. These inconsistent results were expected as this model does not consider the probability in its constraints nor has a straightforward confidence level parameter, compromising its feasibility level or low EOA guarantees. On the contrary, the integrated CC (green labels) has a risk measure in its constraints but performs poorly in low confidence levels and has a less consistent approximation error than the JCC model. This performance was expected as the integrated does not consider any implicit measure of feasibility frequency of the JCC, focusing on the expected infeasibility and not necessarily it's chance.

In summary, these results confirm the reliability and robustness of data-driven JCC model and support it as our primary choice for the proposed regression-driven optimization methodology. The following section compares the proposed regression-driven JCC with the non-stochastic regression-driven model and the current company methodology.

7.2.4 The trade-off between stochastic and non-stochastic optimization

Last, the selected stochastic regression-driven approach with JCCs is to be compared with the non-stochastic regression-driven optimization and the company's methodology that uses plans for the workover duration with average estimations according to the workover group, enabling the assessment of the gains of employing the final proposed regression-driven optimization methodology. These stochastic models were tested in instances with 20 scenarios (stochastic model only), a planning horizon of 360 days, a release date density of 0.50, and two random seeds.

Table 7.4 compares the three models (first column: "Plan" for the company current methodology that uses the workover conservative duration to estimate a rig schedule; "Non-stoc." for the data-driven optimization model proposed in Chapter 6; "JCC" for the final stochastic regression-driven JCC model selected in the previous Section) according the number of wells (15 or 25, second column) and the number of rigs (2 or 3, third column) for the metrics: GAP, solving time, OF, EEV, EOA, and feasibility frequency in percentage (%) of the second-stage simulating 500 scenarios.

Model	Wells	Rigs	GAP	Time	OF	EEV	EOA	Feasibility	EOA
			(%)	(s)	(M)	(M)	(M)	Frequency (%)	(%)
Plan	15	2	0%	1	435.2	344.2	91.1	100%	21%
		3	0%	0	363.8	259.7	104.2	100%	29%
	25	2	0%	18	797.1	687.5	109.6	100%	14%
		3	0%	5	651.9	501.6	150.3	100%	23%
Non-stoc.	15	2	0%	17	232.2	237.2	-5.1	26%	-2%
		3	0%	1	196.4	213.1	-16.6	32%	-8%
	25	2	27%	3600	537.4	547.1	-9.7	11%	-2%
		3	0%	3152	344.9	346.6	-1.8	3%	-1%
JCC	15	2	21%	3600	293.5	288.8	4.7	86%	2%
		3	0%	14	230.1	234.8	-4.7	94%	-2%

Table 7.4 continued from previous page

Model	Wells	Rigs	GAP	Time	OF	EEV	EOA	Feasibility	EOA
			(%)	(s)	(M)	(M)	(M)	Frequency (%)	(%)
	25	2	30%	3602	624.7	631.2	-6.5	92%	-1%
		3	23%	3600	405.0	414.4	-9.4	89%	-2%

Table 7.4: Results comparing the stochastic and non-stochastic approaches.

The results from Table 7.4 reinforce the power of the proposed regression-driven methodology in generating models with less approximation error. By using the average workover group duration in a deterministic model, the company original methodology results in extremely conservative solutions, leading to an inefficient solution and a significant error of approximation. Alternatively, the non-stochastic regression-driven optimization model does not have a reliable feasibility frequency when simulated. The only alternative that achieved simultaneously high feasibility levels and low errors of approximation was the proposed stochastic regression-driven JCC. Even though it requires more computational effort, it is much less conservative than the company's current approach, with solutions close to reality and guaranteeing a good confidence level for the solution's feasibility. The comparison of the stochastic and non-stochastic models can easy be seen in Figure 7.5, where the labels are the model strategy (red for the stochastic regression-driven JCC, green for the company's current approach, and blue for the non-stochastic regression-driven), the y-axis is the error of approximating (EOA) in percentage (%), and the x-axis is the feasibility frequency, also in percentage (%). Note that the ideal spot of the model result is the bottom-right (EOA and feasibility frequency close to 0% and 100%, respectively).

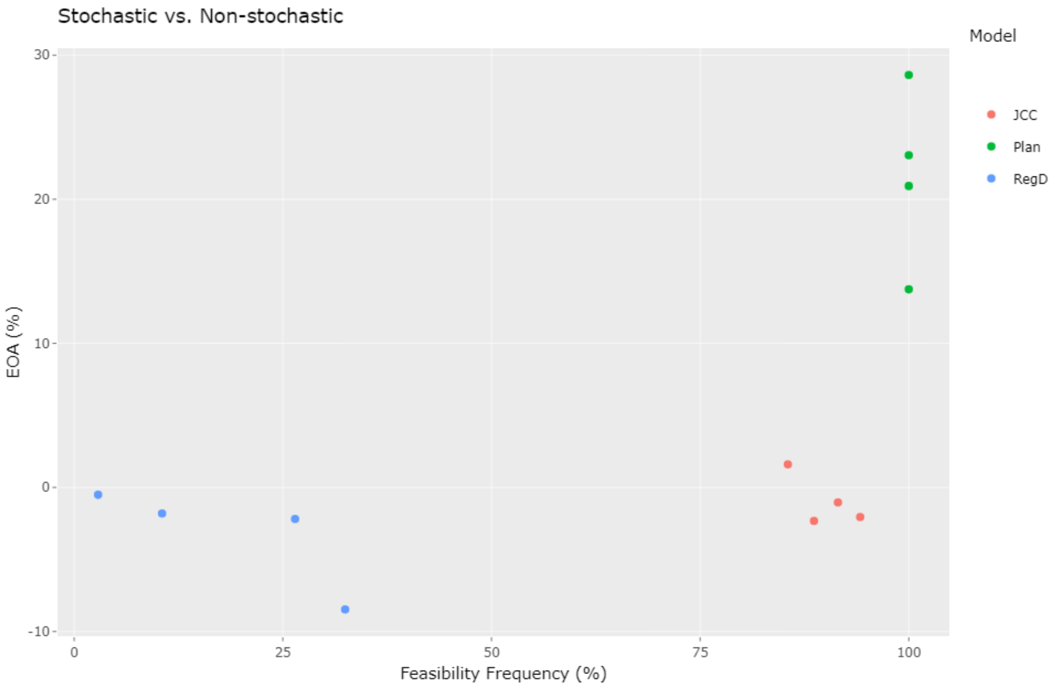


Figure 7.5: Performance comparison of the stochastic and non-stochastic regression-driven models and current company approach.

Analyzing Figure 7.5, we can confirm the superiority of the regression-driven stochastic JCC model. Meanwhile, the results of the non-stochastic models are outside the 'ideal zone', with high approximation errors or high rate of infeasible solutions, the results of the stochastic regression-driven JCC models are concentrated in the 'ideal zone', with EOA and feasibility frequency close to 0% and 100%, respectively. Moreover, the non-stochastic results are scattered and seem much more variable than the proposed stochastic model. Not only the proposed stochastic regression-driven JCC model is accurate but also precise and outperforms the others models that writhe with a lack of precision and reliability.

These results reinforce the importance of considering uncertainty in the mathematical modeling. The data-driven methodology is extremely helpful in obtaining models more accurate to reality. However, underneath any data-driven approach, there is still uncertainty related to the data models and other endogenous uncertainties. These uncertainties affect the precision of the model. Therefore, considering the uncertainty of the data-driven optimization model is crucial for obtaining accurate and precise data-driven models. This data-driven optimization paradigm was also mentioned by Ning and You (2019) in their review of data-driven optimization under uncertainty studies.

Some final considerations, suggestions for future studies and other future developments are discussed in the Section 8.

8
Final considerations

Oil rigs are expensive and scarce resources used in the exploration and production of oil & gas. They are crucial for several wells operations, such as workover, which is an intervention in the wells' operations for maintenance or inspection. The proper planning of these operations is critical for the wells' profitability due to the high costs associated with rigs and oil production losses of an inactive well waiting for a workover. Moreover, the workover operations occur in an environment surrounded by uncertainties, such as whether workovers will be needed and their durations. In this environment, the workover rig scheduling problem (WRSP) emerges as the problem of selecting a fleet of workover rigs and a set of workover wells to be served, allocating the wells to the rigs, and scheduling the operations, while minimizing the costs associated with the rigs hiring and the oil production loss of the wells waiting for service. The WRSP is a particular case of the rig scheduling problem (RSP). Both problems are extremely important for oil companies and have attracted a lot of attention from the literature, motivating *RQ 1*, which addresses trends and gaps that can be observed in the RSP literature from academic and industrial perspectives.

Despite the risky environment in which the WRSP and the RSP are inserted, few studies applying optimization under uncertainty or data-driven optimization were detected in the systematic literature review (SLR) from Chapter 2. Another opportunity noticed in the SLR was the gap between academic research and industry needs, with insufficient studies implemented or validated in the industry. Many advanced optimization techniques are available in the literature and should be applied in the WRSP with realistic assumptions, such as a heterogeneous fleet of rigs, multi-objective functions, and uncertain parameters. As a result of the aforementioned gaps detected in answering *RQ 1*, we posed *RQ 2* of how the optimization under uncertainty could be applied in WRSP to mitigate any uncertainty-related issues. The SLR and RSP classifications that were presented in Chapter 2 resulted in the publishing of an article "A Systematic Literature review for the rig scheduling problem: Classification and state-of-the-art" in the journal Computers & Chemical Engineering (Santos et al. 2021). Aiming to answer *RQ 2* and fulfill the literature gaps, this thesis addresses real-world-based instances of the WRSP and proposes a data-driven optimization methodology with clustering algorithms, ridge regression, stochastic programming, and chance-constrained optimization.

This proposed methodology was developed after reviewing the state-of-art of the data-driven optimization under uncertainty in Chapter 3, in which a lack of linear reformulations for data-driven joint chance-constrained models was detected. As explained in Chapter 4, the methodology is divided into three major phases: data treatment, predictive modeling, and optimization. The data treatment uses text mining and clustering techniques to refine and retrieve information from the data. In the predictive modeling, ridge regression and generalized linear models are used to estimate not only the workover duration but also the endogenous uncertainties in the model. Last, in the optimization phase, the regression prediction and error are inserted in a scenario-based joint chance-constrained (JCC) model, generating solutions more resilient to uncertainties.

One of the main contributions of this thesis was applying the data-driven optimization methodology in a WRSP with real-world-based instances. Chapter 5 focused on the assumptions of the WRSP from the offshore oil industry perspective, motivating *RQ 3*. The case study is an offshore workover rig scheduling problem in which the goal is to select the wells requiring workover that will be served and an optimal fleet of heterogeneous rigs to be hired and to schedule the wells to the rigs, minimizing the oil production losses associated with the well waiting for service and the hiring costs of the rigs' operation, respecting the machine eligibility and workover release date. An innovative assumption of this WRSP is the uncertainty in the intervention durations of the wells. These durations are difficult to estimate and depend not only on the well's characteristics but also on the rig that performs the operation, and other unknown factors, such as weather conditions. The current methodology of the studied company is a deterministic approach using conservative estimation according to the workover type that is required. However, the company stores past information on the wells, rigs, and workover operations that is possibly useful in the workover planning. These records were gathered for this study and several real-world-based instances were created based on them.

In the first part of the work, the past workover dataset was analyzed following rigorous data science practice. Text mining (stemming, stop-words, and string similarity measures) and clustering algorithms (k-means) were used to treat the data by cleaning, simplifying, and labeling the workover historical data. This treated data was used as input for the training of the GLM and ridge regression models, which estimate the duration. Text mining and clustering procedures proved to be an efficient way of labeling the historical data and acquiring hidden information, discovering new features to be used in the predictive models. Overall, combining these data science techniques with the regression model improved the prediction of the workover duration, which is currently poorly estimated by the studied company.

Later, in Chapter 6, the workover duration estimations were used in the linear programming model for the WRSP, following the proposed data-driven optimization methodology and tackling *RQ 4*, which queries how we can model the WRSP for a practical case. 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 15% and, therefore, requires considerably less rescheduling. Meanwhile, the current approach employed by the company usually has deviations of 20 to 120%, requiring more frequent rescheduling. These results indicate 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. However, 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 was suitable for the problem. Nonetheless, the uncertain nature of the regression is an important feature to be considered as it can represent other exogenous uncertainties that are incredibly complex to estimate and that might affect the solution's feasibility. This regression-driven optimization methodology proposed in Chapter 6 resulted in the publishing of an article "A data-driven optimization model for the workover rig scheduling problem: Case study in an oil company" in the journal Computers & Chemical Engineering (Santos et al. 2023).

