Using Business Intelligence to leverage field workforce capacity planning in Telco

Salomé Pinto de Meneses Monteiro

Master's Dissertation

Supervisor at FEUP: Prof. Bernardo Almada Lobo
Supervisor at NOS: João Miguel Santos Figueira

U. PORTO

FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

Master in Industrial Engineering and Management

2021-06-28

Abstract

When it comes to operations management in service industries, the focus tends to be placed on the operational optimization of the resources available by assigning them to the tasks scheduled. This work, within the business context of a telecommunications company, focused on working with the available data, through Business Intelligence tools, at a higher level of planning. The goal is to introduce tools and models focused on exploring technician distribution and supporting capacity planning decisions.

To this end, the project developed can be divided into three core stages.

The first one consisted essentially of database modeling, using MySQL, and focused on shedding light on the distribution of resources – both in geographical and time dimensions – so that current available capacity could be understood. Furthermore, it joined this information with the records of interventions performed, to understand the balance between the available and occupied capacity, that is to say, the utilization rate of the resources.

The second one aimed at incorporating a forecasting component, required for a full planning process in which the capacity available should meet projected demand. The foundation of a Machine Learning model was built using the Random Forest algorithm. However, with accuracy results indicating a Mean Absolute Percentage Error of 55,91%, it still has considerably large room for improvement, particularly when predictions are made for areas with lower population density, as the model performed better when predictions targeted urban areas.

Finally, the project was delivered to its key clients – the responsible persons for the planning process, in negotiation with the outsourcing partners that supply the workforce – through a simulation tool, developed in Microsoft Excel, that allows for the development of scenarios regarding three different levers that can impact responsiveness: fluctuations in future demand, alterations in the geographical distribution of the workforce and the establishment of service level performance targets.

Although the implementation of this support tool in the work processes of the operations management team should be evaluated in the long-term, it is expected that it will constitute a favorable improvement in the capacity planning decisions made in the company, as it provides information that was not available before, at the same time as it incorporates a structured path in the decision making processes.

Resumo

No que diz respeito à gestão de operações aplicada aos serviços, verifica-se um foco generalizado na otimização operacional do processo de alocação dos recursos disponíveis às tarefas planeadas. Este projecto, desenvolvido no contexto de uma empresa de telecomunicações, foca-se, por outro lado, na utilzação de ferramentas de Business Intelligence para auxiliar o planeamento tático da distribuição da força técnica. O objetivo é que tal seja conseguido através do desenvolvimento de modelos e ferramentas direccionados ao apoio de decisões relacionadas com o planeamento de capacidade.

Para tal, o estudo desenvolvido foi dividido em três fases principais.

A primeira fase cingiu-se à modelação da base de dados da empresa, através da linguagem de MySQL, com o objetivo de dar visibilidade à situação atual de distribuição da capacidade disponível, tanto na sua componente geográfica como temporal. Esta informação foi, então, compilada com os registos de intervenções executadas, de modo a ilustrar também a comparação entre capacidade disponível e ocupada. Isto resultou, assim, na modelação da taxa de utilização dos recursos.

Posteriormente, a segunda fase do projeto teve por objetivo a incorporação de uma componente preditiva no mesmo, uma vez que um processo de planeamento de capacidade holístico deve-se assegurar que existe uma equivalência entre a procura projetada e a capacidade disponível. A construção de um modelo de *Machine Learning* foi iniciada, através do algoritmo *Random Forest*. No entanto, com um Erro Médio Absoluto Percentual (MAPE) de 55,91%, foi concluído que a margem de melhorias do modelo era ainda grande – particularmente em áreas de menor densidade populacional, uma vez que a performance do modelo foi consideravelmente melhor apenas em áreas urbanas.

Por último, o projeto teve como clientes finais os responsáveis pelo processo de planeamento, que o realizam através de negociação com os parceiros estratégicos que fornecem a mão-de-obra técnica. A eles foi, então, entregue uma ferramenta de simulação desenvolvida em Microsoft Excel e suportada no modelo descrito na primeira fase, que permitia que cenários fossem simulados relativamente a oscilações na procura prevista, alterações na distribuição geográfica dos técnicos ou estabelecimento de objetivos de responsividade na realização das intervenções.

Embora a implementação desta ferramenta de apoio nos processos de trabalho da equipa de gestão de operações deva ser algo cujo sucesso é avaliado a longo prazo, pode-se afirmar que constituiu um desenvolvimento favorável, no sentido em que dotou os envolvidos de informação nova, ao mesmo tempo que estruturou o processo de decisão suportado na mesma.

Acknowledgments

This appreciation is directed at everyone that, in the most diverse ways, has contributed to this project or to my personal and professional development to this point.

First, I would like to thank everyone at NOS for this opportunity. Particularly to my supervisor João Figueira, for the time, support and friendship put into welcoming me into his team and allowing this project to grow. To him and to everyone in the I&M Department who I got to meet throughout this journey: It was not easy to adjust to this dissertation experience in a purely remote environment, but I know it would have been much harder, had it not been for your good spirits and willingness to help.

This work represents the wrap-up of five years in FEUP and, although this journey had its ups and downs (as all do), I am so grateful for everyone who had something to teach me along the way. To my professors, and especially to my supervisor Prof. Bernardo Almada Lobo, thank you for the availability to support me and for all the wisdom you shared with me during this period.

And because reflecting on my journey would not make sense without mentioning one of the biggest parts of it: to ESTIEM and to everyone I got to meet through this network, thank you for the memories and knowledge shared and for teaching me that the most incredible experiences are just one bold decision away.

To all my friends, the ones I made in university and I know I can count on wherever I go next, and the ones I was smart enough to choose many years ago and have not left my side since (even when life gets busy and a few months go by in between coffee dates).

To Nina, who throughout our university lives, and especially in the past few months, has been there whenever the mood calls for a work partner or for a glass of wine and a Netflix episode.

And finally, my wholehearted thank you to my father, Manuel, my safest supporter, to my twin brother, Rodrigo, my first and best friend no matter what, and to Flor, my chosen family.

Contents

Acronyms and Symbols

AID	Average Intervention Duration
AIT	Average Intervention Time
B2B	Business-to-Business
B2C	Business-to-Customer
BI	Business Intelligence
BI&A	Business Intelligence and Analytics
BGO	Background Optimization
CRISP-DM	Cross Industry Standard Process for Data Mining
CVA	Cross-Validation Accuracy
DSS	Decision Support Systems
DTH	Direct-to-Home
DW	Data Warehouse
ETL	Extraction, Transformation and Loading
FTTH	Fiber-to-the-Home
HFC	Hybrid Fiber-Coaxial
I&M	Installations and Maintenance
IT	Information Technology
KPI	Key Performance Indicator
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
RDBMS	Relational Database Management Systems
RMSE	Root Mean Squared Error
SLA	Service Level Agreement
SP	Service Provider
Telco	Telecommunications Company
TRSP	Technician Routing and Scheduling Problem
UML	Universal Modeling Language
VRP	Vehicle Routing Problem
WO	Work Order
WMAPE	Weighted Mean Absolute Percentage Error
WSRP	Workforce Scheduling and Routing Problem

List of Figures

List of Tables

Chapter 1

Introduction

In this first chapter, the focus is on presenting briefly the motivation and scope of the problem tackled, as well as the business context in which the analysis took place.

1.1 Project Motivation

The problem tackled by this project relates to the planning of the field workforce capacity of a Telecommunications company. Although, there is a vast literature on the operational optimization of field technicians assignment to tasks, the knowledge on how to plan the distribution of these resources is still very sparse.

Within the industry context here presented, there is, also, a layer of complexity added by the fact that the technician workforce is generally subcontracted, and the strategic partners have conflicting performance goals driving their tactical planning: while the telecommunications company, ideally, wants the best responsiveness possible – as it directly affects Customer Satisfaction – the partners value that their labor force has a high utilization rate so there is no waste of resources.

Thus, this project focuses on using Business Intelligence tools to track the evolution of these opposite metrics, based on the workforce made available in each geographical region. The goal is to empower the persons in the company involved in capacity planning decisions, with the right information, so that it can be used when monitoring the performance and negotiating with the Service Providers where technicians should be placed.

In the case of this project, the information deemed relevant to be treated and delivered to these stakeholders consists of, both a more thorough reporting of current performance, through the calculation of new performance metrics, but also an additional component of capacity planning that was identified as missing: the forecast of future job demand and its impact on future labor force requirements.

This project, thus, aims to tackle capacity planning for field operations from both the lens of Business Intelligence reporting and predictive modeling, establishing a basis on how these two dimensions can be incorporated into the decision making process.

1.2 NOS and the Telecommunications Market

The telecommunications company that is the target of this study is "NOS SGPS", from hereon referred to as just NOS, which is one of the biggest players in this sector in Portugal. A brief description of the company and its market context is, then, here presented.

The NOS brand was launched in 2014. This was the culmination of the merger process between Zon Multimedia and Optimus Telecommunications, that had been initiated the previous year. The goal of this alliance was to adjust the value both companies were providing by integrating their offers of fixed and mobile solutions for television, Internet and telephone, which are now offered as service packages known as *multiple play*.

Currently, the NOS Group is the full-owner of ten businesses, present mainly in Portugal and Portuguese speaking countries. These include services such as Cinemas and Publicity, although this study focuses on its core business and the one it is mostly known for: communications. Within this sector the company tackles the Residential market segment, in a Business-to-Customer (B2C) approach, as well as the Business-to-Business (B2B) market segments, divided into Corporate and Mass Business, depending if the solutions required are, respectively, more or less complex.

Within the Portuguese telecommunications (Telco) market, NOS is among the top three players that, together, are responsible for over 95% of the national *multiple play* subscriptions, according to APRITEL (2019). This same study shows that pricing for telecommunications in Portugal is one of the lowest compared to the European average and that the top three competitors all have their offers priced similarly. This means that competition in this market is happening mainly through technological differentiation and customer relationship management.

1.3 Project Scope and Stakeholders

This project is integrated within the Business Intelligence team for operations relating to Installations and Maintenance (I&M), Logistics, Technical Support and Terminal Management. Among other responsibilities, this is where field operations are managed, which represent the key focus area of this study.

NOS, currently, subcontracts its technical workforce to multiple Service Providers (SP). These partners are responsible for the recruitment and training of the technicians that will be used by NOS in the multiple operations that require direct intervention at the residence of the customer, or at the company site, in the case of a business client.