Aiming to hedge against these potential losses and minimize their risk to an acceptable level, a better approach with a complete analysis of the data and considering the uncertainty in the workover operations is crucial. A chance-constrained formulation was proposed in Chapter 7 for the WRSP to be used in our regression-driven optimization methodology. Usually, joint chance constraints result in non-linear programming models. Based on the stochastic reformulations of Luedtke and Ahmed (2008), Nemirovski and Shapiro

(2007), and Nikzad et al. (2019), a regression-driven stochastic joint chance-constrained model was proposed for the WRSP. In that, the regression errors are used as an uncertain parameter in the mathematical model. A Normal distribution is estimated for the regression error and used in a Monte Carlo simulation to generate scenarios that are reduced to a practical sample size with a Wasserstein distance-based method. This regression-driven stochastic JCC model allowed us to tackle *RQ 5*, which inquires how we can combine chance-constrained models and data-driven optimization to solve large and realistic problems such as the WRSP. The regression-driven stochastic JCC model was compared not only with the company's current approach, but also with other variations of the regression-driven methodology, such as integrated CC, a budget-constrained, and the model from Chapter 6. Meanwhile, the proposed regression-driven methodology achieves solutions with a much lower error of approximation, being more reliable to achieve efficient schedules. As to the different formulation alternatives for the regression-driven methodology, the data-driven stochastic JCC model outperformed the other models, obtaining solutions that were reliable and robust, not only with lower expected objective functions and low error of approximating (between -5% and 5%) but also a guaranteed feasibility level. Nonetheless, the regression-driven stochastic JCC model requires more computational effort than the regression-driven model and, as a result, has more difficulty closing optimally gaps in larger instances. Therefore, we recommend using the regression-driven methodology with the stochastic JCC model for small and medium instances and, if needed, in large instances, the non-stochastic data-driven model is a good alternative with disturbances not greater than 10%. Next, we suggest future studies related to this thesis, its methodology, and the RSP literature.

8.1 Future research and suggestions

Data-driven optimization under uncertainty is a new trend in the Operations Research community with a lot of potential. The regression-driven methodology enhanced the workover estimations considerably. Nonetheless, future improvements could still be possible. For instance, the R^2 values of the selected predictor are still in the order of 0.5, which ensues further efforts in improving the prediction accuracy. This improvement can be achieved by, e.g., investigating additional feature engineering strategies. Furthermore, alternative prediction methods can be tested and compared with the current data-driven methodology, such as LASSO regression, gradient-boosted trees, random forest, or support vector clustering.

As to the optimization phase, new strategies of joint chance-constraint formulations and stochastic programming could be tested. For instance, the non-linear joint chance-constrained representation (Section 4.1.2.1) could be treated using a Branch-price-and-cut solving formulation for MINLP. More comparison of the regression-driven stochastic JCC method with others data-driven joint chance-constrained models, such as Calfa et al. (2015)'s kernel-based one, also needs to be tested. Kernel density estimations can also be used as a distribution for the scenario generation, simulating the regression residuals or the workover duration. Another possibility is to evaluate the impact of the choice of the type of optimization under uncertainty, comparing the joint chance-constrained WRSP with a stochastic programming WRSP or a robust programming WRSP. Further experiments also can be developed comparing different stochastic programming approaches, such as sample average approximation and others simulation and scenario reduction techniques.

Another optimization perspective would be to adapt the data-driven optimization methodology to approximate algorithms, such as metaheuristics, simheuristics, and matheuristics, instead of exact methods (mathematical programming). These approaches have the potential of reducing the computational effort considerably, enabling the consideration of more complex assumptions in the problem.

The WRSP is highly complex decision-making problem subjected to several risks and sharing ties with other resources and decisions of the petroleum E&P. This thesis approached a problem considering uncertainty in the workover duration, which is one of the most significant uncertainties affecting the workover planning. Nonetheless, other uncertainties could also be considered in the problem, such as the occurrence of the workover in a well, known as the dynamic-WRSP, its release date, the fleet availability can also be subjected to disruptions, weather conditions affecting the operation, the costs associated with oil production loss, and the rigs hiring costs.

Furthermore, new assumptions can be incorporated into the modeling, such as due dates, fleet availability, objective functions with net present values, different types of rig hiring contracts, learning curves for the workover duration in a rig, and travel times between the wells, possibly varying according to the rig type. As seen in the SLR, learning curves are an important feature from the industry perspective that has received very low attention in the literature due to its complexity. Others assets should also be integrated with the WRSP decision-making from a resource-planning perspective. For instance: the crew availability or scheduling; equipment availability; offshore support vessels usage. As mentioned in the SLR of RSP, to close the gap between the industry demands and the academy studies, it is essential to consider realistic assumptions and integrate the RSP decisions.

Last, the proposed data-driven methodology could be applied to similar scheduling problems, especially problems with sufficient amounts of historical data for the use of predictive models.

Bibliography

Abu-Marrul, V., Martinelli, R., and Hamacher, S. (2020). Scheduling pipe laying support vessels with non-anticipatory family setup times and intersections between sets of operations. *International Journal of Production Research*, 0(0):1–15.

Accioly, R. M. S., Marcellino, F. J. M., and Kobayashi, H. K. (2002). Uma aplicação da programação por restrições no escalonamento de atividades em poços de petróleo. In *Proceedings of the XXXIV Brazilian Symposium on Operations Research*, pages 1–10, Rio de Janeiro.

Achkar, V. G., Cafaro, V. G., Méndez, C. A., and Cafaro, D. C. (2019a). A discrete-time MILP formulation for the optimal scheduling of maintenance tasks on oil and gas wells and surface facilities. *Computer Aided Chemical Engineering*, 46:727–732.

Achkar, V. G., Cafaro, V. G., Méndez, C. A., and Cafaro, D. C. (2019b). Discrete-time milp formulation for the optimal scheduling of maintenance tasks on oil and gas production assets. *Industrial & Engineering Chemistry Research*, 58(19):8231–8245.

Agarwal, M., Sharma, R., and Mathew, L. (2016). Challenges in supply chain management in upstream sector of oil and gas industry. In *Agro Supply Chain Conference (ASCC)*, pages 1–17, University of Petroleum & Energy Studies, Dehradun.

Ahmed, S. and Shapiro, A. (2008). *Solving Chance-Constrained Stochastic Programs via Sampling and Integer Programming*, chapter Chapter 12, pages 261–269.

Al-Azani, K. H. (2014). Drilling rigs structure: a comparison between onshore and offshore drilling rigs with proposed future developments. Report, King Fahd University of Petroleum and Minerals, Saudi Arabia.

Al Gharbi, S. H. (2011). Drilling rig schedule optimization. Master's thesis, King Fahd University of Petroleum and Minerals, Saudi Arabia.

Aloise, D., de Noronha, T. F., Maia, R. S., Bittencourt, V. G., and Aloise, D. J. (2002). Heurísticas de colônia de formigas com path-relinking para o problema de otimização da alocação de sondas de produção terrestre–spt. In *Proceedings of the XXXIV Brazilian Symposium on Operations Research*, Rio de Janeiro.

Aloise, D. J., Aloise, D., Rocha, C. T. M., Ribeiro, C. C., Ribeiro Filho, J. C., and Moura, L. S. S. (2006). Scheduling workover rigs for onshore oil production. *Discrete Applied Mathematics*, 154:695–702.

Alves, V. R. F. M. and Ferreira Filho, V. J. M. (2006). Proposta de algoritmo genético para a solução do problema de roteamento e sequenciamento de sondas de manutenção. In *Proceedings of the XXXVIII Brazilian Symposium on Operations Research*, pages 1837–1848, Goiânia, Brazil.

Amer, M., Ilya, I., Ali, A., Amir, A., and Muhaisen, A. (2016). A fair and competitive drilling and workover services assignment process. In *IADC/SPE Asia Pacific Drilling Technology Conference*, Singapore.

Amrideswaran, H., Ali Al-Sada, A., and Mohammed, A. A. M. S. (2015). Risk assessment suite - an innovative approach of risk mitigation measures prioritization for well integrity assurance. In *Proceedings of the SPE Middle East Oil and Gas Show and Conference, MEOS*, pages 675–692.

Arnaout, A., Heber, M., Wolf-Zoellner, P., and Thonhauser, G. (2017). A new method for rig move optimization - case study on moving land rigs. *Oil Gas European Magazine*, 43(1):28–30.

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

Aronofsky, J. S. (1962). Linear programming a problem-solving tool for petroleum industry management. *Journal of Petroleum Technology*, 14(7):729–736.

Aseeri, A., Gorman, P., and Bagajewicz, M. J. (2004). Financial risk management in offshore oil infrastructure planning and scheduling. *Industrial and Engineering Chemistry Research*, 43(12):3063–3072.

Attia, A. M., Ghaithan, A. M., and Duffuaa, S. O. (2019). A multi-objective optimization model for tactical planning of upstream oil & gas supply chains. *Computers & Chemical Engineering*, 128:216–227.

Aurachman, R., Ajiguno, T., and Putri, E. M. (2020). Well maintenance scheduling using dynamic programming approach: influence diagram. *Journal of Physics: Conference Series*, 1477:052024.

Baker, R. (1996). *A primer of oilwell drilling: a basic text of oil and gas drilling*. Petroleum Extension Service, 4th edition.

Bakker, S., Aarlott, M., Tomasgard, A., and Midthun, K. (2017). Planning of an offshore well plugging campaign: A vehicle routing approach. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10572:158–173.

Bakker, S., Vrålstad, T., and Tomasgard, A. (2019). An optimization model for the planning of offshore plug and abandonment campaigns. *Journal of Petroleum Science and Engineering*, 180:369–379.

Bakker, S. J. (2020). *Optimization models for the plugging and abandoning of offshore oil and gas fields*. Phd thesis, NTNU, Norwegian University of Science and Technology, Trondheim, Norway.

Bakker, S. J., Wang, A., and Gounaris, C. E. (2021). Vehicle routing with endogenous learning: Application to offshore plug and abandonment campaign planning. *European Journal of Operational Research*, 289(1):93–106.

Barnes, J. W., Brennan, J. J., and Knap, R. M. (1977). Scheduling a backlog of oilwell workovers. *Journal of Petroleum Technology*, 29:1651–1653.

Barnes, R. J. and Kokossis, A. (2007). A mathematical programming approach to the analysis, design and scheduling of offshore oilfields. *Computer Aided Chemical Engineering*, 24:503–508.

Bassi, H. V. (2010). Simulação-otimização e reconexão por caminhos aplicadas ao gerenciamento de sondas de intervenção. Master's thesis, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.

Bassi, H. V., Ferreira Filho, V. J. M., and Bahiense, L. (2012). Planning and scheduling a fleet of rigs using simulation-optimization. *Computers & Industrial Engineering*, 63:1074–1088.

Ben-Tal, A., Bhadra, S., Bhattacharyya, C., and Saketha Nath, J. (2011). Chance constrained uncertain classification via robust optimization. *Mathematical programming*, 127(1):145–173.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust optimization*, volume 28. Princeton university press.

Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 5(Sep):1089–1105.

Bertsimas, D., Gupta, V., and Kallus, N. (2018). Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292.

Bertsimas, D. and Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044.

Bertsimas, D. and McCord, C. (2018). Optimization over continuous and multi-dimensional decisions with observational data. *Advances in neural information processing systems*, 31.

Bertsimas, D. and Sim, M. (2004). The price of robustness. *Operations research*, 52(1):35–53.

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

Bissoli, D. C. (2014). Uma abordagem heurística para o problema de roteamento de sondas de intervenção bi-objetivo. Master's thesis, Universidade Federal do Espírito Santo, Vitória, Brazil.

Bissoli, D. C., Chaves, G. L. D., and Ribeiro, G. M. (2016). Drivers to the workover rig problem. *Journal of Petroleum Science and Engineering*, 139:13–22.

Bissoli, D. C., Vieria, B. S., Chaves, G. L. D., and Ribeiro, G. M. (2014). Um ALNS para o problema de roteamento de sondas de intervenção bi-objetivo. In *Proceedings of the XLVI Brazilian Symposium on Operations Research*, pages 1218–1230, Salvador, Brazil.

Biswal, M., Sahoo, N., and Li, D. (2005). Probabilistic linearly constrained programming problems with lognormal random variables. *OPSEARCH*, 42.

Bourgoyne, A., Millheim, K., Chenevert, M., and Young, F. (2016). *Applied drilling engineering*. Society of Petroleum Engineers, second edition.

Calafiore, G. and Campi, M. C. (2005). Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, 102(1):25–46.

Calafiore, G. C. and Campi, M. C. (2006). The scenario approach to robust control design. *IEEE Transactions on automatic control*, 51(5):742–753.

Calafiore, G. C. and Ghaoui, L. E. (2006). On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130(1):1–22.

Calderón, A. J. and Pekney, N. J. (2020). Optimization of enhanced oil recovery operations in unconventional reservoirs. *Applied Energy*, 258:114072.

Calfa, B., Grossmann, I., Agarwal, A., Bury, S., and Wassick, J. (2015). Data-driven individual and joint chance-constrained optimization via kernel smoothing. *Computers & Chemical Engineering*, 78:51–69.

Carrilho, L. M., Andrade, T., Ribas, G., and Hamacher, S. (2018). Dimensionamento e sequenciamento de sondas usando a formulação bucket-indexed. In *Proceedings of the XLIII Brazilian Symposium on Operations Research*, Rio de Janeiro, Brazil.

Carrilho, L. M. and Villas Boas, M. A. C. (2016). *CORE: decision support system for oil rig scheduling*. Undergraduate final project, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brazil.

Carvalho, M. C. A. and Pinto, J. M. (2006). An MILP model and solution technique for the planning of infrastructure in offshore oilfields. *Journal of Petroleum Science and Engineering*, 51(1-2):97–110.

Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. P., Basto, J. P., and Alcalá, S. G. S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137:106024.

Castiñeira, D., Toronyi, R., and Saleri, N. (2018). Machine learning and natural language processing for automated analysis of drilling and completion data. In *SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition (SATS)*, Saudi Arabia.