Between 2019 and early 2020, NOS has managed to centralize the operational management of this subcontracted workforce, by adopting a new software: Click Field Service Edge, commonly designated as just *Click*, is a platform provided by ClickSoftware that allows for a direct optimization of the jobs assigned to each technician. Among other functionalities, *Click* runs recurrently an optimization algorithm that given the amount of technicians available and the number of jobs scheduled for a certain time period, establishes the optimal routes for each technician.

This project is, therefore, based on the information collected from this software regarding the technicians available and the interventions occurring and aims at using this data to support the capacity planning process and decisions within NOS by increasing the amount of information accessible by those involved in the direct negotiation with SPs. These internal stakeholders and their accountability levels will be further detailed in Chapter 3.

As for the field activities covered by this project, it is important to note, that these are solely the ones that work under the assumptions here explained. This means that field interventions that are neither outsourced nor providing a direct service to the client are considered to be out of the scope of this problem as they would require different operational targets.

1.4 Key Objectives

In a more concrete manner, and based on the aforementioned, it can be established that the main goal of this project is to create the necessary tools and data infrastructure to help NOS with its field workforce management, by presenting information in a way that will shed light on its geographical situation in terms of technician supply and demand.

Three key objectives can, then, be set as the fundamental dimensions that should be covered by this project so its core goal can be achieved:

1. Improving the data infrastructure to allow reporting on the utilization and idle time of field technicians.
2. Incorporating the forecasting of future demand for interventions in the capacity planning process.
3. Creating a tool that will support the stakeholders of the capacity planning process in their decision making.

1.5 Project Methodology

The problem stated and the proposed solution require the application of techniques heavily connected with data analysis and the direct usage of a data mining technique for predictive analytics. For this reason, the methodology used as a reference is the Cross Industry Standard Process for Data Mining (CRISP-DM).

This process is described in Olson and Delen (2008) and it consists of six key stages – Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment – as shown in Figure 1.1.

The framework can, also, be visualized as integrated with the three previously presented objectives with different stages meeting each.

Nonetheless, the project will require the construction of more than one data model to achieve its objectives and, therefore, this iterative cycle is taken as a general guideline for each of its stages, which will be more thoroughly explained in Chapter 4.

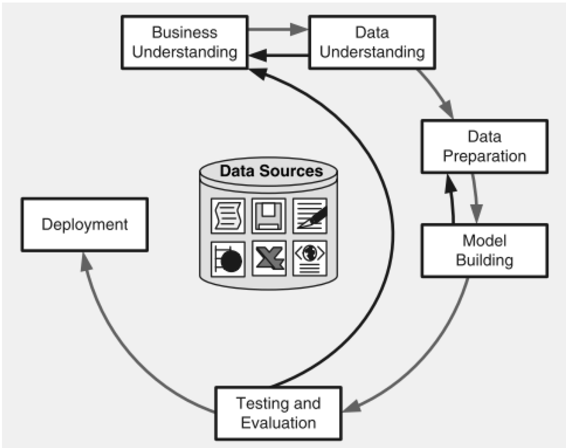

Figure 1.1: CRISP-DM Methodology. Source: Olson and Delen (2008).

1.6 Document Structure

This dissertation is divided into six main chapters.

Chapter 1, here presented, consists of a general summary of the case study tackled. The framework and goals of the project can also be read here.

Chapter 2, focuses on presenting a Theoretical Framework, presenting the academic concepts used for this approach, as well as the state-of-art research done through them. Furthermore, an analysis is provided on how these concepts have been applied in the Telecommunications industry.

Chapter 3 consists of a more thorough study on the present context of the company, particularly on the operations here tackled. It is, also here, that can be seen a summary of the available data before its modeling is delved into. Using the CRISP-DM methodology, this chapter can be perceived as both the Business and Data Understanding stages.

Chapter 4 is where the project implementation is presented, in its main three components – each of them connecting directly to one of the above cited objectives.

The results and main outcomes from this implementation, as well as some of its implications and potential limitations, are presented and analyzed in Chapter 5.

Finally, Chapter 6 will be, based on these results, reflecting on what was achieved and proposing further work that can be developed on the topics presented.

Chapter 2

Theoretical Framing

In this chapter the state-of-art regarding topics associated with this study can be found, starting with the problem under study: resource capacity planning, and, afterward, delving deeper into some of the Analytics tools that will be required for the solution proposed.

Some general remarks and examples of the connection between Business Intelligence and Analytics topics and Telco industry applications is also reflected upon, together with recent work developed in this field.

2.1 Technician Capacity Management

Resource planning can be layered in strategical, tactical and operational decision making levels, as summarized in Figure 2.1. Within applications in Telco, for which the resource subject to planning should be considered the technicians, Kassem et al. (2012), establishes the difference between these levels by stating that, strategical decisions consist of a high level match between workforce available and expected demand, while tactical planning is more detailed, as it already focuses on understanding which technicians are available and what interventions they have the skills to perform.

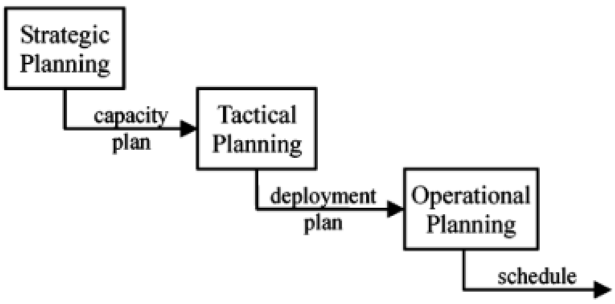

Figure 2.1: The stages of resource planning. Source: Kern et al. (2009)

As for operational planning, although not the focus of this study, it is where the most literature is available and consists of the direct assignment of jobs to the technician that should perform them, through an optimization perspective. This optimization and the studies being produced on it will be further delved into in the Technician Routing and Scheduling Problems section. This stage of the workforce management process is also being subject to a great degree of optimization where software platforms mostly handle it directly (Kassem et al. 2012). The challenge is, then, how to make these platforms reflect accordingly the real-world scenarios and manage their uncertainty.

2.1.1 Tactical Resource Planning

Service chain management differs from supply chain management, as, unlike physical goods produced, intangible services are always heterogeneous in quality as they require direct interaction between employee and client. This variability becomes more significant when diverse services occur, something that is a strong characteristic of the Telco industry (Kern et al. 2009).

As stated by Mohamed et al. (2012), resource planning is a fundamental part of Service Chain Management, in which the target should be to ensure the necessary capacity to meet service level is maintained at the lowest possible costs.

As previously mentioned, within Telco applications, the resources subject to planning are the field technicians. Akkermans and Vos (2003) asserts as one of the particular challenges of this industry the low flexibility when implementing reserve capacity, due to the fact that technicians have to be hired and trained ahead, which puts more stress on the planning process. The same study – which focused on the analysis of amplification effect (equivalent to bullwhip effect) in service chains – found that information sharing regarding sales projections and commercial campaigns could yield improvements for operations management performance.

As for studies focused on the development of solutions for resource planning in Telco, it is worth highlighting Kern et al. (2009), Kassem et al. (2012) and Mohamed et al. (2012), as some interesting automated solutions to the issue. All of this state-of-art literature was, however, produced in the particular context of British Telecom and, thus, is adjusted to the workforce planning systems of the company, which can be seen as a limitation to the research developed in this field.

Both the later enumerated studies include a user-defined set of rules as to what prioritization to set in the planning process. These inputs, in the analysis done by Mohamed et al. (2012) can be counted as:

1. Distance;
2. Area Preference;
3. Skill Preference;
4. Duration of the tasks;
5. Importance of the task;
6. Task density;
7. Expected working time of the technician;
8. Distance between the technician at a certain point in time and their finish location (their home).

2.1.2 Technician Routing and Scheduling

Also referred by some authors, such as Castillo-Salazar et al. (2016), as Workforce Scheduling and Routing Problem (WSRP), the Technician Routing and Scheduling Problem (TRSP), as described by Pekel (2020), can essentially be described as an optimization problem where the goal is to have a set of technicians carry out a number of tasks at a minimum cost while meeting certain resource constraints. It consists of an extension of the Vehicle Routing Problem (VRP) in combination with classical scheduling problems (Castillo-Salazar et al. 2016).

Several real-world applications have been studied, with different particularities to each, from which a few will be here cited. Decerle et al. (2017) proposes a general multi-objective model to be used in the optimization of home health care services. Here, constraints associated with the workforce expertise become particularly relevant. Other variants of the problem include the patrol scheduling for security purposes, applied by Misir et al. (2014) to the monitoring of a network of train stations, as well as the maintenance of wind turbines, studied by Froger et al. (2018). The latter, similarly to what happens in the case here in study, includes the additional challenge of managing an outsourced workforce and the need to balance stakeholders with different objectives: while the service owner is interested in high responsiveness, the maintenance is subcontracted to a partner whose interest is in minimizing its costs.

The telecommunications industry, in particular, presents itself as one of the drivers of research when it comes to this problem. In 2007, the French Operations Research Society proposed an approach to the TRSP as their yearly challenge, with a set of real-world datasets provided by France Telecom. This research topic was inspired by the work initially presented by Dutot et al. (2006), where the French company suggests its first draft of what should be an analytical system to support allocation decisions that, up until that point, were done manually by the supervisor in charge of a set of technicians. While this challenge included a strong constraint that is not always applicable - that tasks are performed by a team rather than a single technician and, as a team is formed, it must stay together for the duration of the workday - it, nonetheless, provides a good basis to explain the general factors that influence the problem within this context.

The problem definition presented in the aforementioned challenge is recovered by Cordeau et al. (2010), explaining that a set of N tasks, which are characterized by, among other factors, a duration (d) and a priority level (p) and a set of required skills. These should, then, be matched to a set of T technicians that are characterized by a *skill vector* defining the skills they are proficient in. The routing aspect of the problem is, here, simplified by the addition of an outsourcing cost factor (c) associated with each task. However, more recent studies, such as the proposed solution presented by Pereira et al. (2020) incorporate the geographical component, with each intervention being associated with a vertex V, and distance between vertexes taken into account when assigning consecutive tasks. This article also focuses on the possibility of including precedence relationships between tasks as a constraint.

From this basis, other factors can be explored that will increase the performance of these models in real-world settings. Specifically, some of the promising work recently developed in TRSPs consists of attempts to reflect human psychology effects in the models developed. Chen (2016) introduces the effect of learning as a phenomenon that leads to a gradual increase in technician productivity. Similarly, L. van Eck et al. (2017) explores the effect some tasks may have on the overall productivity of a technician, as they may be more tiring or stressful and, therefore, have a negative impact on the performance of the technician they are assigned to.

2.2 Business Intelligence and Analytics

According to Philips-Wren et al. (2021), while the term has been in use since the 1990s, there is not, to this day, a single, clear, definition of what Business Intelligence (BI) is. Rather, the concept of Business Intelligence and Analytics (BI&A) – which can be aggregated into one, as a competitive BI system does incorporate a strong analytical component (Davenport 2005) – is generally used to define all capabilities of converting large volumes of data into insights and knowledge used in decision-making. Arnott et al. (2017) explains that BI, in an organizational context, "is often used as the umbrella term for large-scale Decision Support Systems (DSS)" and currently represents the largest Information Technology (IT) investment area for enterprises.

As for its historical evolution, Chen et al. (2012) defines three main generations for BI&A systems. The first, BI&A1.0, originated from an evolution in Database Management Systems and encompassed the concepts of Extraction, Transformation and Loading (ETL) of information, as well as reporting performed through statistical analysis and intuitive visualizations. These systems rely on Relational Database Management Systems (RDBMS) and are still often used by enterprises to monitor and manage business performance. BI&A2.0, on the other hand, arose at the beginning of the millennium with the upsurge of the generalized use of the Internet and is characterized by Web-based data collection and unstructured databases. BI&A3.0 incorporates another layer of complexity with data being collected from mobile devices and sensors.

Based on their organizational scope, BI systems can be classified as either enterprise BI or functional BI. The first consists of a combination of data that is relevant across many organizational areas and is, therefore, a much more complex system developed by the IT team of the organization. As for functional BI systems, they tend to serve a single business unit, thus, being of their own, internal accountability (Arnott et al. 2017).

2.3 Machine Learning Forecasting Techniques

In general terms, predictive analytics consists of any technique used with the objective of making predictions (Han et al. 2012). This can be done through classical statistics, using regression models, however, with the introduction of data mining algorithms, a much broader set of possibilities is now available, as these were able to incorporate hundreds of predictors within a model, something that would have been far too complex for multiple regression models (Thelen et al. 2004).

Pavlyshenko (2019) enhances the fact that supervised machine learning algorithms allow for complex patterns in a time series to be captured. However, the same work also mentions that one of the key complexities of time series analysis is how accuracy is highly dependant on the time period for which historical data is available, and its consequent ability to record seasonal patterns.

2.3.1 The Data Mining Process

There is a strong connection between BI technologies and data mining, which will be explored throughout this study. Han et al. (2012) explains that effective BI systems, that allow an enterprise to retrieve meaningful information about its operations, should be able to present this information in three time dimensions: historical, current and predictive. This last component is where the two terms converge, as BI analytical tools should rely on data mining predictive techniques.

As for its definition, Bose and Mahapatra (2001) write that "Data Mining (...) is the process of discovering interesting patterns in databases that are useful in decision making". Although this definition is relatively broad, the same article thoroughly explains the connection between this concept and Machine Learning, stating that machine learning techniques, as computational tools designed to automate the learning process, are fundamental for analyzing data and discovering patterns, and, therefore, play a major role in the development of data mining applications. The added value, as reaffirmed by Xu et al. (2021) lays in the ability these techniques have of saving time and computational resources when large amounts of data are being processed.

Fayyad et al. (1996) initially presented the Knowledge Discovery in Databases (KDD) process, which can be seen represented in Figure 2.2.

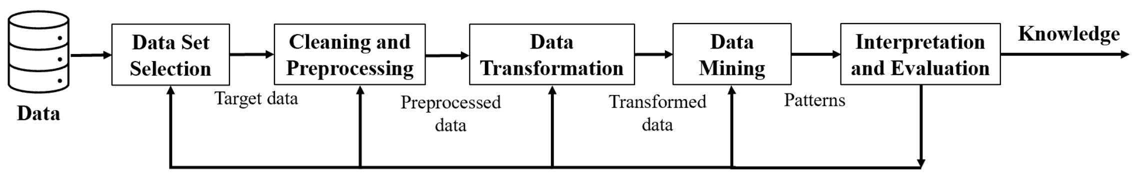

Figure 2.2: The KDD process. Adapted from: Fayyad et al. (1996)

In this framework, and in the explanation for its creation given by Fayyad et al. (1996), data mining is seen as a phase of the process and constitutes exclusively the application of tools to extract patterns from data. The distinction between the term designating specifically this stage or the full approach is, thus, often confused in the literature (Azevedo and Santos 2008). However, as mentioned in Olson and Delen (2008) "In order to be applied successfully, the data mining solution must be viewed as a process rather than a set of tools or techniques" and, therefore, one can view the KDD process as an initial proposal of a data mining framework.

In an attempt to reach more consensual guidelines for the implementation of data mining applications, standard processes have, since, been suggested. The most popular ones and the ones more often studied in the literature are CRISP-DM and SEMMA (Azevedo and Santos 2008).

CRISP-DM, as shown previously in Figure 1.1 consists of six key stages and aims at providing an industry standard for data mining. In turn, SEMMA stands for an acronym for Sample, Explore, Modify, Model, Assess.

All three mentioned framework models are designed to be driven by iterative experimentation, meaning that once the Assessment/Evaluation stage is reached, improvements should be made by going back to the initial phases and refining the data (Olson and Delen 2008). A comparison between them was made by Azevedo and Santos (2008), and the conclusion reached is that the phases in each can be directly equaled. Therefore, they do not constitute opposing proposals but rather complementary ones. Nonetheless, CRISP-DM may be considered more thorough as it includes the Business Understanding stage, as shown in Figure 2.3.

KDD	SEMMA	CRISP-DM
Pre KDD	-------------	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	-------------	Deployment

Figure 2.3: Comparative study of Data Mining Processes. Source: Azevedo and Santos (2008)

2.3.2 Random Forest Algorithm

The Random Forest algorithm was initially presented by Breiman (2001), as part of a series of new studies on ensemble supervised learning algorithms. From this first study, the technique has always been adaptable to both categorical responses, known as classification, and continuous responses, otherwise named regression.

Supervised learning algorithms are defined as the ones where the data is trained by fully labeled data, and, in the case of regression, these labels correspond to numerical values (Romero et al. 2019). In turn, ensemble algorithms, as defined by Han et al. (2012), who focus their work on the application of models for multi-class classification, consist of a combination of base classifiers that all contribute to the final prediction. The tendency is that the ensemble will be more accurate than its base models as the effect of any errors inflicted by one of the models will be minimized by the majority. Figure 2.4 illustrates precisely this process.

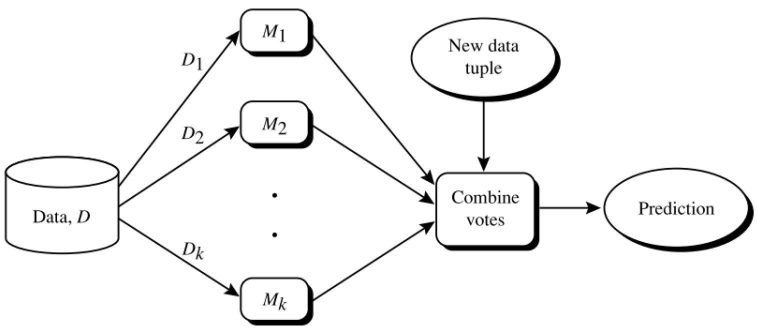

Figure 2.4: How ensemble methods increase Classification Accuracy. Source: Han et al. (2012).

As for this particular algorithm, in its introduction, Breiman (2001) explains that "Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.". This, essentially means that the ensemble process, here, occurs through a combination of base algorithms known as decision trees.

In turn, these decision trees consist of structures comparable flowcharts, where each node tests a particular predictor variable and performs a binary partition (Cutler et al. 2012). These predictor variables can be both continuous values, for which a split point (c) is defined, or categorical ones, where the test performed consists of checking if the value belongs to a subset (S) of the possible categories. These two possibilities are illustrated in Figure 2.5. The divisions, then, run recursively, until a stopping criterion is met and the prediction value is obtained from the terminal nodes (Cutler et al. 2012).

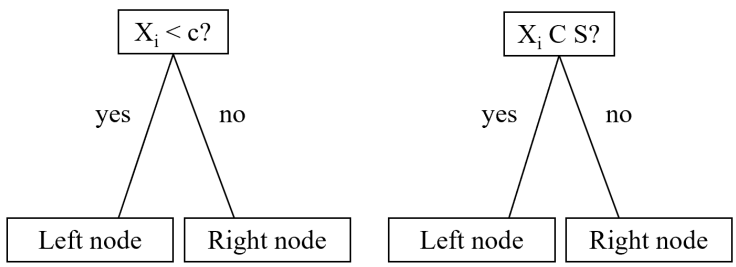

Figure 2.5: Visual representation of decision tree node partitions for continuous or categorical predictors. Adapted from: Cutler et al. (2012).

The key difference between classification and regression random forest is, too, explained by Breiman (2001), when presenting the latter. While in classification algorithms, the end result obtained in each terminal node of the trees is a particular class, for regression algorithms it consists of a numerical value. From there, in classification, the prediction obtained is given by a majority vote of all trees (as previously illustrated in Figure 2.4), while for regression the prediction emerges from calculating the average value of the predictions resulting from all trees.

Romero et al. (2019) further summarizes the algorithm in three key steps:

1. A set of bootstrap samples is picked from the original training set.

2. A decision tree is built for each of these sub-samples and a random subset of cuts is chosen from each of them.

3. A prediction is obtained through the aggregation of all individual predictions.

As an algorithm designed to increase prediction accuracy, Lingjun et al. (2018) highlights the advantages of Random Forest, stating that, not only is this algorithm suitable for many forms of data with different volumes, but it also requires fairly low computational costs, while producing results that are highly accurate and easily interpreted. Furthermore, its models remain successful in situations where the number of predictors is large in comparison to the sample size – a situation that could constitute a problem in the implementation of other algorithms. These advantages are also confirmed by Cutler et al. (2012), from who it can be added that, while individual decision trees would normally have to be subjected to pruning, in order to prevent overfitting the model, such process is not necessary for Random Forest and the risk and effects of overfitting are relatively small.

In a comparative study, Vairagade et al. (2019) applied Random Forest to demand forecasting for applications relating to Supply Chain Management, using as predictors a combination of time, pricing, geographical and other information. The results were compared to the ones generated through a different Data Mining technique – in this case, Ketas Neural Networks – and the conclusion was that Random Forest proved to be the most efficient. Similar conclusions were also achieved by Romero et al. (2019) and Brentan et al. (2017).