Chaari, T., Chaabane, S., Aissani, N., and Trentesaux, D. (2014). Scheduling under uncertainty: Survey and research directions. In *2014 International Conference on Advanced Logistics and Transport*, pages 229–234.

Charnes, A. and Cooper, W. W. (1959). Chance-constrained programming. *Management Science*, 6(1):73–79.

Charnes, A. and Cooper, W. W. (1963). Deterministic equivalents for optimizing and satisficing under chance constraints. *Operations Research*, 11(1):18–39.

Charnes, A., Cooper, W. W., and Symonds, G. H. (1958). Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil. *Management Science*, 4(3):235–263.

Chen, Y., Guo, Q., Sun, H., Li, Z., Wu, W., and Li, Z. (2018a). A distributionally robust optimization model for unit commitment based on kullback–leibler divergence. *IEEE Transactions on Power Systems*, 33(5):5147–5160.

Chen, Z., Peng, S., and Liu, J. (2018b). Data-driven robust chance constrained problems: a mixture model approach. *Journal of Optimization Theory and Applications*, 179(3):1065–1085.

Chowdhury, S. (2016). Optimizing rig move time and activity schedule using critical path analysis. In Chowdhury, S., editor, *Optimization and Business Improvement Studies in Upstream Oil and Gas Industry*, pages 137–166.

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

Cong, Z., Bakshi, A., Prasanna, V., Da Sie, W., and Bourgeois, B. (2008). A framework for design space exploration in oilfield asset development. In *Proceedings of the Intelligent Energy Conference and Exhibition*, pages 1155–1161, Amsterdam, Netherlands.

Costa, L. R. (2005). Soluções para o problema de otimização de itinerário de sondas. Master's thesis, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.

Costa, L. R. and 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*, page 1844–1853, São João del Rei, Brazil.

Costa, L. R. and 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*, pages 2176–2187, Gramado, Brazil.

Cui, X., Zhu, S., Sun, X., and Li, D. (2013). Nonlinear portfolio selection using approximate parametric value-at-risk. *Journal of Banking & Finance*, 37(6):2124–2139.

Currie, J. C., Novotnak, J. F., Aasboee, B. T., and Kennedy, C. J. (1997a). Optimized reservoir management using mixed linear programming. In *Proceedings of the Hydrocarbon Economics and Evaluation Symposium*, pages 235–241, Dallas, United States of America.

Currie, J. C., Novotnak, J. F., Aasboee, B. T., and Kennedy, C. J. (1997b). *Optimized reservoir management with mixed linear programming*. Number 9.

Danach, K. (2016). *Hyperheuristics in logistics.* PhD thesis, Ecole Centrale de Lille, Lille, France.

Davidson, J. E., Yang, Y., Netemeyer, S., Banki, A., and Do, L. (2009). An object-oriented approach to goal-based well management. In *Proceedings of the SPE Reservoir Simulation Symposium*, volume 1, pages 521–531, The Woodlands, United States of America.

de Andrade Filho, A. C. B. (1994). Optimal scheduling of development in an oil field. Master's thesis, Stanford University, Stanford, USA.

Dentcheva, D. and Martinez, G. (2013). Regularization methods for optimization problems with probabilistic constraints. *Mathematical Programming*, 138(1):223–251.

Dentcheva, D., Prékopa, A., and Ruszczynski, A. (2000). Concavity and efficient points of discrete distributions in probabilistic programming. *Mathematical programming*, 89(1):55–77.

Desai, J. N., Pandian, S., and Vij, R. K. (2020). Big data analytics in upstream oil and gas industries for sustainable exploration and development: A review. *Environmental Technology & Innovation*, page 101186.

Dewan, M. A. A., Lin, F., and Kinshuk (2016). Dynamic pricing mechanism for multi-agent based system of well scheduling. In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 1104–1108.

Diwekar, U. (2008). Introduction to applied optimization.

Douro, R. F. and Lorenzoni, L. L. (2009). Um algoritmo genético-2opt aplicado ao problema de otimização de itinerário de sondas de produção terrestre. In *Proceedings of the XLI Brazilian Symposium on Operations Research*, page 2121–2132, Porto Seguro, Brazil.

Drouven, M. G. and Grossmann, I. E. (2016). Multi-period planning, design, and strategic models for long-term, quality-sensitive shale gas development. *AIChE Journal*, 62(7):2296–2323.

Duhamel, C., Santos, A. C., and Guedes, L. M. (2012). Models and hybrid methods for the onshore wells maintenance problem. *Computers and Operations Research*, 39(12):2944–2953.

Dupačová, J., Gröwe-Kuska, N., and Römisch, W. (2003). Scenario reduction in stochastic programming. *Mathematical programming*, 95(3):493–511.

D'Ambrosio, C., Frangioni, A., Liberti, L., and Lodi, A. (2012). A storm of feasibility pumps for nonconvex minlp. *Mathematical programming*, 136(2):375–402.

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

Eksioglu, B., Vural, A. V., and Reisman, A. (2009). The vehicle routing problem: a taxonomic review. *Computers & Industrial Engineering*, 57(4):1472–1483.

Engebretsen, S. and Bohlin, J. (2019). Statistical predictions with glmnet. *Clinical epigenetics*, 11(1):1–3.

Falex, A. (2009). *Planejamento da frota de sondas para atendimento de uma campanha de perfuração de um campo*. Undergraduate final project, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.

Fernández Pérez, M. A., Oliveira, F., and Hamacher, S. (2018). Optimizing workover rig fleet sizing and scheduling using deterministic and stochastic programming models. *Industrial & Engineering Chemistry Research*, 57(22):7544–7554.

Flager, F. (2014). A method to optimize onshore drilling rig fleet size and schedule considering both reservoir management and operational objectives. *Journal of Project Production Management*, 1.

Fu, X., Guo, Q., Sun, H., Pan, Z., Xiong, W., and Wang, L. (2017). Typical scenario set generation algorithm for an integrated energy system based on the wasserstein distance metric. *Energy*, 135:153–70.

Georgiadis, G. P., Elekidis, A. P., and Georgiadis, M. C. (2019). Optimization-based scheduling for the process industries: from theory to real-life industrial applications. *Processes*, 7(7). 438.

Glinz, I. and Berumen, L. (2009). Optimization model for an oil well drilling program: Mexico case. *Oil and Gas Business*.

Gonçalves, R. K. (2009). *Otimização de alocação de sondas para exploração off-shore de hidrocarbonetos por algoritmos genéticos*. Specialization in business intelligence, CCE, PUC-Rio, Rio de Janeiro, Brazil.

Gouvêa, E., Goldbarg, M., and Costa, W. (2002). Algoritmos evolucionários na solução do problema da programação de sondas de produção. In *Proceedings of the XXXIV Brazilian Symposium on Operations Research*, page 1–13, Rio de Janeiro, Brazil.

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

Gutleber, D. S., Heiberger, E. M., and Morris, T. D. (1995). Simulation analysis for integrated evaluation of technical and commercial risk. *Journal of Petroleum Technology*, 47:1062–1067.

Haneveld, W. K. (1986). *On Integrated Chance Constraints*, pages 113–138. Springer, Berlin, Heidelberg.

Hartsock, J. H. and Greaney, W. A. (1971). A stochastic inventory model for scheduling development drilling. *Society of Petroleum Engineers Journal*.

Hasle, G., Haut, R., Johansen, B., and Ølberg, T. (1996). Well activity scheduling-an application of constraint reasoning. *Artificial Intelligence in the Petroleum Industry: Symbolic and Computational Applications II*, pages 209–228.

Haugland, D., Jornsten, K., and Shayan, E. (1991). Modelling petroleum fields with movable platforms. *Applied Mathematical Modelling*, 15(1):33–39.

Haugland, D. and Tjøstheim, B. P. (2015). Optimal intake and routing of floating oil rigs in the north sea. In Murty, K., editor, *Case Studies in Operations Research*.

International Series in Operations Research & Management Science, vol. 212, page 315–336. Springer, New York, United States of America.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software, Articles*, 27(5):1–32.

He, Y. and Li, H. (2018). Probability density forecasting of wind power using quantile regression neural network and kernel density estimation. *Energy conversion and management*, 164:374–384.

Hebbali, A. and Hebbali, M. A. (2017). Package 'olsrr'.

Heitsch, H. and Römisch, W. (2003). Scenario reduction algorithms in stochastic programming. *Computational optimization and applications*, 24(2):187–206.

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

Holmager, M. and Redda, M. (2013). *OffshoreBook: An Introduction to the Offshore Industry*. Offshore Center Danmark, Oslo, third edition.

Hong, L. J., Luo, J., and Nelson, B. L. (2015). Chance constrained selection of the best. *INFORMS Journal on Computing*, 27(2):317–334.

Horton, G. and Dedigama, T. (2006). Drilling and petroleum engineering program and project management at santos ltd. In *Proceedings of the SPE Asia Pacific Oil and Gas Conference and Exhibition 2006: Thriving on Volatility*, pages 1108–1123.

Hota, A. R., Cherukuri, A., and Lygeros, J. (2019). Data-driven chance constrained optimization under wasserstein ambiguity sets. In *2019 American Control Conference (ACC)*, pages 1501–1506. IEEE.

Hu, Z. and Hong, L. J. (2013). Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, pages 1695–1724.

Husni, M. H. (2008). *A multiperiod optimization model to schedule large-scale petroleum development projects*. PhD thesis, Texas A&M University.

Iachan, R. (2009). A Brazilian experience: 40 years using operations research at Petrobras. *International Transactions in Operational Research*, 16(5):585–593.

IADC (2015). Still sunny, but clouds are brewing. https://www.drillingcontractor.org/wp-content/uploads/2015/01/PVD-11.jpg. (Accessed 14 January 2020).

IFP School (2015). What are the main steps of an oil or gas field development project. https://ferasa.net/what-are-the-main-steps-of-an-oil-or-gas-field-development-project. (Accessed 17 December 2019).

IOM3 (2015). An introduction to oil & gas drilling and well operations: educational material from the IOM3 oil and division. http://www.iom3.org/. (Accessed 16 December 2019).

Irani, B. (2007). Technology update: Pemex uses data management as a vital part of decision making in the oil patch. *Journal of Petroleum Technology*, 59(3).

Irgens, M., Guzman, R. P., Stamatopoulos, J., and Jackson, K. (2008). Optimization for operational decision support: the rig fleet management case. In *Proceedings of the SPE Annual Technical Conference and Exhibition*, volume 6, pages 4254–4267, Denver, United States of America.

Irgens, M. and Lavenue, W. L. (2007). Use of advanced optimization techniques to manage a complex drilling schedule. In *Proceedings of the SPE Annual Technical Conference and Exhibition*, volume 6, pages 3769–3777, Anaheim, United States of America.

Iyer, R. R., Grossmann, I. E., Vasantharajan, S., and Cullick, A. S. (1998). Optimal planning and scheduling of offshore oil field infrastructure investment and operations. *Industrial and Engineering Chemistry Research*, 37(4):1380–1397.

Ji, R. and Lejeune, M. (2018). Data-driven distributionally robust chance-constrained programming with wasserstein metric. *Available at Optimization Online*.

Jiang, R. and Guan, Y. (2016). Data-driven chance constrained stochastic program. *Math. Program.*, 158:291–327.

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

Kaiser, M. J. and Snyder, B. (2013). The five offshore drilling rig markets. *Marine Police*, 39:201–214.

Kaut, M. (2021). Scenario generation by selection from historical data. *Computational Management Science*, 18(3):411–429.

Kaut, M. and Stein, W. (2003). *Evaluation of scenario-generation methods for stochastic programming*. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät ....

KDnuggets (2019). Comparison of the text distance metrics. https://www.kdnuggets.com/2019/01/comparison-text-distance-metrics.html. Last time accessed in 14^{th} November 2022.

Kelly, J. D., Menezes, B. C., and Grossmann, I. E. (2017). Decision automation for oil and gas well startup scheduling using milp. *Computer Aided Chemical Engineering*, 40:1399–1404.

Khodro Diesel, K. (2019). Trucks technical specifications. https://www.tatratrucks.com. (Accessed 14 January 2020).

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

Kogan, A. and Lejeune, M. A. (2014). Threshold boolean form for joint probabilistic constraints with random technology matrix. *Mathematical Programming*, 147(1):391–427.

Kromodihardjo, S. and Kromodihardjo, E. S. (2016). Modeling of well service and workover to optimize scheduling of oil well maintenance. *Applied Mechanics and Materials*, 836:311–316.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer.

Kulachenko, I. and Kononova, P. (2020). *A matheuristic for the drilling rig routing problem*. Springer International Publishing, Cham.

Lang, D. T. (2004). Word stemming in r. *Department of Statistics, UC Davis*.

Lange, G. and Lin, F. (2014). Modeling well scheduling as a virtual enterprise with intelligent agents. In *2014 IEEE 17th International Conference on Computational Science and Engineering*, pages 89–96.

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

Lasschuit, W. and Thijssen, N. (2004). Supporting supply chain planning and scheduling decisions in the oil and chemical industry. *Computers & Chemical Engineering*, 28(6):863–870. FOCAPO 2003 Special issue.

Lasserre, J. B. and Weisser, T. (2021). Distributionally robust polynomial chance-constraints under mixture ambiguity sets. *Mathematical Programming*, 185(1):409–453.

Laud, P. W. and Ibrahim, J. G. (1995). Predictive model selection. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):247–262.