2.3.3 Measuring Accuracy in Forecasting Models

Different methods of evaluating the performance of Machine Learning algorithms can be established. From these, three will here be highlighted, as they can all be appropriate for the evaluation of regression algorithms, and, in particular, of forecasting models.

First **Holdout**, as mentioned in Olson and Delen (2008), consists of a simple split of the data available into a train and a test set. Typically the proportions used are two thirds of the data used for training and the rest for testing. The train set is then used for the model fitting and, subsequently, the model is applied to make predictions for the test set. These predictions are compared with the real values and error metrics can be calculated.

From these metrics, it is worth highlighting the use of Mean Absolute Error (MAE), which consists of the average difference between the real value (y_i) and its prediction ($\hat{y}_i$) for all N observations. This is the metric used for the conclusions drawn in many of the Random Forest analysis above-cited, such as Romero et al. (2019).

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{2.1}$$

Alternatively, Root Mean Squared Error (RMSE) is also frequently used. However, according to Hyndman and Koehler (2006), this metric is particularly sensitive to outliers and, therefore, may not be the best indicator for the performance of a forecasting model.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{2.2}$$

Both MAE and RMSE are only applicable in situations where all the data sets available are on the same scale as the values provided can only be compared if that is the case (Hyndman and Koehler 2006).

For this reason, complementary to the first one, one can use the Mean Absolute Percentage Error (MAPE) to aid the analysis of results as, unlike the previous, its performance does not depend on the magnitude of the figure subject to forecast.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{\hat{y}_i} \tag{2.3}$$

Additionally, to evaluate the error of a combination of different forecasts and understand the contribution of each depending on their magnitude, one can use Weighted MAPE (WMAPE) to complement the analysis.

$$WMAPE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{\sum_{i=1}^{N}\hat{y}_i} \tag{2.4}$$

Olson and Delen (2008) establishes as the main criticism to Holdout method the fact that it assumes that the data behaves in such a way that, regardless of the random partition on the data, both sets will behave similarly. This may not be true and, therefore, stratified sampling methods are advised.

The first of these stratified sampling methods, also explained by Olson and Delen (2008) is the ***k*-fold Cross-Validation**. This method, similar to Holdout, consists of the splitting of data in a train and a test set. However, instead of just doing this process once, the complete data set is split into k mutually exclusive sets of approximately the same size and each of them, in its turn, used as the test set while the other $(k-1)$ parts are used for training the model. This process is graphically represented in Figure 2.6.

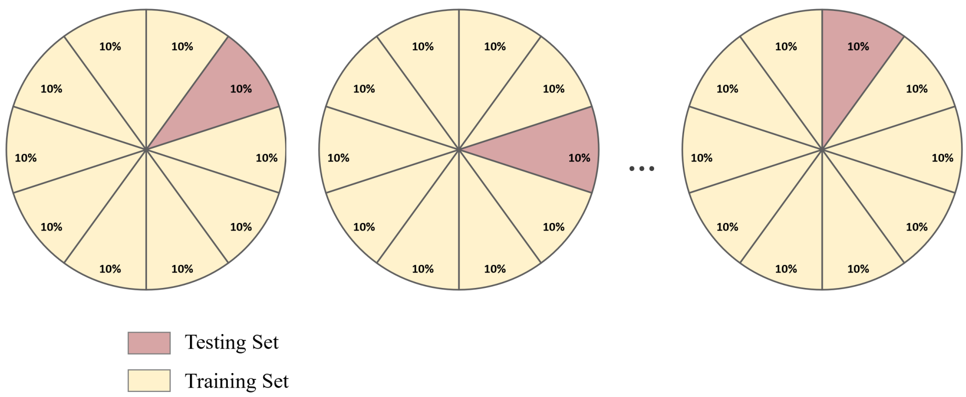

Figure 2.6: k-fold Cross-Validation for $k = 10$. Adapted from: Olson and Delen (2008)

Cross-Validation Accuracy (CVA) is, then, obtained by the average result of the individual measures of the accuracy metric chosen (A_i) obtained in each fold (Olson and Delen 2008). This accuracy metric can be one of the aforementioned (MAE, RMSE or MAPE) or any other metric coherent with the problem concerned.

$$CVA = \frac{1}{k}\sum_{i=1}^{k}A_i \tag{2.5}$$

Lastly, when speaking of models that result from time series forecasting, it may be relevant to evaluate the inherent correlation between observations that are near in time. For this reason, it may be relevant to test a **Time-Series Cross-Validation**. This sampling method, although derived from k-fold Cross-Validation, differs in how the data set is divided (Vairagade et al. 2019). The data, is then, divided, as represented in Figure 2.7, in such a way that the testing set is given by the observations chronologically immediately after the training set. In the next fold, the training set comprises both sets from the previous one, and so forth. Similar to k-fold Cross-Validation, Equation (2.5) also serves to combine the final accuracy metrics obtained.

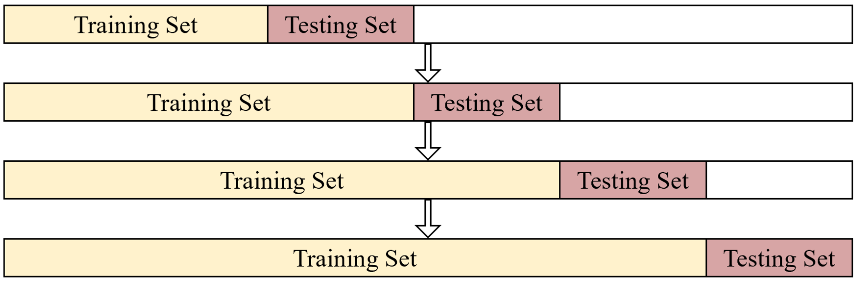

Figure 2.7: Time-Series Cross Validation. Adapted from: Vairagade et al. (2019)

2.4 Analytics in Telco Applications

Ashraf and Khan (2015) states that most of the existing literature dedicated to the analysis of potential usages of Telco data is precisely directed at data mining, and, particularly, at data mining used for making predictions.

Data used and its applications can, however, change a lot from one work to the next, according to Bughin (2016), due to the major diversity in data collection sources. Indeed, as stated in this work, companies that started collecting call and, later, text message data, now also have the benefits of being internet providers, and, through this service, being able to capture data from a variety of internet broadband applications.

Applications of this data can range from more classical ones such as the time series forecasting of outgoing calls, presented by Mastorocostas and Hilas (2012), to the approach of big business issues such as customer churn prediction, as is the case with Huang et al. (2015) who empirically demonstrated how big data makes this calculation considerably easier through an analysis focused on the three V's perspective ("Volume, Variety and Velocity"). Ashraf and Khan (2015), too, refers that Marketing and Sales teams are usually some of the biggest consumers of BI&A applications within Telco.

Bughin (2016) found, however, that, although this is a great area of investment for companies, the profit obtained from it was yet to be proportional, which leads to the conclusion that, even if this was one of the first industries to focus and largely invest on data analytics, its potential benefits can still be much greater than what has been achieved on these early stages of implementation.

Another issue worth mentioning associated with Telco data has to do with the increasingly prominent topic that is privacy and the (legally enforced) respect for personal information. One of the big topics related to this area is the concept of Differential Privacy. Jain et al. (2018) explains that it consists of the existence of mechanisms that add "noise" between the Database and the analyst, so that no personal information is provided about single individuals, but refers that the avoidance of data leakages can come at the expense of some accuracy for the models developed.

2.5 Final Remarks

The literature found within the topic of field workforce management in Telco showed that there is a generalized need for more in-depth research on data-based approaches to tactical resource planning. Indeed, when compared to operational management within the same applications, the latter has much more vast content to be examined. Therefore, this project aims at filling this gap, by studying how Business Intelligence and Analytics can contribute to support such planning decisions.

The, previously cited, three dimensions of an effective BI system (historical, current and predictive), established by Han et al. (2012), will serve as a basis for this study. The goal is the creation of a single concise deliverable, which can present a combination of meaningful information in the three fronts to decision-makers involved in tactical resource planning, and, thus, apply this data to future decisions regarding workforce dimensioning.

Because it is the one where deeper insights are being provided by academic research, the predictive component – and the contributions of Data Mining and Machine Learning models to this front – were also here presented in this Chapter. The project will, then, use these insights to explore the application of a Random Forest algorithm (Breiman 2001) in the Time Series forecast of intervention demand. Accuracy results measured through some of the aforementioned techniques will be provided, so that conclusions can be drawn regarding the fitness of the model to the Telco business and data context under study.

Chapter 3

Problem Scope and Description

The purpose of this chapter is to provide a clear understanding of the processes that have a direct impact on the problem tackled, establishing both the business and data context around the field operations that are the subject of this study.

The scope of the problem is, also, here established, as, due to the high complexity of the operations analyzed, the implementation of an initial solution required a narrower focus.

3.1 Business Context

As briefly explained in Chapter 1, the project here described, focuses on the I&M Operations Department that is responsible for managing all field operations where technicians are sent to specific client locations to perform specific jobs. The majority of these operations consist of Residential ones, for the B2C market segment, however, the same workforce also works in B2B solutions, for the Mass Business solutions, as these are technologically very similar to Residential installments.

The focus of this section is to explain the current processes associated with strategic and tactical management of this workforce.

3.1.1 Field Workforce Management

The technicians involved in these field operations are part of a subcontracted workforce, meaning that they are provided to NOS through strategic outsourcing partners, the SPs. These SPs are responsible for the recruitment and training of the technicians, and all fixed labor costs are ascribed to them, while NOS functions on a activity based approach, where its costs are directly related to the amount of interventions performed.

This generates an issue of conflicting business goals to be managed between the SP and NOS when tactical decisions, regarding the workforce capacity in each region, are made. Indeed, the partnership works in such a way that the SPs are responsible for all the fixed labor costs associated with the workforce and, therefore, is interested in minimizing them. On the contrary, NOS is directly impacted by the responsiveness to intervention requests, as delays can impact Customer

Satisfaction and even, in the case of Installations, cause sales to fall through. This results in a complicated balance between maintaining a high utilization rate for all resources but still being able to allocate all intervention requests within a short time span.

It is worth mentioning that NOS does have its internal teams of technicians, known as the SP NOS and the Urgent Squad. These consist of smaller groups of technicians with a higher amount of training that are responsible for interventions deemed of high importance. The creation of these interventions can be, for example, triggered by situations where other technicians have gone to the residence of a certain client but were not able to solve the problem reported. They are also responsible for the Corporate market segment, with more technologically complex installments.

After interviews with the managers of both teams to understand further how they operate, the planning of their capacity is considered out of the scope of this problem and, therefore, their operations will be isolated within the analysis. This was motivated by the fact that they function in a rather customized approach and so, not only do they constitute a residual portion of the total number of interventions but also one where patterns would be difficult to identify. The technicians in these teams tend to cover a large geographical area (such as the entire North of Portugal, for example) and, thus, the capacity provided by them is not directly assignable to the much narrower areas here under study. The planning process below explained, also, does not apply to them, as they are under different accountability within the I&M Department.

3.1.2 Internal Capacity Planning Process

In order to further understand how capacity related business decisions are made, interviews were held with the different persons, within the I&M Department, responsible for direct contact with the SPs. Table 3.1 summarizes who these are and their role within the process.

Table 3.1: Summary of the Technician Management Accountability in NOS

Time Period	**Performance Monitoring**	**Direct Negotiation with SPs**
Long-term	Partners Manager	Key Account Managers
Short-term	Interventions Manager	Capacity and Agenda Manager

Key Account Managers (KAMs) are directly responsible for negotiating long-term (that here is considered as a time period spanning from a full year to three months of antecedence) capacity decisions with SPs. They are organized in a team of four people, covering different geographical regions and with different SPs under the personal accountability of each. The team is directly supervised by the Partners Manager who is also the point of contact between the I&M Department and the Commercial Department.

As for the short-term, the Capacity and Agenda Manager is the person responsible for inserting the technician data – such as their schedules and days off – provided by the partners into *Click* and can, therefore, compare directly the capacity provided and the intervention requests placed in the software. To allow for higher adaptability, there is, thus, direct contact between this person and the SPs so that short-term changes (with one month to one week of antecedence) can be arranged, which may help improve performance in areas that are raising warnings due to low responsiveness. The Interventions Management team, on the other hand, is responsible for the quality assurance in present-day operations by handling immediate problems – such as absent clients or technician delays – which, essentially, means that any errors incurred during the workforce planning will be directly felt and subsequently reported through their operations.

Although all these stakeholders were crucial for the understanding of the problem, the key scope of it lays within the responsibility of KAMs – tactical planning.

As for the information this planning is based on, the process was found to be mostly reactive. As of the moment this project was initiated, predictions were only occurring for the demand of one specific type of intervention: the Installations. The discrepancy is justified by the fact that this data comes from a different department within the company. The Commercial Department, being responsible for direct client sales, are doing their own, monthly, forecast of new clients acquired, which directly reflects on how many Installations will have to occur. The communication is, then, established in such a way that these forecasts are delivered to the Partners Manager who oversees that the KAMs are made aware of them and, together with the SPs, adjust their plans accordingly.

A conversation was, then, held with the Analytics responsible for the Commercial Department to understand how their forecasting solution is implemented. Findings included the fact that the technology being sold is considered relevant as different patterns occur when selling cable (HFC) or fiber (FTTH) solutions. For the latter, there is also a relevant distinction between sales occurring in preexisting areas or sales occurring due to new openings: as the FTTH infrastructural network is still under expansion in Portugal, there are, on a monthly basis, new areas gaining access to this service. While some of the infrastructures are built and owned by NOS, in many cases, they derive from a joint venture between more Telcos or are owned by other companies whose core business is only focused on building infrastructure. In both these cases, not only is the opening date not fully under the control of NOS, and, thus, delays may occur which will affect planning, but a fast response is particularly crucial, as competitors are generally expanding to the area simultaneously.

Furthermore, it was found that installations forecast requires that an activation rate is taken into consideration, in order to estimate how a sale translates into an intervention. This is not necessarily a 1:1 relationship as some clients may have initially accepted the offer but end up canceling the service before it is installed and activated, an effect that is increased when responsiveness lowers.

Lastly, it was revealed that the forecasting process done by the Commercial Department involves several time horizons, so that accuracy can increase as antecedence lowers. Although results are calculated computationally through time series and market analysis, all obtained numbers are subject to human validation from those responsible for direct sales, so that non-quantifiable factors are not dismissed.

3.2 Operations Overview

As briefly explained before, since early 2020, NOS uses a solution known as *Click* to solve its daily TRSP. In general terms, this system is responsible for optimizing the number of tasks performed by each technician and their routes, while ensuring that all scheduled tasks are fulfilled. The software also consists of an integrated solution, meaning that in addition to the back-office optimization, it incorporates a direct communication with the platform where tasks are scheduled and handles the connection with technicians through an interface where they are able to see and report on their assigned interventions.

This process as well as the key variables used in it will be detailed in this section. To the end, the metrics used to evaluate its performance are also presented.

3.2.1 Intervention Scheduling and Assignment Process

The optimization and assignment process for field interventions includes three key moments.

First, there is the moment of creation of the task in the system. When a client requests an intervention, typically, a four hours time slot is agreed with them, corresponding to a certain morning, afternoon or evening, within which they should expect a technician in their home. As this happens, *Click* automatically attributes the intervention to a technician, so that, if this attribution fails, a warning is raised, showing there is no available capacity to meet the request.

The other two moments happen for a shorter time window, when interventions are processed in batch, in order to reassign them in an optimized manner. The first of the two is the Background Optimization (BGO), which consists of a TRSP algorithm that runs during the night and optimizes routes for the interventions scheduled for the next day.

This is, then, complemented by an in-day optimization that allows for readjustments to the calculated assignments. It runs recurrently throughout the day and is necessary to ensure that the system is recalculating the optimal solution in case of unforeseen events, such as delays in a certain intervention, belated cancellations or the absence of either a technician or a client.

To maximize the adaptability of this in-day optimization, most technicians are only notified ahead of the next two interventions they are expected to perform, rather than receive their full schedule at the beginning of the working day.

3.2.2 Intervention Characterization and Assignment Criteria

When an intervention is scheduled, a Work Order (WO) is immediately created in *Click*. In order to understand when and by whom it should be performed, it is, then, necessary that, as this scheduling happens, some parameters are filled that will characterize the WO. Table 3.2 shows a general overview of the most relevant ones.

As far as the Geographical Positioning is concerned, it is, also relevant to mention that this is established by associating the intervention with the Housing Unit (HU) where it will occur. NOS has a record of HUs that includes all of the locations its customers have services installed in as well as other registered residencies. When a WO is created, it has to be associated with one of these HUs, which provides all the geographical information necessary for the assignment process.

In regard to Type of Intervention and Technology, both these characteristics are particularly relevant as the categories a WO fits into will have a direct impact on its assignment and execution

Table 3.2: Characteristics of a WO

Group	Criteria	Description
General Categories	Type of Intervention	General descriptor of what is to be done in that WO.
	Technology	What kind of broadband solution is implemented in the residence where the WO will occur.
Geographical Positioning	Coordinates	Latitude and longitude that of the residence of the client that requested the intervention.
	Cell	Small sections in which the country is divided due to field technological infrastructures.
	Zone	Broader areas in which the country is divided, depending on the strategy of each SP.
Time Related Characteristics	Planned Duration	How much time the WO is expected to take to be fulfilled. Automatically filled as a function of the type of intervention and technology.
	Scheduled Slot	Establishes the time slot agreed upon with the client when the intervention should occur. Usually one of three available four hours slots.

process, as well as on the business targets it will impact. Therefore, a more thorough description of the possible categories will be bellow presented.

As for the types of interventions, a WO can be categorized as belonging to one of the following groups:

1. **Installations and Service Re-establishments.** These take place whenever a new client is acquired. The difference between the two lays simply on the fact that a Re-establishment happens when the client has had the service before and, at some point, it had been shut down. Nonetheless, the technological requirements are fairly similar and, therefore, the two are usually grouped in reporting applications.

2. **Service Alterations.** They occur when there is a change in the equipment a client has installed, the most common case being due to negotiations when a contract is renewed.

3. **Maintenance Operations.** These result from a technical failure in the service provided.

4. **Shutdowns and Scheduled Shutdowns.** When a client that had installed a cable solution is lost, the equipment has to be physically disconnected. The two types differ as regular Shutdowns can be done from the exterior while Scheduled Shutdowns demand that some equipment is recovered and, therefore, a time needs to be agreed with the customer.

5. **Audits.** These consist of quality checks performed during or after other interventions take place. Their occurrence can be either with legal motivations or simply due to an internal interest in improving the services provided.

6. **Budgeting.** These consist of less frequent activities, typically only occurring for certain Corporate Clients that are looking to purchase more complex solutions.

From all of these operations, Shutdowns, Audits and Budgeting interventions are to be considered out of the scope of the problem. This is due to the fact that they are, mostly, performed by technicians specialized in these particular jobs, so they do not directly impact the capacity available for others activities. Furthermore, with the exception of Budgeting, they do not directly serve the end-customer rather than internal needs of the company, and, consequently, the service level is not a high-priority factor when making decisions regarding these operations so assumptions used for this problem do not directly apply to them.

Besides the type of intervention, a certain field operation is, also, characterized by the type of technology it is to be performed in. NOS currently has three different solutions in use:

1. **Direct-to-Home (DTH)**. This is the oldest solution offered and still in use in some particular areas. With this alternative, services are provided via Satellite communication with a DTH antenna.

2. **Hybrid Fiber-Coaxial (HFC)**. These solutions are, also, often referred to, for simplification purposes, as just Cable. As the name suggests, it consists of a hybrid network with both coaxial cable and optical fiber being used to carry the broadband content.

3. **Fiber-to-the-Home (FTTH)**. Consists of a purely optical fiber solution. NOS, similarly to its competitors, is currently still expanding its FTTH network. For this reason, many of the Installations occur when this solution starts being available in new geographical regions, which often occurs for more than one Telco simultaneously. Therefore, providing a good service level in installations for these new FTTH openings consists of a key factor for competitive advantage.

Within the FTTH category, a further differentiation is necessary depending on how the infrastructure is built. Three types can be asserted: NOS Fiber, when the infrastructure is fully owned by the company), SWAP Fiber, when it results from a joint venture between NOS and another Telco and DST Fiber, bought from a company with the same name that is specialized in this service, and that typically sells to NOS and its competitors simultaneously.

Lastly, besides the characteristics defined at the moment of its creation, an important factor to be considered in any analysis of the operations is the WO status, which continually changes from the moment a WO is created, as displayed in the flowchart represented in Figure 3.1.