Lejeune, M. and Noyan, N. (2010). Mathematical programming approaches for generating p-efficient points. *European Journal of Operational Research*, 207(2):590–600.

Lejeune, M. A. (2012). Pattern definition of the p-efficiency concept. *Annals of Operations Research*, 200(1):23–36.

Lejeune, M. A. and Margot, F. (2016). Solving chance-constrained optimization problems with stochastic quadratic inequalities. *Operations Research*, 64(4):939–957.

Lejeune, M. A. and Prékopa, A. (2018). Relaxations for probabilistically constrained stochastic programming problems: review and extensions. *Annals of Operations Research*, pages 1–22.

Lejeune, M. A. and Ruszczyński, A. (2007). An efficient trajectory method for probabilistic production-inventory-distribution problems. *Operations Research*, 55(2):378–394.

Li, Z. and Li, Z. (2015). Chance constrained planning and scheduling under uncertainty using robust optimization approximation. *IFAC-PapersOnLine*, 48(8):1156–1161.

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

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

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

Litvak, M., Gane, B., Williams, G., Mansfield, M., Angert, P., Macdonald, C., McMurray, L., Skinner, R., and Walker, G. J. (2007). Field development optimization technology. pages 400–409, Houston, TX, USA.

Litvak, M., Onwunalu, J., and Baxter, J. (2011). Field development optimization with subsurface uncertainties. In *Proceedings of the SPE Annual Technical Conference and Exhibition*, volume 2, pages 1691–1702, Denver, United States of America.

Litvak, M. L. and Angert, P. F. (2009). Field development optimization applied to giant oil fields. In *Proceedings of the SPE Reservoir Simulation Symposium*, The Woodlands, United States of America.

Lorenzoni, L. L. and Polycarpo, W. M. (2010). Scatter search aplicado ao problema de otimização de itinerários de sondas de produção terrestre. In *Proceedings of the XXX National Production Engineering Meeting*, pages 1–14, São Carlos, Brazil.

Luedtke, J. and Ahmed, S. (2008). A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19(2):674–699.

Ma, Z., Vajargah, A. K., Lee, H., Kansao, R., Darabi, H., and Castineira, D. (2018). Applications of machine learning and data mining in speedwise® drilling analytics: a case study. In *Proceedings of the Abu Dhabi International Petroleum Exhibition & Conference*, Abu Dhabi, United Arab Emirates.

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

Maia, R. S., Gonzaga, C. S. M., Lima Junior, F. C., and Bittencourt, V. G. (2002). Otimização das intervenções em poços de petróleo por sondas de produção terrestre: busca tabu. In *Proceedings of the XXXIV Brazilian Symposium on Operations Research*, Rio de Janeiro, Brazil.

Marchesi, J. F., Carrilho, L. M., Gelli, J. G. M., Hamacher, S., and Sousa, P. H. (2019). Otimização do planejamento de projetos: aplicação á construção de poços marítimos em uma indústria de éleo e gás. In *Proceedings of the LI Brazilian Symposium on Operations Research*, Campinas, Brazil.

Markit, I. (2019). Petrodata offshore rig day rate trends. https://www.ihs.com/products/oil-gas-drilling-rigs-offshore-day-rates.html. Accesed 16 December 2019.

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

Marques, L. C., Machado, F. A. P. P., Oliveira, F. C., and 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*, pages 2538–2547, Salvador, Brazil.

Martin, W. W., Walters, J. V., and Woodard, P. W. (2010). Application of a well slot optimization process to drilling large numbers of wells in clusters on artificial islands. In *Proceedings of the Abu Dhabi International Petroleum Exhibition and Conference*, pages 1094–1099, Abu Dhabi, United Arab Emirates.

Mazzini, F. F., de Melo, R., Accioly, S., and Vasconcelos, R. V. J. (2002). Dimensionamento de equipamentos críticos da cadeia de suprimento da perfuração e completação de poços. In *Proceedings of the XLII Brazilian Symposium on Operations Research*, page 2243–2250, Bento Gonçalves, Brazil.

McCullagh, P. and Nelder, J. A. (2019). *Generalized linear models*. Routledge.

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):93–100.

McKechnie, B., Gardner, K., and Dawson, B. (2002). Rig management system (RMS ii). In *Proceedings of the IADC/SPE Drilling Conference*, pages 207–211, Dallas, United States of America.

Miller, B. L. and Wagner, H. M. (1965). Chance constrained programming with joint constraints. *Operations Research*, 13(6):930–945.

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

Moura, A. V., Pereira, R. A., and De Souza, C. C. (2008). Scheduling activities at oil wells with resource displacement. *International Transactions in Operational Research*, 15(6):659–683.

Namakshenas, M. and Pishvaee, M. S. (2019). Data-driven robust optimization. *Robust and Constrained Optimization: Methods and Applications*, pages 1–40.

Nascimento, J. M. (2002). Ferramentas computacionais híbridas para a otimização da produção de petróleo em águas profundas. Master's thesis, Universidade Estadual de Campinas, Campinas, Brazil.

Neiro, S. M. S. and Pinto, J. M. (2004). A general modeling framework for the operational planning of petroleum supply chains. *Computers & Chemical Engineering*, 28(6):871–896. FOCAPO 2003 Special issue.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Nemirovski, A. (2012). On safe tractable approximations of chance constraints. *European Journal of Operational Research*, 219(3):707–718.

Nemirovski, A. and Shapiro, A. (2007). Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996.

Nesvold, R. L., Herring, T. R., and Currie, J. C. (1996). Field development optimization using linear programming coupled with reservoir simulation - Ekofisk Field. In *Proceedings of the European Petroleum Conference*, volume 1, pages 507–519, Milan, Italy.

Neves, T. A. (2007). Heurísticas com memória adaptativa aplicadas ao problema de roteamento e scheduling de sondas de manutenção. Master's thesis, Universidade Federal Fluminense, Niterói, Brazil.

Neves, T. A. and Ochi, L. S. (2006). GRASP com memória adaptativa na solução de um problema de roteamento de veículos com múltiplas origens. In *Proceedings of the XXXVIII Brazilian Symposium on Operations Research*, pages 1323–1332, Goiânia, Brazil.

Neves, T. A. and Ochi, L. S. (2007). GRASP com memória adaptativa aplicado ao problema de roteamento e scheduling de sondas de manutencao. In *Proceedings of the XXVII Brazilian Society of Computing*, pages 1–10.

Nikzad, E., Bashiri, M., and Oliveira, F. (2019). Two-stage stochastic programming approach for the medical drug inventory routing problem under uncertainty. *Computers & Industrial Engineering*, 128:358–370.

Ning, C. and You, F. (2019). Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming. *Computers & Chemical Engineering*, 125:434–448.

Noronha, T. and Aloise, D. (2001). Algoritmos e estratégias e solução para o problema do gerenciamento de sondas de produção terrestre na bacia petrolífera potiguar. *Revista Eletrônica de Iniciação Científica*, 1:1–11.

NORWEP (2019). Oil & gas upstream. https://www.norwep.com/Partners/Oil-Gas-Upstream. (Accessed 14 January 2020).

Oliveira, E. F., Pagoto, F. B., Silva, F. T., and Lorenzoni, L. L. (2007). Scatter search aplicado ao problema de otimização da alocação de sondas de produção em poços de petróleo. In *Proceedings of XXVII National Production Engineering Meeting*, volume 0, pages 1–10, Foz do Iguaçu, Brazil.

Oliveira, F., Gupta, V., Hamacher, S., and Grossmann, I. E. (2013). A lagrangean decomposition approach for oil supply chain investment planning under uncertainty with risk considerations. *Computers & Chemical Engineering*, 50:184–195.

Omosebi, O., Osisanya, S. O., and Ahmed, R. (2014). Integrated model-based approach to drilling project management. In *Proceedings of the SPE Nigeria Annual International Conference and Exhibition*, pages 519–531, Lagos, Nigeria.

Ondeck, A., Drouven, M., Blandino, N., and Grossmann, I. E. (2019). Multi-operational planning of shale gas pad development. *Computers & Chemical Engineering*, 126:83–101.

Onwunalu, J., Litvak, M., Durlofsky, L. J., and Aziz, K. (2008). Application of statistical proxies to speed up field development optimization procedures. In *Proceedings of the Abu Dhabi International Petroleum Exhibition and Conference*, volume 1, pages 342–355, Abu Dhabi, United Arab Emirates.

Osmundsen, E., Roll, K. H., and Tveterås, R. (2010). Exploration drilling productivity at the Norwegian shelf. *Journal of Petroleum Science and Engineering*, 73:122–128.

Pacheco, A. V. F. (2011). *Métodos de solução para o problema da alocação de sondas a poços de petróleo*. Undergraduate final project, Universidade Federal do Espírito Santo, São Mateus, Brazil.

Pacheco, A. V. F., Dias Filho, A. C. T., and Ribeiro, G. M. (2009a). Uma heurística para o problema da alocação e sondas de produção em poços de petróleo. In *Proceedings of the National Production Engineering Meeting*, page 1–12, Salvador, Brazil.

Pacheco, A. V. F., Ribeiro, G. M., and Mauri, G. R. (2009b). Novas soluções para o problema da alocação de sondas de produção a poços de petróleo com um grasp+ path-relinking. In *Proceedings of the XLII Brazilian Symposium on Operations Research*, page 3129–3138, Bento Gonçalves, Brazil.

Pacheco, A. V. F., Ribeiro, G. M., and Mauri, G. R. (2010). A grasp with path-relinking for the workover rig scheduling problem. *International Journal of Natural Computing Research*, 1(2):1–14.

Paiva, R. O. (1997). Otimização do itinerário de sondas de intervenção. Master's thesis, Universidade Estadual de Campinas (Unicamp), Campinas, Brazil.

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

Pandolfi, D., Villagra, A., de San Pedro, E., Lasso, M., and Leguizamón, G. (2010). An experimental study of an evolutionary tool for scheduling in oil wells. In García-Pedrajas, N., Herrera, F., Fyfe, C., M., B. J., and Ali, M., editors, *Trends in Applied Intelligent Systems. IEA/AIE 2010. Lecture Notes in Computer Science, vol. 6098*, pages 576–585. Springer, Berlin, Germany.

Pereira, R. A. (2005). Scheduling of development activities of oil wells: Grasp. Master's thesis, Universidade Estadual de Campinas (Unicamp), Campinas, SP, Brazil.

Pereira, R. A., Arnaldo, V. M., and De Souza, C. C. (2005a). GRASP strategies for scheduling activities at oil wells with resource displacement. Technical report, Universidade Estadual de Campinas (Unicamp). Campinas, SP, Brazil.

Pereira, R. A., Moura, A. V., and De Souza, C. C. (2005b). Comparative experiments with grasp and constraint programming for the oil well drilling problem. *Lecture Notes in Computer Science*, 3503:328–340.

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

Pessôa Filho, P. d. A., Santos, F. L. S., and Mansoori, G. A. (2006). An update on the developments in petroleum production research in Brazil. *Journal of Petroleum Science and Engineering*, 51(1):1–5. Special Issue on Petroleum Production Research in Brazil.

Petrobras (2014). Comparativo entre os diferentes tipos de plataformas. http://www.petrobras.com.br/infograficos/tipos-de-plataformas/desktop/index.html. Accesed 16 December 2019.

Pflug, G. C. (2001). Scenario tree generation for multiperiod financial optimization by optimal discretization. *Mathematical programming*, 89(2):251–271.

Pittman, J. (1985). Computer speeds offshore well planning, rig scheduling. *Oil & gas journal*, 83:84–97.

Postek, K., Ben-Tal, A., Den Hertog, D., and Melenberg, B. (2018). Robust optimization with ambiguous stochastic constraints under mean and dispersion information. *Operations Research*, 66(3):814–833.

Prékopa, A. (1990). Dual method for a one-stage stochastic programming problem with random rhs obeying a discrete probability distribution. *Z. Oper. Res*, 34:441–461.

Prékopa, A. (2003). Probabilistic programming. *Handbooks in operations research and management science*, 10:267–351.

Prékopa, A. (2013). *Stochastic programming*, volume 324. Springer Science & Business Media.

Prékopa, A. (2015). On probabilistic constrained programming. In *Proceedings of the Princeton Symposium on Mathematical Programming*, pages 113–138.

Prékopa, A. (1971). *On Probabilistic Constrained Programming*, pages 113–138. Princeton University Press.

Prékopa, A. (1973). Contributions to the theory of stochastic programming. *Mathematical Programming*, 4(1):202–221.

Pérez, M. F., Oliveira, F., and Hamacher, S. (2019). Otimizaćão do itinerário e dimensionamento de sondas workover usando modelos de programação matemática. In *Proceedings of the LI Brazilian Symposium on Operations Research*, Campinas, Brazil.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Racine, J. S. (2008). *Nonparametric econometrics: A primer*, volume 4. Now Publishers Inc.

Rahimian, H. and Pagnoncelli, B. (2020). Contextual chance-constrained programming. *Available on Optimization Online*.

Ribeiro, G. M., Desaulniers, G., and Desrosiers, J. (2012a). A branch-price-and-cut algorithm for the workover rig routing problem. *Computers and Operations Research*, 39(12):3305–3315.

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

Ribeiro, G. M., Laporte, G., and Mauri, G. R. (2012b). A comparison of three metaheuristics for the workover rig routing problem. *European Journal of Operational Research*, 220(1):28–36.