When an intervention is scheduled with the client, the WO is automatically created as Assigned or Not Assigned depending if it is allocated to a technician or not. Before its dispatching, a WO can change multiple times between these two statuses due to the continuous optimization processes happening in *Click*.

During the day a WO is scheduled for, it is initially dispatched and received by the technician it is, in the end, assigned to. Then, the technician marks the order as In Transit, meaning that they

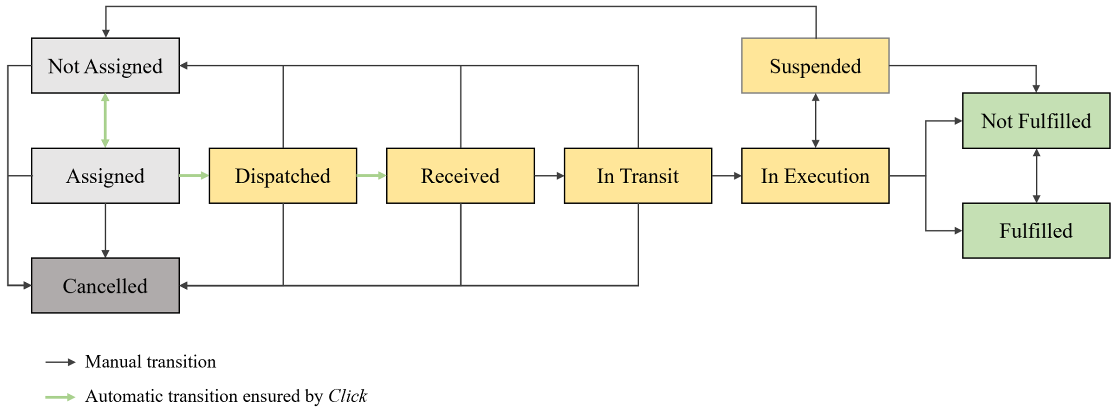

Figure 3.1: Flow of the status of a WO as implemented in *Click*

are currently in dislocation for the residence where the WO is to be executed. At any point up to this stage, the WO may be Cancelled.

During its Execution, a WO may be Suspended, which occurs when the technician has to stop the work, usually to communicate with their supervisors and quickly solve a particular issue related to the intervention. As the execution finishes, the WO is marked as either Fulfilled or Not Fulfilled depending on whether the objective of the intervention was achieved or not.

For the purpose of this analysis, both Fulfilled and Not Fulfilled WOs are considered, as they both occupy capacity. WOs that have reached either of these two statuses will also be referred to as Closed WOs through the course of this analysis. Contrarily to Closed ones, Cancelled WOs generally assume that status before a technician has to spend any of their time on them and are, therefore, not to be considered when studying occupation and capacity.

In order to understand how capacity can be measured for the operations outlined above, it is necessary to study the criteria that a technician is required to meet to qualify to be assigned a specific WO. These criteria can be categorized into three groups: general attributes of the technician, geographical positioning and work schedule. Based on this assortment, Table 3.3 shows the different sets of properties that characterize a technician and how they reflect the criteria that *Click* has to ensure are met when an intervention is assigned.

One of the main takeaways from meeting with the KAMs – responsible for the long term capacity negotiations with SPs – was establishing that this negotiation is done through the analysis of each municipality individually. This was, therefore, defined as the geographical unit used for all calculations involved in this project.

To further illustrate how these municipalities and their capacity correlate with the assignment criteria presented, Figure 3.2 provides a visual example of the geographical rules applied when checking if a technician is able to be assigned a WO. It does not represent a real, configured Zone or a real technician. Nonetheless, supposing a technician is configured with the represented departure point, operating radius and associated to the Zone represented, they would only be able to receive WOs registered within the areas covered by both conditions. For example, they cannot work in Vila Nova de Gaia, as it is outside their radius, nor in São João da Madeira, as it is in

Table 3.3: Assignment criteria that qualify a technician to perform a certain WO

Group	Criteria	Description
General Attributes	Active	Flag that essentially indicates if the technician is presently working for NOS or not.
	Skills	Set of expertise a technician has. These should match the required skills for each WO. Among other more specific ones, a technician must a skill indicating that they have received training in the technology used for the WO.
	Efficiency	Ratio that indicates if the technician is expected to be faster (Efficiency > 1) or slower (Efficiency < 1) than the average expected time for a certain WO. While this criterion does not disqualify a technician from performing a task it will weight in on the optimization so, more efficient technicians are expected to be prioritized.
Geographical Positioning	Departure Point Coordinates	Latitude and longitude that define the point where the technician is based, typically their residence.
	Operating Radius	Maximum distance the technician can travel, from their departure coordinates, to perform a WO.
	Zone	Broader areas in which the country is divided, depending on the strategy of each SP. For a technician to qualify for a WO, it has to be scheduled within their zone.
	Temporary Areas	Departure coordinates and operating radius may vary based on the day of the week. If this occurs, the technician has assigned a different set of these values for a specific weekday. For days and technicians where this does not happen the default coordinates and radius apply.
Work Schedule	Schedule	Start and finish time of the workday of a technician, plus their lunch break. One can, then, only be assigned WOs that fit within this period.
	Optional Work Schedule	Essentially consists of overtime work. During this period – which starts at the end of the regular schedule and has a fixed duration – *Click* allows tasks already in Execution to be finished, but no new ones can be initiated.
	Unavailability	Configuration that sets time periods in which a technician is not available in exception to their regular schedule. Depending on the reason, these can last from a few hours to a few weeks. Technicians cannot be assigned a WO if it overlaps with an unavailability.

a different Zone. In the example presented, their work hours should, therefore, count only as capacity available in Santa Maria da Feira, Espinho and Ovar.

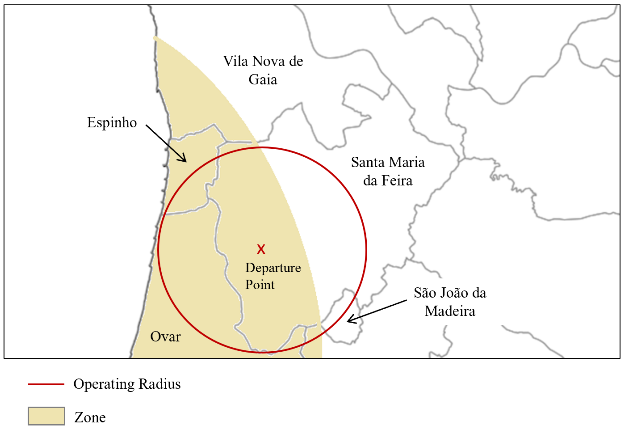

Figure 3.2: Schematic representation of geographical positioning rules. Background map obtained from: DGT (2020)

Furthermore, it was found that from the existing unavailability information, the motive could impact how they should be interpreted for this project. Some are scheduled with motive set as 'Reserve' and that essentially means they are manually placed in *Click* by the Capacity and Agenda Manager, to prevent the BGO from assigning ahead WOs to some technicians so that they are free to receive any last minute reassignments that may occur during the in-day optimization. These are normally, then, planned to be removed either after the BGO runs or throughout the day. For this reason, they will be excluded from this analysis as they do not constitute a capacity restriction.

Regarding the schedules of technicians, it is relevant to add that while this project was taking place, a standardization of these schedules was also occurring. While up until this point each technician could have a different one, the goal is that each SP should only have a limited amount of schedule options available for their technicians. It is estimated that this will decrease the number of different schedule definitions from about nine hundred to just forty different ones, once an agreement has been reached with all partners.

3.2.3 Current Performance Metrics

As mentioned before, the main concern for NOS as far as its field operations and the scope of this project are concerned, is to have a good service level, represented by high responsiveness. This is measured and reported through three main Key Performance Indicators (KPI):

1. **Average Intervention Time (AIT).** It represents the average amount of time taken between scheduling a WO and it being Fulfilled.

2. **Percentage of WOs out of Service Level Agreement (SLA).** Established a certain target AIT, here referred to as the SLA, this metric defines what percentage of WOs are not fulfilling this goal.

3. **Average Intervention Duration (AID)** It consists of a measure of technician productivity and indicates the average amount of time a WO spends in Execution. Although the two parameters are manually configured, ideally, Intervention Duration should be equal to the result of multiplying the Planned Duration of a WO by the Efficiency of a Technician, hence the importance of tracking this parameter, as technicians may, as a result of the wrong configuration, be under or over assigned.

These three KPIs are currently already measured and reported by Municipality, Technology, SP and Intervention Type which allows for different targets to be established on each of these levels. Division by intervention type is particularly relevant, as different targets must necessarily be established for each, depending on their urgency. For example, when it comes to Maintenance operations the SLA should be very narrow – normally around two days – as the amount of time a customer has their service down is something to be reduced as much as possible.

Additionally, a fourth KPI that will be relevant for this project corresponds to **Responsiveness** itself. As a metric, it defines the number of days it would take to give flow to all pending WOs at a certain point in time, assuming no others would arrive in the interim. It can be used as a compliment of AIT, as the first one may be affected by other variables that are not necessarily related to technician availability, such as problems in the technician allocation performed by *Click* or simply the availability of the clients to schedule interventions immediately. The comparison of the two can, therefore, provide important operational insights.

3.3 Data Characterization

This last section, then, aims at providing an overview of the current data available, its organization and characterization. Lastly, the creation of a structure required to obtain complementary data is, also, explained.

3.3.1 Data Structures

The majority of the data used in this study consists of information that is diffused directly from *Click* to Data Warehouse (DW). This data is then accessible through a MySQL RDBMS, where the model associated with this project was constructed. Figure 3.3 shows, in a simplified manner, how this data diffusion takes place and its periodicity. From here, it is important to note the fact that, while some data is diffused in real time and therefore available with twenty four hours or less of delay, others, in particular the WO status history, will take two days to be available.

There is, additionally a second database that includes all information stored as it was reported previously to the implementation of *Click*. The two are separated because information was managed very differently and most of the previously presented rules for assignment of interventions were not similar in the previous software, and, therefore, this one is not very often used for current reporting. Nonetheless, it is still also being updated with data diffusions and it does include further historical tracking of clients which will be required for this analysis.

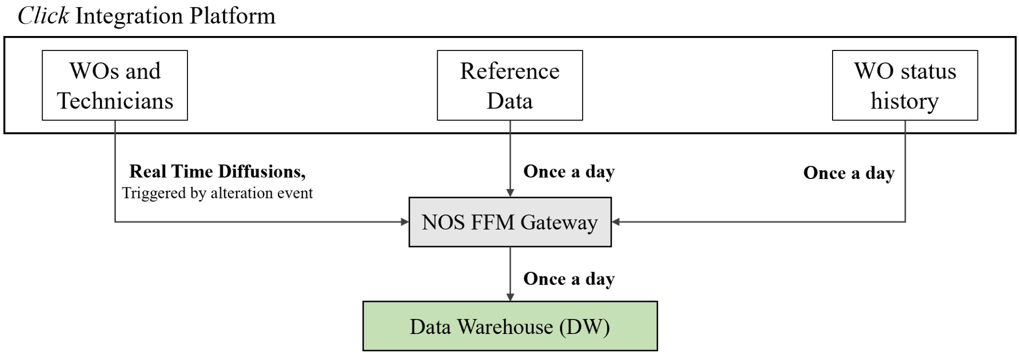

Figure 3.3: Simplified schematic representation of diffusions between *Click* and DW

Lastly, it was identified that some crucial information was missing from the extractions above explained, as the data, although processed by *Click* was not being diffused to the DW. This triggered the beginning of a parallel project to engineer this diffusion, however, in order to continue with the development of this study an alternative, semi-manual solution was built, which will be further explained towards the end of this section.

3.3.2 Data Description

The information used for this project, as qualitatively described above, relates to two key dimensions: the historical WO records, with the characteristics presented in Table 3.2, and the technician characteristics that will impact their availability to perform a certain intervention, enumerated in Table 3.3.

Concerning how the information is accessible, Table 3.4 presents a summary of the initially available tables in the RDBMS of NOS that will be relevant for this analysis.

Table 3.4: Summary of the Tables in the I&M database relevant to the model

Category	Table	Description
WOs History	ACTIVITIES	Summary of every WO that is created in *Click* and its progress. Every WO represents a single record.
	ADDITIONAL_AUDITS_INFO	Table where the record of In Transit status is compiled for every WO.
	ALTERATIONS_HISTORY	Where it is stored all records of alterations applied to a WO such as technician attributed or status. Every alteration represents a record.
Technician Information	TECHNICIANS	Historic data of the configurations of technicians.
	TECHNICIAN_SKILLS	Relation between each technician and their skills.
	UNAVAILABILITY	Historic data of all unavailable schedules registered for every technician
Others	HUs	Record of all HUs belonging to clients, as well as their geographic information.

From these, the historic WO data registered in the ACTIVITIES table will be used not only to calculate past occupation rates but, also, for the forecasting of future WO demand. For this reason, and to provide some more visibility on the existing data, an exploratory analysis of its records was performed. Tables 3.5 and 3.6 show a general statistical analysis of the weekly number of WOs by municipality, for both the types of interventions within the scope of this study and the existing Technologies in the municipalities where they are implemented.

Table 3.5: Exploratory analysis of the number of weekly WOs in each municipality by type of intervention

Type of Intervention	**Mean**	**Standard Deviation**	**Coefficient of Variation**	**Min**	**Median**	**Max**
Service Alteration	32.7	84.0	2.6	1.0	7.0	1355.0
Installation + Reestablishment	26.5	54.3	2.1	1.0	9.0	974.0
Maintenance	46.5	122.1	2.6	1.0	11.0	1862.0

Table 3.6: Exploratory analysis of the number of weekly WOs in each municipality by Technology

Type of Intervention	**Mean**	**Standard Deviation**	**Coefficient of Variation**	**Min**	**Median**	**Max**
DTH	10.3	10.9	1.1	1.0	7.0	119.0
NOS Fiber	36.3	37.3	1.0	1.0	24.0	308.0
DST Fiber	10.0	10.9	1.1	1.0	7.0	134.0
SWAP Fiber	9.4	10.9	1.2	1.0	6.0	112.0
HFC	184.8	400.4	2.2	1.0	42.0	4714.0

As can be observed, the majority of the WOs is made up of Maintenance Operations, while the predominant Technology tends to be HFC. This second one is due to the fact that FTTH infrastructure is still being constructed and not yet implemented in some of the biggest urban areas in the country.

High standard deviations can be explained by the difference in population density in the penetration rate of NOS in different municipalities. This further justifies the need to weight as a predictor this additional geographical factor.

3.3.3 Data Extractions

As before explained, when reviewing the data that would be necessary for this analysis, it was identified that some information, although existent and used by *Click*, was not integrated in the diffused data arriving at DW. This data, specifically, corresponded to:

1. **Technician Efficiency.** Although there was a TECHNICIANS table being updated regularly, this attribute was not included in it.

2. **Temporary Areas.** In this case, all information related to this topic, explained in Table 3.3, was missing.

3. **Active Unavailability.** The problem here was that a single diffusion of each unavailability was occurring at the moment of its creation. If an unavailability was altered or deleted, this information would not be transmitted.

4. **Schedules.** The name of the schedule a technician has associated is diffused as an attribute of TECHNICIANS, however, a complementary table is required to interpret them. There should, then, be an established association between each schedule name and its corresponding start, finish and optional work times.

Because this data was crucial for an accurate analysis of the workforce capacity, an initial step of this process, before modeling, had to be ensuring this data was available in the Database. The solution found was a semi-manual process described in Figure 3.4.

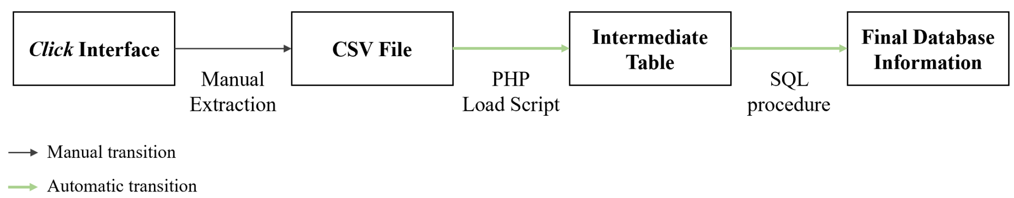

Figure 3.4: Solution found for the information extraction process.

This process requires that, with a certain periodicity, a user manually logs into the *Click* back-end management interface and in the respective visualizations of the aforementioned data manually extracts it as a CSV file.

The file should be saved in a specific shared folder which a PHP script is programmed to scan periodically (with a periodicity here set for every ten minutes). Given the name of the file, the script is able to detect which of the four possibilities it corresponds to and, based on them, executes the respective "LOAD DATA" MySQL query. From there, the data is stored in an intermediate table that contains only the records that were just uploaded.

The same script, having read and loaded the file, will trigger a procedure in the Database that will transform the data from this intermediate table, inserting them into another one where some fields are calculated by the procedure.

From the occurring transformations in this procedure, it can be noted that for the case of Technician Efficiency and Temporary Areas the focus is on building historical records, and, therefore, having the process working for a prolonged period of time, different technician configurations will be visible, with a validity time range associated with each. As for the Unavailability, the goal is to find out whether the diffused ones still occur and, thus, a copy of the UNAVAILABILITY table is made, where, for the period extracted, if the unavailability no longer exists in *Click* and, consequently, in the uploaded table, it is marked as inactive.

As for the Schedules, the data extracted corresponded to the start and finish time on a specific period of time. However, as mentioned before, it was known that there was a project in place to standardize technician schedules and, therefore, the procedure activated here focused on finding the patterns on the data that could be transposed for any future date. An example of one of these patterns could be: "every odd week, on Tuesdays, technicians with this schedule will start working at 8am, finish at 5pm and do one hour of optional work".

All of the first three extractions were functioning before project implementation began. As for the schedules, because their standardization had not been finished, this information was often replaced in the model by a default value.

This process, although allowing for information to be retrieved so the models can use it, has several limitations, from which the main ones worth noting are:

- The fact that there is no historical knowledge of past configurations means that the capabilities of the model to study past performance is limited.
- Data availability depends on a person triggering the process by extracting the files and placing them on the shared folder. This means that information gaps may occur if extraction frequency and accountability are not well defined.
- Every extraction corresponds to static information from the moment it occurred. If, in between extractions, a certain configuration is set for a small period of time, that information will never be received in the DW.

It is, therefore, expected, that a more robust alternative is engineered and, this project should be implemented having present that the models created should be adaptable to the data source of the here mentioned attributes having to, at some point in the future, be adjusted.

Chapter 4

Project Implementation

This chapter aims at describing the key steps taken in the execution of this project. The sections presented correspond each to one of the objectives established for the project.

The methodology and timeline of the implementation of each of these phases are described in Figure 4.1. As can be seen here, the three stages had to be comprised into a shorter time span, with a slight overlap, rather than a fully sequential process which would have occurred in an ideal scenario, as they were dependent on the data preparation, within which the data extractions process explained in Chapter 3 is included.

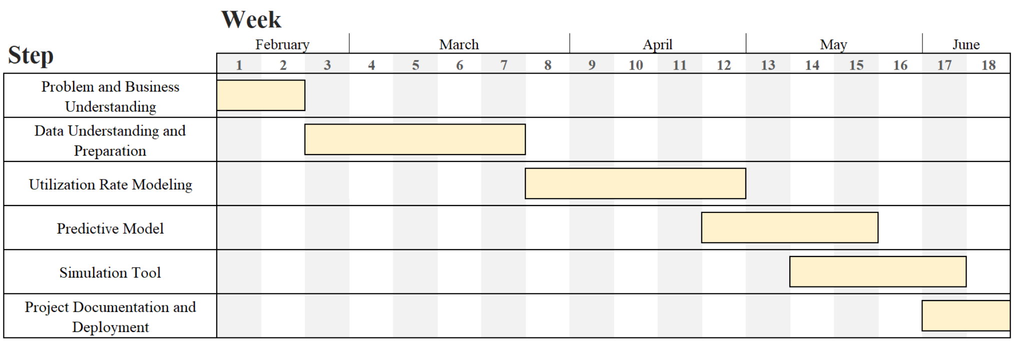

Figure 4.1: Gantt chart representing the project implementation timeline.

4.1 Utilization Rate Modeling

The first step of this project was established as being the construction of a data model that proposes to transform the existing database in a way that can calculate the geographical capacity and occupation rates. The aim is to both use this data in the subsequent steps of this project but also to create the data infrastructure surrounding this information that will be automatically updated with field information, so that future reports or other applications can be built based on it.

In order to do so, it was identified three key aspects needed to be made visible, and that will be, in this section, individually tackled. They can be summarized in the following questions:

1. Which municipalities should the work hours of each technician be ascribed to as capacity?
2. How many work hours did each technician spend working on a certain day?
3. How much time did the WOs performed on a certain day require?