Ribeiro, G. M., Mauri, G. R., and Lorena, L. A. N. (2011). A simple and robust simulated annealing algorithm for scheduling workover rigs on onshore oil fields. *Computers and Industrial Engineering*, 60(4):519–526.

Rocha, C. T. M., Aloise, D., Aloise, D. J., and Melo, J. D. (2003). Heurísticas paralelas de busca em vizinhança variável para o problema de otimização do emprego de sondas de produção terrestre. In *Proceedings of the XXXV Brazilian Symposium on Operations Research*, page 1972–1982, Natal, Brazil.

Rokach, L. and Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Ruszczyński, A. (2002). Probabilistic programming with discrete distributions and precedence constrained knapsack polyhedra. *Mathematical Programming*, 93(2):195–215.

Sabry, G. A., Goldbarg, M. C., and Goldbarg, E. F. G. (2012). Um estudo algorítmico do problema da programação de sondas de produção. In *Proceedings of the XVI CLAIO – Congresso Latino Iberoamericano de Investigation Operativa*, pages 2757–2768.

Sahoo, N. P. and Biswal, M. P. (2005). Computation of probabilistic linear programming problems involving normal and log-normal random variables with a joint constraint. *International Journal of Computer Mathematics*, 82(11):1323–1338.

Santos, I. M. (2018). Mathematical programming models and local search algorithms for the offshore rig scheduling problem. Master's thesis, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brazil.

Santos, I. M., Carrilho, L. M., Oliveira, F. L. C., Andrade, T., and Ribas, G. (2017). Offshore oil rig scheduling simulation: a multi-perspective approach. In *Proceedings of the XLIX Brazilian Symposium on Operations Research*.

Santos, I. M., Hamacher, S., and Oliveira, F. (2021). A systematic literature review for the rig scheduling problem: Classification and state-of-the-art. *Computers & Chemical Engineering*, 153:107443.

Santos, I. M., Hamacher, S., and Oliveira, F. (2023). A data-driven optimization model for the workover rig scheduling problem: Case study in an oil company. *Computers & Chemical Engineering*, 170:108088.

Sarica, S. and Luo, J. (2021). Stopwords in technical language processing. *Plos one*, 16(8):e0254937.

Saxena, A., Goyal, V., and Lejeune, M. A. (2010). Mip reformulations of the probabilistic set covering problem. *Mathematical programming*, 121(1):1–31.

Serra, T. (2012). Programação por restrições e escalonamento baseado em restrições: Um estudo de caso na programação de recursos para o desenvolvimento de poços de petróleo. Master's thesis, Universidade de São Paulo, São Paulo, Brazil.

Serra, T., Nishioka, G., and Marcellino, F. J. M. (2011). A constraint-based scheduling of offshore well development activities with inventory management. In *Proceedings of the Brazilian Symposium on Operations Research*, pages 3459–3469, Ubatuba, Brazil.

Serra, T., Nishioka, G., and Marcellino, F. J. M. (2012a). Integrated project selection and resource scheduling of offshore oil well developments: An evaluation of CP models and heuristic assumptions. In *ICAPS 2012 - 22nd International Conference on Automated Planning and Scheduling; COPLAS 2012 - Proceedings of the Workshop on Constraint Satisfaction Techniques for Planning and Scheduling Problems*, pages 51–58.

Serra, T., Nishioka, G., and Marcellino, F. J. M. (2012b). The offshore resources scheduling problem: Detailing a constraint programming approach. In *International Conference on Principles and Practice of Constraint Programming*, volume 7514, pages 823–839.

Serra, T., Nishioka, G., and Marcellino, F. J. M. (2012c). On estimating the return of resource aquisitions through scheduling: An evaluation of continuous-time milp models to approach the development of offshore oil wells. In *Proceedings of the 6th Scheduling and Planning Applications Workshop*, Atibaia, Brazil.

Shaji, N., Sundar, C., Jagyasi, B., and 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., and Pal, S., editors, *Pattern Recognition and Machine Intelligence. PReMI 2019. Lecture Notes in Computer Science, vol. 11941*, pages 385–394. Springer, Cham, Switzerland.

Shapiro, A. (2003). Monte carlo sampling methods. *Handbooks in operations research and management science*, 10:353–425.

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

Silva, F. T. and 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. *Brazilian Journal of Production Engineering*, 4:169–184.

Silva, L. M. R., Santos, A. M. P., and Guedes Soares, C. (2016). A mixed integer formulation for the offshore rig scheduling problem. In *Proceedings of the 3rd International Conference on Maritime Technology and Engineering*, volume 2, pages 1005–1012, Lisbon, Portugal.

Smith, W. E. (1956). Various optimizers for single-stage production. *Naval Research Logistics Quarterly*, 3(1-2):59–66.

Soares, W. K. S., Costa, A. P. C. S., and Aloise, D. J. (2011). Considerações sobre o problema do agendamento de sondas de manutenção "onshore" e proposição de instâncias. In *Proceedings of the XLIII Brazilian Symposium on Operations Research*, Ubatuba, Brazil.

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

Street, J. O., Carroll, R. J., and Ruppert, D. (1988). A note on computing robust regression estimates via iteratively reweighted least squares. *The American Statistician*, 42(2):152–154.

Sumaida, A. S. B., El Gebaly, M. A., Loayza, A. F. J., Al Hammadi, A., and Abu Aish, A. N. (2013). Systematic approach to optimize the rig move within adco onshore field. In *Proceedings of the SPE/IADC Middle East Drilling Technology Conference and Exhibition*, pages 780–788, Dubai, United Arab Emirates.

Sun, Y., Ma, L., and Wang, S. (2015). A comparative evaluation of string similarity metrics for ontology alignment. *Journal of Information &Computational Science*, 12(3):957–964.

Suslick, S. B. and Schiozer, D. J. (2004). Risk analysis applied to petroleum exploration and production: an overview. *Journal of Petroleum Science and Engineering*, 44(1):1–9.

Suslick, S. B., Schiozer, D. J., and Rodriquez, M. R. (2009). Uncertainty and risk analysis in petroleum exploration and production. *Terræ*, 6:30–41.

Tavallali, M. S., Bakhtazma, F., Meymandpour, A., Sadeghi, F., Hamed, M., and Karimi, I. A. (2015). A drilling scheduling toolbox for oil and gas reservoirs. *Computer Aided Chemical Engineering*, 37:2453–2458.

Tavallali, M. S. and Karimi, I. A. (2014). Perspectives on the design and planning of oil field infrastructure. *Computer Aided Chemical Engineering*, 34:163–172.

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

Tavallali, M. S. and Zare, M. (2018). Planning the drilling rig activities – routing and allocation. *Computer Aided Chemical Engineering*, 43:1219–1224.

Thome, A. M. T., Scavarda, L. F., and Scavarda, A. J. (2016). Conducting systematic literature review in operations management. *Production Planning & Control*, 27(5):408–420.

TOYO (2019). Oil & gas development. https://www.toyo-eng.com/jp/en/products/oil_and_gas_development. (Accessed 14 January 2020).

Tozzo, E., Costa, A. P. S., and Lins, I. D. (2020). A hybrid multi-objective genetic algorithm for scheduling heterogeneous workover rigs on onshore oil fields. *Journal of Petroleum Science and Engineering*, 195:107935.

Trindade, V. A. (2005). Desenvolvimento e análise experimental da metaheurística GRASP para um problema de planejamento de sondas de manutenção. Master's thesis, Universidade Federal Fluminense, Niterói, Brazil.

Trindade, V. A. and Ochi, L. S. (2004). Proposta e avaliação experimental de heurísticas GRASP para um problema de escalonamento de veículos. In *Proceedings of the XXXVI Brazilian Symposium on Operations Research,*, São João Del Rey, Brazil.

Trindade, V. A. and Ochi, L. S. (2005). Hybrid adaptive memory programming using GRASP and path relinking for the scheduling workover rigs for onshore oil production. In *Proceedings of the HIS 2005: Fifth International Conference on Hybrid Intelligent Systems*, pages 500–502.

Van Den Heever, S. A. and Grossmann, I. E. (2000). An iterative aggregation/disaggregation approach for the solution of a mixed-integer nonlinear oilfield infrastructure planning model. *Industrial and Engineering Chemistry Research*, 39(6):1955–1971.

Van Den Heever, S. A. and Grossmann, I. E. (2006). A mixed-integer nonlinear programming approach to the optimal planning of offshore oilfield infrastructures. In Appa, G., Pitsoulis, L., and Williams, H. P., editors, *Handbook on Modelling for Discrete Optimization. International Series in Operations research & Management Science. vol. 88*, pages 291–315. Springer, Boston, United States of America.

Vasconcellos, R. V. J. and Ferreira Filho, V. J. M. (2006). Algoritmo genético para o problema de scheduling de projetos com restrição de recurso: uma aplicação em operações em poços de petróleo. *Proceedings of the Brazilian Symposium on Operations Research*, pages 1795–1802.

Vasconcelos, D., Nogueira, E., Sousa, S., and 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*, pages 1705–1713, Rio da Janeiro, Brazil.

Verderame, P. M., Elia, J. A., Li, J., and Floudas, C. A. (2010). Planning and scheduling under uncertainty: a review across multiple sectors. *Industrial & engineering chemistry research*, 49(9):3993–4017.

Villagra, A., Pandolfi, D., and Leguizamón, G. (2013). Handling constraints with an evolutionary tool for scheduling oil wells maintenance visits. *Engineering Optimization*, 45(8):963–981.

Wang, D. and Li, M. (2017). Robust stochastic configuration networks with kernel density estimation for uncertain data regression. *Information Sciences*, 412:210–222.

Wang, Y., Estefen, S. F., Lourenço, M. I., and Hong, C. (2019). Optimal design and scheduling for offshore oil-field development. *Computers & Chemical Engineering*, 123:300–316.

Wets, R. J. (2002). Stochastic programming models: wait-and-see versus here-and-now. In *Decision Making Under Uncertainty*, pages 1–15. Springer.

Wigwe, M. E., Bougre, E. S., Watson, M., and Giussani, A. (2020). Comparative evaluation of multi-basin production performance and application of spatio-temporal models for unconventional oil and gas production prediction. *Journal of Petroleum Exploration and Production Technology*, 10(8):3091–3110.

Xie, W. (2019). On distributionally robust chance constrained programs with wasserstein distance. *Mathematical Programming*, pages 1–41.

Xie, W. (2021). On distributionally robust chance constrained programs with wasserstein distance. *Mathematical Programming*, 186(1):115–155.

Xie, W. and Ahmed, S. (2018). On deterministic reformulations of distributionally robust joint chance constrained optimization problems. *SIAM Journal on Optimization*, 28(2):1151–1182.

Yan, X. and Su, X. (2009). *Linear regression analysis: theory and computing*. World Scientific.

Yang, W. and Xu, H. (2016). Distributionally robust chance constraints for non-linear uncertainties. *Mathematical Programming*, 155(1):231–265.

Yuan, Z. and Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association*, 100(472):1202–1214.

Yujian, L. and Bo, L. (2007). A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095.

Zahran, H. F. and Al-Fardan, F. I. (2014). Automation of ADCO well delivery process - a dream that has become a reality. In *Proceedings of the Abu Dhabi International Petroleum Exhibition and Conference*, volume 5, pages 3909–3914, Abu Dhabi, United Arab Emirates.

Zhang, S., Jia, R., He, D., and Chu, F. (2022). Data-driven robust optimization based on principle component analysis and cutting plane methods. *Industrial & Engineering Chemistry Research*, 61(5):2167–2182.

Zhang, Y., Feng, Y., and Rong, G. (2016). Data-driven chance constrained and robust optimization under matrix uncertainty. *Industrial & Engineering Chemistry Research*, 55(21):6145–6160.

Zhao, P. and Xiao, Q. (2016). Portfolio selection problem with value-at-risk constraints under non-extensive statistical mechanics. *Journal of computational and applied mathematics*, 298:64–71.

Zhao, Z. and Kumar, M. (2017). Split-bernstein approach to chance-constrained optimal control. *Journal of Guidance, Control, and Dynamics*, 40(11):2782–2795.

Zymler, S., Kuhn, D., and Rustem, B. (2013). Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1):167–198.

A
Search query generation

Figure A.1 shows the word plot that supported the search query used in this SLR.

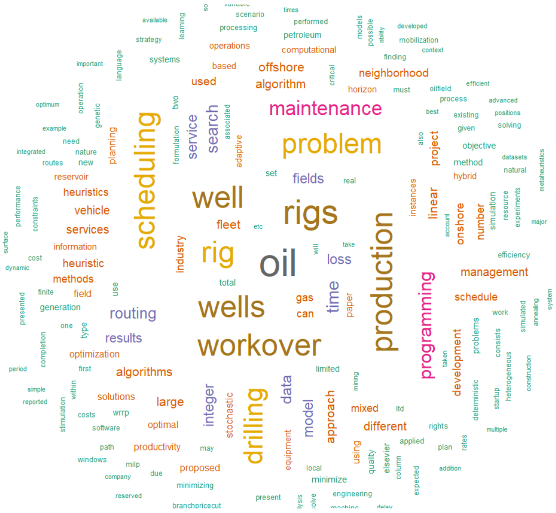


Figure A.1: Word clouds plot.

Table A.1 contains the complete search query used in the Scopus database during this SLR.

(TITLE-ABS-KEY(schedul*) OR TITLE-ABS-KEY(reschedul*) OR TITLE-ABS-KEY(rout*) OR TITLE-ABS-KEY(mobilization) OR TITLE-ABS-KEY("move*") OR TITLE-ABS-KEY("programing")) AND (TITLE-ABS-KEY(simulat*) OR TITLE-ABS-KEY(optimiz*) OR TITLE-ABS-KEY(model*) OR TITLE-ABS-KEY(algorithm) OR TITLE-ABS-KEY(*heuristic*) OR TITLE-ABS-KEY(procedure) OR TITLE-ABS-KEY(technique) OR TITLE-ABS-KEY(system)) AND (TITLE-ABS-KEY(wells) OR TITLE-ABS-KEY("oil well") OR TITLE-ABS-KEY(oil) OR TITLE-ABS-KEY (oilfield) OR TITLE-ABS-KEY(petroleum)) AND (((TITLE-ABS-KEY(onshore) OR TITLE-ABS-KEY(offshore)) AND (TITLE-ABS-KEY(workover) OR TITLE-ABS-KEY(drilling) OR TITLE-ABS-KEY("well completion") OR TITLE-ABS-KEY("downhole completion") OR TITLE-ABS-KEY ("reservoir evaluation"))) OR ((TITLE-ABS-KEY(onshore) OR TITLE-ABS-KEY(offshore)) AND (TITLE-ABS-KEY(rig) OR TITLE-ABS-KEY(rigs))) OR ((TITLE-ABS-KEY(rig) OR TITLE-ABS-KEY(rigs)) AND (TITLE-ABS-KEY(workover) OR TITLE-ABS-KEY (drilling) OR TITLE-ABS-KEY("well completion") OR TITLE-ABS-KEY("downhole completion") OR TITLE-ABS-KEY("reservoir evaluation")))) OR ((((TITLE(rig) OR TITLE(rigs) OR TITLE(vessel)) AND (TITLE(workover) OR TITLE(drilling) OR TITLE("well completion") OR TITLE("downhole completion") OR TITLE("reservoir evaluation") OR TITLE(well))) AND ((TITLE-ABS-KEY(schedul*) OR TITLE-ABS-KEY(reschedul*) OR TITLE-ABS-KEY(rout*) OR TITLE-ABS-KEY(mobilization) OR TITLE-ABS-KEY("move*") OR TITLE-ABS-KEY(planning) OR TITLE(problem)) OR (TITLE-ABS-KEY(simulat*) OR TITLE-ABS-KEY(optimi?*) OR TITLE-ABS-KEY(model*) OR TITLE-ABS-KEY (algorithm) OR TITLE-ABS-KEY(*heuristic*) OR TITLE-ABS-KEY(programming)))) OR (((TITLE(rig) OR TITLE(rigs)) OR (TITLE(workover) OR TITLE(drilling) OR TITLE("well completion") OR TITLE("downhole completion") OR TITLE("reservoir evaluation") OR TITLE(well))) AND ((TITLE(schedul*) OR TITLE(reschedul*) OR TITLE (rout*) OR TITLE(mobilization) OR TITLE("move*") OR TITLE(planning)) AND (TITLE(simulat*) OR TITLE(optimi?*) OR TITLE(model*) OR TITLE(algorithm) OR TITLE(*heuristic*) OR TITLE(programming) OR TITLE(problem))))) OR (((TITLE(rig) OR TITLE(rigs)) OR (TITLE(workover) OR TITLE(drilling) OR TITLE("well completion") OR TITLE("downhole completion") OR TITLE("reservoir evaluation")) OR (TITLE(well))) AND (TITLE-ABS-KEY(schedul*) OR TITLE-ABS-KEY(reschedul*) OR TITLE-ABS-KEY(rout*) OR TITLE-ABS-KEY(mobilization) OR TITLE-ABS-KEY("move*")) AND (TITLE-ABS-KEY (oil) OR TITLE-ABS-KEY(oilfield) OR TITLE-ABS-KEY(petroleum)) AND (TITLE-ABS-KEY(simulat*) OR TITLE-ABS-KEY(optimi?*) OR TITLE-ABS-KEY(model*) OR TITLE-ABS-KEY(algorithm) OR TITLE-ABS-KEY(*heuristic*) OR TITLE-ABS-KEY(programming))) AND (LIMIT-TO(PUBSTAGE,"final")) AND (LIMIT-TO(DOCTYPE,"cp") OR LIMIT-TO(DOCTYPE,"ar") OR LIMIT-TO(DOCTYPE,"cr") OR LIMIT-TO(DOCTYPE,"re") OR LIMIT-TO(DOCTYPE,"ch") OR LIMIT-TO(DOCTYPE,"bk")) AND (LIMIT-TO(LANGUAGE,"English"))

Table A.1: Scopus search query used for the rigs scheduling literature review.

Table A.2 contains the complete search query used in the Web of Science database during this SLR.

```
(TS=(schedul* OR reschedul* OR rout* OR mobilization OR "move*" OR "programing" ) AND
TS=(simulat* OR optimiz* OR model* OR algorithm OR *heuristic* OR procedure OR technique
OR system) AND TS=("wells" OR "oil well" OR "oil" OR oilfield OR petroleum) AND ( (
TS=(onshore OR offshore) AND TS=("workover" OR drilling OR "well completion" OR "downhole
completion" OR "reservoir evaluation") ) OR ( TS=(onshore OR offshore) AND TS=("rig" OR
"rigs") ) OR ( TS=("rig" OR "rigs") AND TS=("workover" OR drilling OR "well completion"
OR "downhole completion" OR "reservoir evaluation") ) ) OR ( ( TI=("rig" OR "rigs" OR
vessel) AND TI=("workover" OR drilling OR "well completion" OR "downhole completion" OR
"reservoir evaluation") AND ( TS=(schedul* OR reschedul* OR rout* OR mobilization OR "move*"
OR planning) OR TI=(problem) OR TI=(simulat* OR optimi?* OR model* OR algorithm
OR *heuristic* OR programming) ) ) OR ( ( TI=("rig" OR "rigs") OR TI=("workover" OR
drilling OR "well completion" OR "downhole completion" OR "reservoir evaluation" OR well) )
AND TI=(schedul* OR reschedul* OR rout* OR mobilization OR "move*" OR planning) AND
TI=(simulat* OR optimi?* OR model* OR algorithm OR *heuristic* OR programming OR
problem) ) ) OR ( ( TI=("rig" OR "rigs") OR TI=("workover" OR drilling OR "well completion" OR
"downhole completion" OR "reservoir evaluation" OR well) ) AND TS=(schedul* OR reschedul* OR
rout* OR mobilization OR "move*") AND TS=("oil" OR oilfield OR petroleum) AND TS=(simulat*
OR optimi?* OR model* OR algorithm OR *heuristic* OR programming) )) AND IDIOMA:
(English) AND TIPOS DE DOCUMENTO: (Article OR Abstract of Published Item OR Book
OR Book Chapter OR Book Review OR Early Access OR Proceedings Paper)
```

Table A.2: Web of Science search query used for the rigs scheduling literature review.

Note: The search query selects only final studies in English published in journals, conferences, and books. And some terms might be in Portuguese.

B
Problems details

Follow, Figures B.1, B.2, B.3, B.4, and B.5 contain the number of papers found for each taxonomy (columns) for the DRSP, WRSP, WRRSP, FP, and RP, respectively. Note that the "Scheduling" and "Routing" taxonomies, respectively, refer to problems that only do scheduling and problems that consider routing with scheduling.

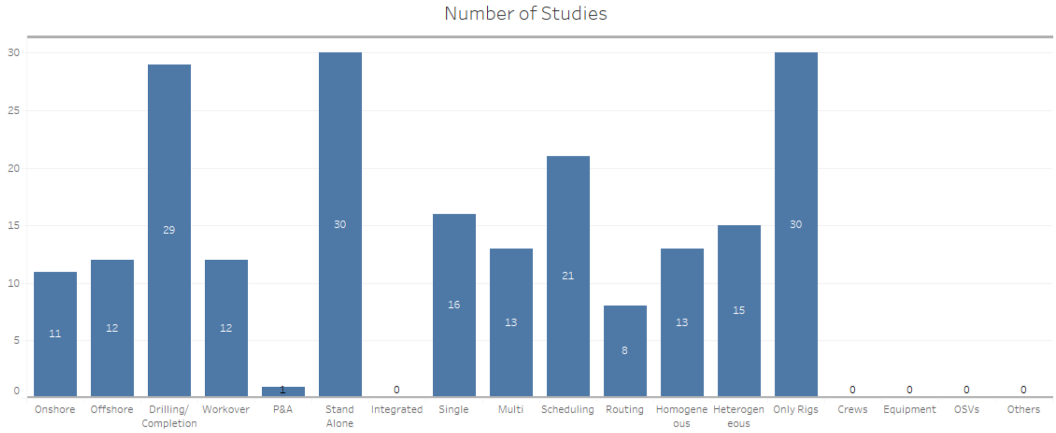


Figure B.1: Distribution of the drilling rig scheduling problem studies according to the taxonomy.

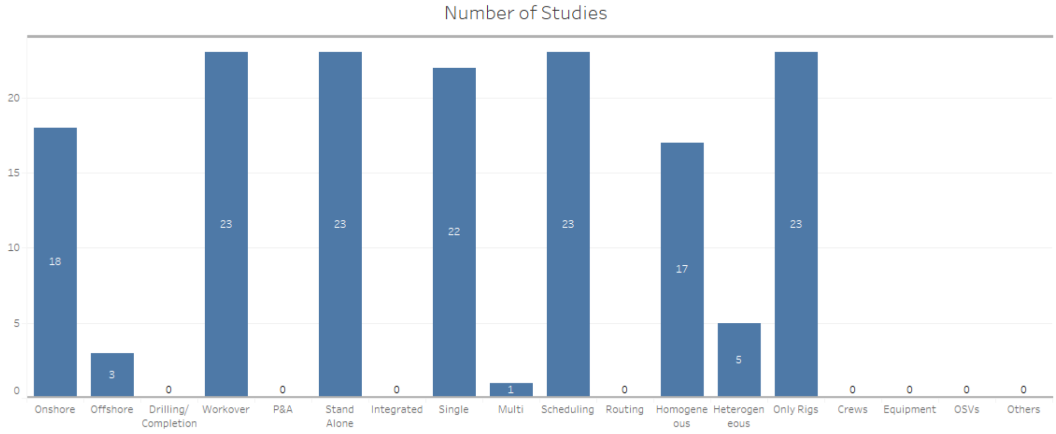


Figure B.2: Distribution of the workover rig scheduling problem studies according to the taxonomy.

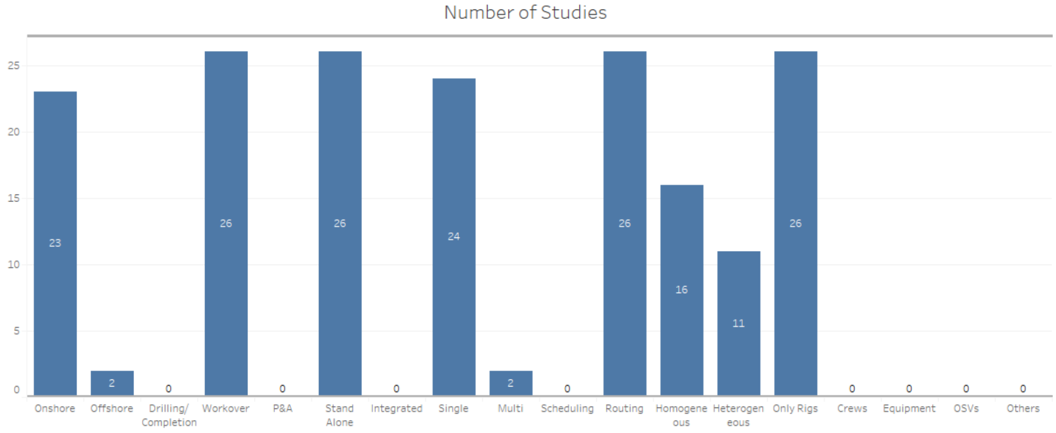


Figure B.3: Distribution of the workover rig routing and scheduling problem studies according to the taxonomy.

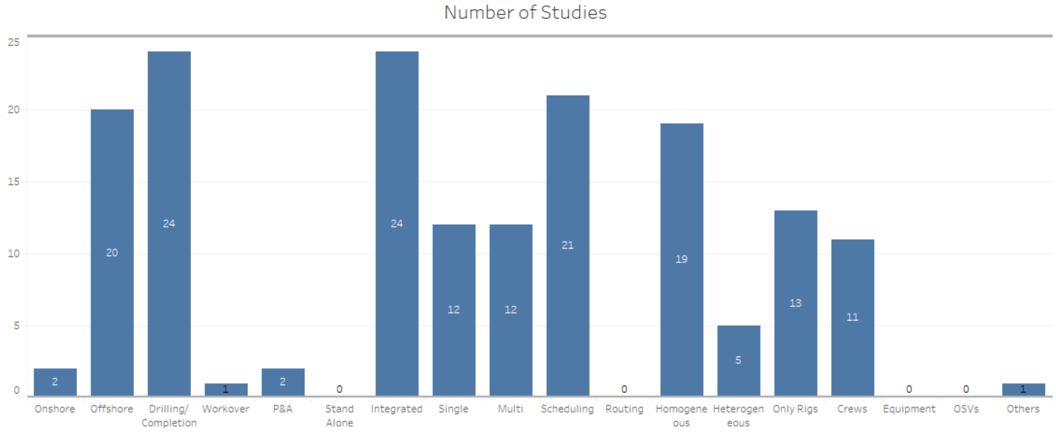


Figure B.4: Distribution of the field planning studies according to the taxonomy.

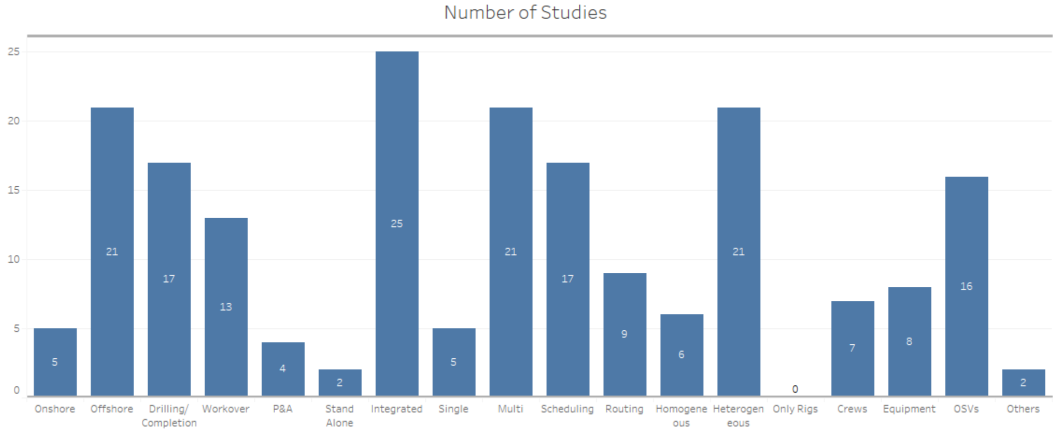


Figure B.5: Distribution of the resource planning studies according to the taxonomy.

C
SLR results details

The following table resumes the rig scheduling publications found in the literature review. The "Rout./Sched." row means "Routing/Scheduling" and note that the "Scheduling" and "Routing" taxonomies, respectively, refer to problems that only do scheduling and problems that consider routing with scheduling. In the "Fleet" row, "Homog." and "Heterog." are shortening for "Homogeneous" and "Heterogeneous", respectively. In the Method row, "Heur.", "Matheur.", "Simu.", "Simu-Opt.", and "Data Driven Opt." refer to "Heuristic", "Matheuristic", "Simulation", "Simulation-Optimization", and "Data-Driven Optimization", respectively. Last, the "O.F." row is the Objective Function row, it's field "Econ. Index" refers to "Economic Index" and "Prod." refers to "Oil Production".

Authors (Year)	Oilfield	Task	Fleet	Rout./Sched.	Method	Section	O.F.
Aronofsky (1962)	-	Drilling	Homog.	Sched.	Exact	2.4.1	Econ. Index
Aronofsky and Williams (1962)	-	Drilling	Homog.	Sched.	Exact	2.4.1	Econ. Index
Hartsock and Greaney (1971)	Offshore	Drilling, Others	Homog.	Sched.	Exact	2.4.1	Econ. Index
Barnes et al. (1977)	-	Workover	Homog.	Sched.	Heur.	2.4.2	Prod. Loss
Haugland et al. (1991)	Offshore	Drilling	Homog.	Rout.	Exact	2.4.1	Econ. Index
de Andrade Filho (1994)	Offshore	Drilling	Homog.	Sched.	Simu-Opt.	2.4.4.1	Econ. Index
Gutleber et al. (1995)	Offshore	Drilling	Homog.	Sched.	Simu.	2.4.1	Econ. Index
Eagle (1996)	Onshore	Drilling	Homog.	Sched.	Heur.	2.4.1	Econ. Index
Hasle et al. (1996)	Offshore	Drilling, Others	Homog.	Sched.	Exact	2.4.4.2	Time
Nesvold et al. (1996)	Offshore	Drilling	Homog.	Sched.	Exact, Heur.	2.4.4.1	Multi
Currie et al. (1997a)	Offshore	Drilling, Others	Homog.	Sched.	Exact	2.4.4.1	Econ. Index
Currie et al. (1997b)	Offshore	Drilling, Others	Homog.	Sched.	Exact	2.4.4.1	Econ. Index
Paiva (1997)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.3	Multi
Iyer et al. (1998)	Offshore	Drilling	Homog.	Sched.	Matheur.	2.4.4.1	Econ. Index
Paiva et al. (2000)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.3	Multi
Van Den Heever and Grossmann (2000)	Offshore	Drilling	Heterog.	-	Matheur.	2.4.4.1	Costs
Noronha and Aloise (2001)	Onshore	Workover	-	Sched.	Heur.	2.4.2	Prod. Loss
Accioly et al. (2002)	Offshore	Drilling, Others	Heterog.	Sched.	Exact	2.4.4.2	Multi
Aloise et al. (2002)	Onshore	Workover	Homog.	Sched.	Heur.	2.4.2	Prod. Loss
Gouvêa et al. (2002)	-	Workover	Homog.	Sched.	Heur.	2.4.2	Prod. Loss
Maia et al. (2002)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.3	Multi
McKechnie et al. (2002)	Offshore	Workover	Heterog.	Sched.	-	2.4.4.2	-
Nascimento (2002)	Offshore	Drilling, Others	Heterog.	Sched.	Exact, Heur.	2.4.4.2	Prod.
Rocha et al. (2003)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.3	Prod. Loss
Aseeri et al. (2004)	Offshore	Drilling	Homog.	Sched.	Exact	2.4.4.1	Econ. Index
Costa and Ferreira Filho (2004)	Onshore	Workover	Homog.	Sched.	Heur.	2.4.2	Prod. Loss
Trindade and Ochi (2004)	Onshore	Workover	Heterog.	Rout.	Heur.	2.4.3	Prod. Loss

Table C.1: Rig scheduling publications found in this literature review

Authors (Year)	Oilfield	Task	Fleet	Rout./Sched.	Method	Section	O.F.
Costa (2005)	Onshore	Workover	Homog.	Sched.	Heur.	2.4.2	Prod. Loss
Costa and Ferreira Filho (2005)	Onshore	Workover	Homog.	Sched.	Heur.	2.4.2	Prod. Loss
Pereira (2005)	Offshore	Drilling; Others	Heterog.	Sched.	Exact; Heur.	2.4.4.2	Prod.
Pereira et al. (2005a)	Offshore	Drilling; Others	Heterog.	Sched.	Exact; Heur.	2.4.4.2	Prod.
Pereira et al. (2005b)	Offshore	Drilling; Others	Heterog.	Sched.	Exact; Heur.	2.4.4.2	Prod.
Trindade (2005)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.3	Prod. Loss
Trindade and Ochi (2005)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.3	Prod. Loss
Aloise et al. (2006)	Onshore	Workover	Heterog.	Rout.	Heur.	2.4.3	Prod. Loss
Alves and Ferreira Filho (2006)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.3	Prod. Loss
Carvalho and Pinto (2006)	Offshore	Drilling	Homog.	Sched.	Mathem.	2.4.4.1	Econ. Index
Horton and Dedigama (2006)	Onshore	Drilling; Others	Heterog.	Sched.	-	2.4.4.2	-
Neves and Ochi (2006)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.3	Prod. Loss
Van Den Heever and Grossmann (2006)	Offshore	Drilling	Heterog.	Sched.	Mathem.	2.4.4.1	Econ. Index
Vasconcellos and Ferreira Filho (2006)	Offshore	Drilling; Others	Heterog.	Sched.	Heur.	2.4.4.2	Time
Barnes and Kokossis (2007)	Offshore	Drilling	Homog.	Sched.	Exact	2.4.4.1	Econ. Index
Irani (2007)	-	Drilling; Others	-	Sched.	-	2.4.1	-
Irgens and Lavenue (2007)	Onshore	Drilling	Heterog.	Sched.	Heur.	2.4.1	Multi
Litvak et al. (2007)	Offshore	Drilling; Others	Homog.	Sched.	Heur.	2.4.4.1	Multi
Neves (2007)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.3	Prod. Loss
Neves and Ochi (2007)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.3	Prod. Loss
Oliveira et al. (2007)	Onshore	Workover	Homog.	Sched.	Heur.	2.4.2	Prod. Loss
Cong et al. (2008)	Offshore	Drilling; Others	Heterog.	Sched.	Simu-Opt.	2.4.4.1	Prod.
Husni (2008)	-	Drilling; Others	Homog.	-	Exact; Heur.	2.4.1	Econ. Index
Irgens et al. (2008)	Onshore	Drilling	Heterog.	Sched.	Heur.	2.4.1	Multi
Lasrado (2008)	Offshore	Workover	Homog.	Sched.	Simu.	2.4.2	Multi
Moura et al. (2008)	Offshore	Completion	Heterog.	-	Exact; Heur.	2.4.4.2	Prod.
Onwunalu et al. (2008)	Offshore	Drilling	Homog.	Sched.	Simu-Opt.	2.4.4.1	Econ. Index
Davidson et al. (2009)	Offshore	Drilling	Homog.	Sched.	Heur.	2.4.4.1	-
Douro and Lorenzoni (2009)	Onshore	Workover	Homog.	Sched.	Heur.	2.4.2	Prod. Loss
Falex (2009)	Offshore	Drilling	Homog.	Sched.	Heur.	2.4.1	Multi
Glinz and Berumen (2009)	Offshore	Drilling	Heterog.	Sched.	Exact	2.4.1	Econ. Index
Gonçalves (2009)	Offshore	Drilling	Heterog.	Rout.	Heur.	2.4.1	Econ. Index
Litvak and Angert (2009)	Offshore	Drilling; Others	Homog.	Sched.	Heur.	2.4.4.1	Econ. Index
Pacheco et al. (2009b)	Onshore	Workover	Homog.	Sched.	Heur.	2.4.2	Prod. Loss
Pacheco et al. (2009a)	Onshore	Workover	Homog.	Sched.	Heur.	2.4.2	Prod. Loss
Bassi (2010)	Offshore	Workover	Heterog.	Rout.	Simu-Opt.	2.4.3	Prod. Loss
Lorenzoni and Polycarpo (2010)	Onshore	Workover	Homog.	Sched.	Heur.	2.4.3	Prod. Loss
Martin et al. (2010)	Offshore	Drilling	Heterog.	-	-	2.4.4.1	Distance
Mazzini et al. (2010)	Offshore	Drilling; Others	Heterog.	Sched.	Exact	2.4.4.2	Costs
Pacheco et al. (2010)	Onshore	Workover	Homog.	Sched.	Heur.	2.4.2	Prod. Loss
Pandolfi et al. (2010)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.4.2	Time
Al Gharbi (2011)	Onshore	Drilling	Homog.	Rout.	Heur.	2.4.1	Costs
Litvak et al. (2011)	Offshore	Drilling; Others	Homog.	Sched.	Simu-Opt.	2.4.4.1	Econ. Index
Pacheco (2011)	Onshore	Workover	Homog.	Sched.	Heur.	2.4.2	Prod. Loss
Ribeiro et al. (2011)	Onshore	Workover	Homog.	Sched.	Heur.	2.4.2	Prod. Loss
Serra et al. (2011)	Offshore	Drilling; Others	Heterog.	Sched.	Exact	2.4.4.2	Multi
Soares et al. (2011)	Onshore	Workover	Heterog.	Rout.	Heur.	2.4.3	Prod. Loss

Table C.1: Rig scheduling publications found in this literature review

Authors (Year)	Oilfield	Task	Fleet	Rout./Sched.	Method	Section	O.F.
Bassi et al. (2012)	Offshore	Workover	Heterog.	Rout.	Simu-Opt.	2.4.3	Prod. Loss
Duhamel et al. (2012)	Onshore	Workover	Homog.	Rout.	Heur.; Matheur.	2.4.3	Prod. Loss
Ribeiro et al. (2012a)	Onshore	Workover	Heterog.	Rout.	Matheur.	2.4.3	Prod. Loss
Ribeiro et al. (2012b)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.3	Prod. Loss
Sabry et al. (2012)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.3	Costs
Serra (2012)	Offshore	Drilling, Others	Heterog.	Sched.	Exact	2.4.4.2	Prod.
Serra et al. (2012a)	Offshore	Drilling, Others	Heterog.	Sched.	Exact	2.4.4.2	Prod.
Serra et al. (2012b)	Offshore	Drilling, Others	Heterog.	Sched.	Exact	2.4.4.2	Prod.
Serra et al. (2012c)	Offshore	Drilling, Others	Heterog.	Sched.	Exact	2.4.4.2	Prod.
Sumaida et al. (2013)	Onshore	Drilling, Others	Heterog.	Rout.	-	2.4.1	-
Villagra et al. (2013)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.4.2	Time
Bissoli (2014)	Onshore	Workover	Homog.+Heterog.	Rout.	Heur.	2.4.3	Multi
Bissoli et al. (2014)	Onshore	Workover	Homog.	Rout.	Heur.	2.4.3	Costs
Lange and Lin (2014)	Onshore	Drilling, Others	Homog.	Sched.	-	2.4.4.1	Costs
Marques et al. (2014)	Offshore	Workover	Homog.	Sched.	Exact	2.4.2	Rigs Fleet
Omosebi et al. (2014)	Offshore	Drilling	Heterog.	Sched.	-	2.4.4.1	-
Ribeiro et al. (2014)	Onshore	Workover	Heterog.	Rout.	Heur.; Matheur.	2.4.3	Prod. Loss
Zahran and Al-Fardan (2014)	-	Drilling, Others	Heterog.	Sched.	Simu.	2.4.1	Multi
Amrideswaran et al. (2015)	Offshore	Workover, Others	Homog.	Sched.	Heur.	2.4.1	-
Haugland and Tjøstheim (2015)	Offshore	Drilling	Heterog.	Rout.	Matheur.	2.4.1	Econ. Index
Monemi et al. (2015)	Onshore	Workover	Heterog.	Sched.	Heur.; Matheur.	2.4.2	Prod. Loss
Tavallali et al. (2015)	-	Drilling	Homog.	-	Exact; Heur.	2.4.4.1	Econ. Index
Amer et al. (2016)	-	Drilling, Others	Heterog.	Sched.	-	2.4.1	-
Carrilho and Villas Boas (2016)	Offshore	Drilling, Others	Heterog.	Sched.	Exact	2.4.1	Multi
Chowdhury (2016)	Onshore	Drilling	Homog.	Sched.	Heur.	2.4.1	Time
Danach (2016)	Onshore	Workover	Heterog.	Sched.	Heur.	2.4.2	Prod.
Dewan et al. (2016)	Onshore	Drilling, Others	Homog.	Sched.	-	2.4.4.1	Costs
Drouven and Grossmann (2016)	Offshore	Drilling, Others	Homog.	Sched.	Exact	2.4.4.2	Econ. Index
Flager (2014)	Onshore	Drilling, Others	Heterog.	Sched.	Simu-Opt.	2.4.1	Multi
Kromodihardjo and Kromodihardjo (2016)	-	Workover	Homog.	Rout.	Heur.	2.4.2	Prod. Loss
Pérez et al. (2016)	Onshore	Workover	Homog.	Sched.	Exact	2.4.2	Prod. Loss
Silva et al. (2016)	Offshore	Drilling, Others	Heterog.	Rout.	Matheur.	2.4.1	Multi
Tavallali et al. (2016)	-	Drilling	Homog.	Sched.	Exact; Heur.	2.4.4.1	Econ. Index
Arnaout et al. (2017)	Onshore	Drilling, Others	Heterog.	Sched.	-	2.4.1	Time
Bakker et al. (2017)	Offshore	P&A	Heterog.	Rout.	Exact	2.4.4.2	Costs
Kelly et al. (2017)	Offshore	Completion	Homog.	Sched.	Exact	2.4.4.1	Prod.
Vasconcelos et al. (2017)	Offshore	Workover	Heterog.	Sched.	Heur.	2.4.2	Time
Carrilho et al. (2018)	Offshore	Drilling, Others	Heterog.	Sched.	Exact	2.4.1	Rigs Fleet

Table C.1: Rig scheduling publications found in this literature review

Authors (Year)	Oilfield	Task	Fleet	Rout./Sched.	Method	Section	O.F.
Castiñeira et al. (2018)	Onshore	Drilling; Others	Heterog.	Sched.	Data Driven Opt.	2.4.1	Multi
Fernández Pérez et al. (2018)	Onshore	Workover	Heterog.	Sched.	Simu-Opt.	2.4.2	Prod. Loss
Ma et al. (2018)	Onshore	Drilling; Others	-	Sched.	Data Driven Opt.	2.4.1	Econ. Index
Santos (2018)	Offshore	Drilling; Others	Homog.	Sched.	Exact; Heur.	2.4.1	Multi
Silva and Silva (2018)	Onshore	Workover	Heterog.	Rout.	Exact	2.4.3	Prod. Loss
Tavallali and Zare (2018)	-	Drilling	Heterog.	Rout.	Exact	2.4.1	Costs
Achkar et al. (2019a)	Onshore	Workover	Heterog.	Rout.	Matheur.	2.4.4.2	Multi
Achkar et al. (2019b)	Onshore	Workover	Heterog.	Rout.	Matheur.	2.4.4.2	Multi
Bakker et al. (2019)	Offshore	P&A	Heterog.	Rout.	Exact	2.4.4.2	Costs
Pérez et al. (2019)	Onshore	Workover	Heterog.	Sched.	Simu-Opt.	2.4.2	Prod. Loss
Marchesi et al. (2019) (2019)	Offshore	Drilling; Others	Homog.	Sched.	Exact	2.4.4.2	Time
Shaji et al. (2019)	Onshore	Workover	Heterog.	Rout.	Heur.	2.4.3	Prod. Loss
Aurachman et al. (2020)	Onshore; Offshore	Workover	Homog.	Rout.	-	2.4.4.2	Prod. Loss
Bakker (2020)	Offshore	P&A	Heterog.	Rout.	Exact; Simu-Opt.	2.4.4.2	Costs
Bakker et al. (2021)	Offshore	P&A	Heterog.	Rout.	Exact	2.4.4.2	Costs
Calderón and Pekney (2020)	Offshore	Drilling	Homog.	Sched.	Exact	2.4.4.1	Econ. Index
Kulachenko and Kononova (2020)	Onshore	Drilling	Homog.	Rout.	Matheur.	2.4.1	Distance
Tozzo et al. (2020)	Onshore	Workover	Heterog.	Rout.	Heur.	2.4.3	Multi

Table C.1: Rig scheduling publications found in this literature review

For more details about each paper, see the complete table in the supplementary file *SupplementaryFile-IuriSantos.xlsx*.

D
Portuguese stopwords

The following table D.1 presents the complete list of the Portuguese stopwords used for the data cleaning process.

"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 D.1: Stopword (in Portuguese) removed from the text.

E
Comparision between k-means and h-cluster

A comparison of the clusters obtained through k-means and h-clusters was used using heatmaps and dendrograms. This comparison was developed with an arbitrary number of clusters and supported the choice of the k-means as our classification algorithm and is shown in Figure E.1.

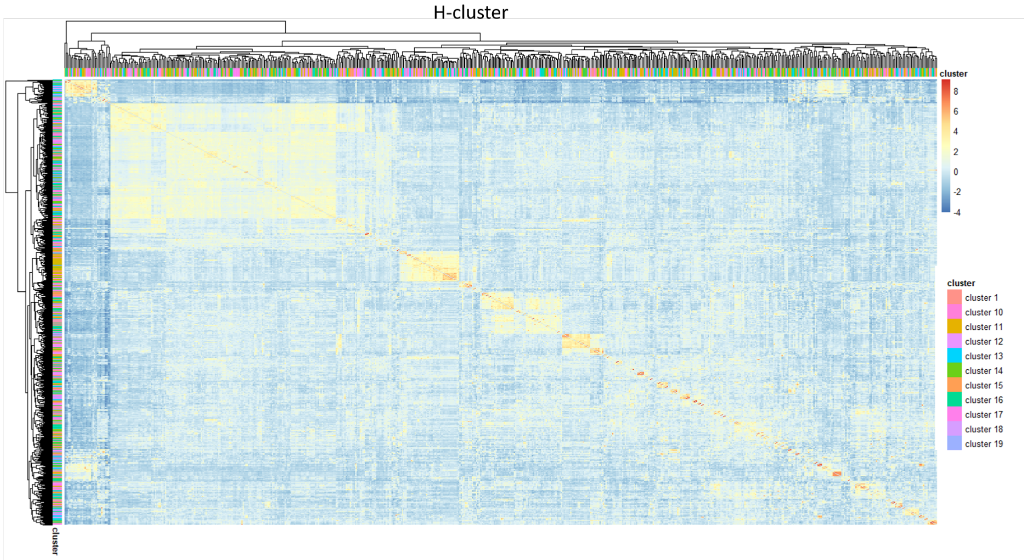


Figure E.1: Heatmaps with dendrogram for k-means (vertical axis) and h-cluster (horizontal axis).

F
Final regression

The final coefficients for the regression of the workover duration are presented in Table F.1. Note that the null coefficients values were ommited from the table.

Parameter	Coefficient
(Intercept)	2.250047e+00
$Depth_i$	-9.305728e-06
$Subpool_{AGP}$	-6.981267e-02
$Subpool_{CORP}$	4.968064e-01
$Subpool_{DPT}$	1.212296e-01
$Subpool_{SSE}$	4.431450e-03
$Basin_{Ceara}$	-2.017823e-02
$Basin_{Espirito}$	-4.528720e-01
$Basin_{Pontiguar}$	2.071196e-01
$Basin_{Sergipe}$	-4.644360e-02
$Clusters_2^{45}$	5.103918782
$Cluster_2^{45}$	6.607015e-02
$Cluster_3^{45}$	-1.754563e-01
$Clusters_4^{45}$	5.203951e-02
$Clusters_7^{45}$	3.480876e-02
$Clusters_8^{45}$	1.103593e-01
$Clusters_9^{45}$	3.791715e-02
$Clusters_1^{45}0$	8.499537e-02
$Clusters_1^{45}3$	-2.468064e-01
$Clusters_1^{45}7$	-2.193741e-01
$Clusters_2^{45}1$	-1.551534e-01
$Clusters_2^{45}5$	3.278956e-01
$Clusters_2^{45}8$	-1.192809e-01
$Clusters_2^{45}9$	1.332856e-01
$Clusters_3^{45}1$	-2.830062e-02
$Clusters_3^{45}3$	4.813594e-02
$Clusters_3^{45}6$	-1.515326e-01

Table F.1: Coefficients for the final regression of $log(d_i^k)$.

Parameter	Coefficient
$Clusters_3^{45}9$	1.102441e-01
$Clusters_4^{45}0$	2.317509e-01
$Clusters_4^{45}2$	-4.002683e-01
$Clusters_4^{45}3$	-8.011800e-02
$Type^{FixedRig}$	5.857882e-01
$Type^{SS/Drillship}$	7.656649e-01

Table F.1: Coefficients for the final regression of $log(d_i^k)$.

G
Kernel-based joint chance-constrained model

This Appendix applies the kernel-based data-driven joint chance-constrained optimization approach from Calfa et al. (2015), described in Section 3.3.6, on the chance-constrained workover rig scheduling problem from Section 7.1.1. First, we reformulate the probabilistic constraint from (7-7) can be reformulated as:

$$\mathbb{P}[S_i - S_j + H(X_{ij}^k - 1) \leq -\tilde{d}_i^k \qquad \forall j, k | i \neq j] \geq \alpha \quad \forall i | i \neq 0 \tag{G-1}$$

$$\mathbb{P}[g_{ij}^k(x) \leq \tilde{\xi}_i^k \qquad \forall i, k | i \neq j] \geq \alpha \quad \forall i | i \neq 0, \tag{G-2}$$

where $g_{ij}^k(x) = S_i - S_j + H(X_{ij}^k - 1)$ and $\tilde{\xi}_i^k = \tilde{d_i}^k$.

According to Calfa et al. (2015), if the distributions for $\tilde{\xi}_i^k$ are independents, uncorrelated, and each one following a Gaussian model, constraint (G-2) can be reformulated using kernel distribution estimation properties as:

$$\sum_{l=1}^{L} \prod_{j \in J, k \in K | i \neq j} \left[\mathcal{K}_i^k \left(\frac{g_{ij}^k(x) - \hat{\xi}_i^{kl}}{h_i^k} \right) \right] \geq \alpha_+^{'} \qquad \forall i \in J | i \neq 0 \tag{G-3}$$

$$\sum_{l=1}^{L} \prod_{j \in J, k \in K | i \neq j} \left[\mathcal{K}_i^k \left(\frac{g_{ij}^k(x) + \hat{d}_i^{kl}}{h_i^k} \right) \right] \geq \alpha_+^{'} \qquad \forall i \in J | i \neq 0, \tag{G-4}$$

where $\hat{\xi}_i^{kl}$ are data points of the uncertainty, $\hat{d}_i^{kl}$ represents the data points l of the duration of the workover i using rig k, h_i^k is the bandwidth selected for the kernel estimation of that type of well of the workover i and rig k, and $\mathcal{K}(\cdot)$ is the kernel or weighting function estimated for that type of well of the workover i and rig k. In this case study, as recommended by Calfa et al. (2015), we use the Gaussian kernel $\mathcal{K}_{Gaussian}(u)$, which leads to the following equations:

$$\sum_{l=1}^{L} \prod_{j \in J, k \in K} \left[\frac{1}{2} + \frac{1}{2} \text{erf} \left(\frac{g_{ij}^k(x) + \hat{d}_i^{kl}}{h_i^k \sqrt{2}} \right) \right] \geq 1 - \alpha_+^{'} \qquad \forall i \in J | i \neq 0 \tag{G-5}$$

$$\sum_{l=1}^{L} \prod_{j \in J, k \in K} \left[\frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{S_i - S_j + H(X_{ij}^k - 1) + \hat{d}_i^{kl}}{h_i^k \sqrt{2}}\right)\right] \geq 1 - \alpha_+^{'} \quad \forall i \in J | i \neq 0 \tag{G-6}$$

As mentioned earlier, this deterministic-equivalent reformulation leads to a non-linear equation that is also non-convex. As a result, the kernel-based JCC-WRSP becomes a non-convex mixed-integer non-linear programming (MINLP) model. Therefore, other alternative methods should be developed. In the next section, we propose alternative data-driven JCC-WRSP method based on regression models.