All of the below mentioned associations and calculations were performed through the creation of a schema in the MySQL RDBMS, that integrated information resulting from the *Click* diffusions as well as other established configurations and the manual extractions explained in Chapter 3.

4.1.1 Geographical Distribution

As previously explained, in order for a technician to qualify to perform a certain WO, it is necessary that it both takes place within the operating radius of the technician and within their zone.

The distance calculation was here performed using the Equidistant cylindrical projection (Snyder 1987), as the simplification can be applicable to relatively small distances within the globe. It essentially consists of a conversion of cylindrical coordinates to a plane and from there, a Euclidean distance formula is applied.

$$distance = \sqrt{(Lat_i - Lat_t)^2 + (Long_i - Long_t)^2 * \cos(Lat_t)} \quad , \tag{4.1}$$

with:
i as the WO
t as the Technician
Lat as the Latitude, converted from degrees to radians
Long as the Longitude, converted from degrees to radians

Based on this, it can be established that a technician qualifies to perform a WO, when their $radius \geq distance$ to said WO.

This radius criteria, however, raises the question, of how to establish the equivalence between technician working hours and the capacity of a municipality, when it can occur that the full municipality is not included in the operating radius or zone of the technician, and, in the case of municipalities with a larger area, using its center coordinates may be a too rough approximation.

The solution found was to use the Cell as the smallest unit of area that can be associated with a single municipality and to perform calculations based on it. Figure 4.2, portrays a simplified study (more attributes and associations may exist in each class), using Universal Modeling Language (UML), of the relationships between the different geographical units as established in the database.

A table was then built establishing the connection between the cell and its corresponding municipality. Additionally, each cell received as attributes, both the number of HUs it included and a centroid point, which got its coordinates from calculating the average latitude and longitude of the HUs of the cell. This way, rather than simply establishing a geometrical middle point, the centroid will tend towards the most populated areas of the cell. The result consists of Table INFO_CELLS, from which a sample can be observed in Table A.1 from Appendix A.

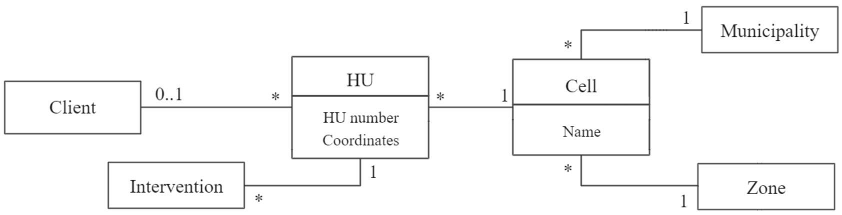

Figure 4.2: Simplified model of the relationships between geographic units.

From here, one can use Equation (4.1), replacing the WO coordinates with the centroid of the cell and, thus, establish, approximately, which cells each technician operates in. After filtering which cells and technicians have the same Zone attributed to them, it is found which municipalities the technician has under their reach, as the cells are directly connected to a single municipality.

Nonetheless, a new problem arises: if the geographical positioning of the technician allows them to cover cells within more than one municipality, how should their work hours be distributed as capacity between those various municipalities? The solution was to calculate the percentage of HUs a technician covers in each municipality ($\%HU_m$) and, from there set a threshold percentage, so that the technician would only be ascribed to municipalities where $\%HU_m$ was higher than said threshold. This way, the work hours of the technician are only to be considered capacity in municipalities where they cover a significant area. During this study, the value used was 15%, however, depending on the business and reporting objectives, this parameter may be adapted.

The calculation of $\%HU_m$ based itself on the previously calculated number of HUs per cell, and can be represented by, for each technician:

$$\%HU_m = \frac{\sum_{c=1}^{C} N_{c,m}}{\sum_{m=1}^{M}\sum_{c=1}^{C} N_{c,m}} \quad , \tag{4.2}$$

with:

$c \in [1, C]$ as the set of cells in a municipality
$m \in [1, M]$ as the set of municipalities covered by the technician
$N_{c,m}$ as the number of HUs in each cell of each municipality

4.1.2 Work Schedule

As shown previously in Table 3.3, the calculation of the number of daily work hours of a technician depends both on their schedule and on whether there is any planned unavailability for the day. In the schedule of a technician, there is also the distinction between Regular Work Hours and Optional Work Hours, however, the two ended up being joined together (by considering the finish time of the technician as the finish time of their optional work schedule), as, although Optional Work does have some scheduling restrictions, it can still be considered as existent capacity.

Having as a calculation basis a specific day, the logic behind the establishment of the number of hours worked by a technician can be explained sequentially. First, the technician schedule is joined with the calendar in such a way that for each day and each technician it is displayed the start time (t_{start}), the finish time (t_{finish}) and the total work hours (Δt) of the technician, given by:

$$\Delta t = \begin{cases} t_{finish} - t_{start}, & \text{if } t_{finish} - t_{start} \leq 5 \\ t_{finish} - t_{start} - 1, & \text{if otherwise} \end{cases} \tag{4.3}$$

There is no diffusion of the configurations regarding lunch breaks, hence using five hours as a threshold to define that, if the workday of a technician lasts longer than that, they are required to take a one hour break.

From there, these three values are adapted, in case there is an unavailability set for the technician in question and overlapping with the day in question, following the logic displayed in Table 4.1, where s represents the schedule and u the unavailability it is being compared against.

Table 4.1: Different unavailability scenarios and how they are treated by the algorithm.

Case	**Changes Made**	**Example of Situation**
$t_{start,u} \leq t_{start,s} \wedge t_{finish,u} < t_{finish,s}$	$t_{start,s} := t_{finish,u}$	Technician takes the morning off.
$t_{start,u} > t_{start,s} \wedge t_{finish,u} \geq t_{finish,s}$	$t_{finish,s} := t_{start,u}$	Technician has to leave work earlier.
$t_{start,u} \leq t_{start,s} \wedge t_{finish,u} \geq t_{finish,s}$	$\Delta t_s := 0$	Technician takes the full day off or is on vacation.
$t_{start,u} > t_{start,s} \wedge t_{finish,u} < t_{finish,s}$	$\Delta t_s := \Delta t_s - \Delta t_u$	Technician leaves for a few hours during the day.

With these alterations and, in the first two cases, recalculating the total work hours using Equation (4.3) it is, afterward, fully known how many hours each technician worked every day.

Furthermore, it may be relevant to include in the model the fact that a technician is configured in *Click* with a parameter correspondent to their Efficiency (ε) that can be used to transform work hours capacity in such a way that:

$$Real\ Work\ Hours = \varepsilon \times \Delta t \tag{4.4}$$

This parameter was modeled in the calculations, however, during this study, total work hours (Δt) will be the variable used to calculate total capacity. This is due to the fact that the Efficiency, although having an impact on the real capacity of the resource, is currently manually configured for each technician and each SP decides what value to attribute to each of their technicians, without the reasoning behind the values attributed being fully known or standardized among them. In order to include this parameter in the final capacity calculations, it would require that it starts being tracked and its accuracy fully studied in order to understand if this estimate is anywhere close to the real value – that would consist of the average ratio between the time the technician takes to perform each WO and the expected duration of the WO.

4.1.3 Occupied Capacity

The third and last necessary component of this study is to understand, from the capacity available, how much of it was used performing interventions.

Records of the occurring WOs by day already exist and are registered in table ACTIVITIES, as previously shown in Table 3.4. The only fields necessary to calculate are, then, the WO duration and the geographical equivalence, attributing WOs to the municipalities where they occurred. This second point was rather direct since, as shown in Table 3.2, the WO already has a cell associated and can, thus, be associated with a municipality, using the, by now created, INFO_CELLS Table.

For the calculation of intervention duration, it is necessary to study the progress of its statuses, which were presented in Figure 3.1. Here, both the In Transit and In Execution status are considered, as both of them generally occupy technician capacity that will be here represented in the form of Dislocation ($\Delta t_{dislocation}$) and Execution ($\Delta t_{execution}$) time, respectively, defined as:

$$\Delta t_{dislocation} = t_{start,\ execution} - t_{start,\ dislocation} \tag{4.5}$$

and,

$$\Delta t_{execution} = t_{end,\ execution} - t_{start,\ execution} \tag{4.6}$$

There is the question of how to consider Suspensions as the amount of time they last can imply different reasons as to why they occurred and, subsequently, different effects on the routine of technicians. For example, when the Suspension only withstands for a few minutes, it was likely a necessary part of the intervention process, while, if it continues for longer, it may be that the technician had to leave the location where the WO was taking place and, in some cases, they may have even executed another WO while one was Suspended. Theoretically, in these second, longer cases – and depending also on the reason for suspension reported by the technician – suspension time should be subtracted from the total time occupied by the WO. However, while the ACTIVITIES and ADITTIONAL_AUDITS_INFO have a single record for each WO, Suspensions would have to be analyzed through table ALTERATIONS_HISTORY , which included multiple records for each WO, and, consequently, much larger volumes of data. Thus, for this analysis, suspension records will not be considered as the computational effort would presumably surpass its benefits.

Alternatively, for both dislocation and execution times, there was an interval established to avoid outliers that could result from the aforementioned suspension cases or from potential technological issues: for example, it was detected that some negative dislocation times occur because of internet connection failures in the devices of the technicians. These values and their alternatives are represented in Table 4.2.

For Dislocation time, the value is replaced with the average of the technician as the amount of time they spend in dislocation is highly dependent on factors such as the geographical area covered or the vehicle used. The number of days used in this average calculation was defined as a parameter that can be manually configured so that it only captures days representative of the current situation of the workforce. For Execution duration, on the other hand, the Planned Duration of the WO is

Table 4.2: Limits outside which a duration is considered an outlier and its respective replacement.

Time Parameter	**Normal Interval (hours)**	**Replacement Value**
Dislocation	[0, 2]	Average Dislocation time of the technician assigned to the WO.
Execution	[0, 4]	Configured Planned Duration for the Type of Intervention and Technology of the WO.

used since, not only is it expected that it constitutes a good estimate of how much time the WO should last, but it is also the time reference *Click* uses when scheduling and assigning WOs and, therefore, will have a direct impact on the amount of WOs the technician performs.

4.1.4 Integration and Utilization Rate Calculation

In order to maintain the schema up to date with the daily data diffusion, three MySQL procedures were created to insert and update info in the tables created from the above explained calculations and that will be further presented in Chapter 5.

Each of the procedures incorporated the logic of one of the above presented sections. All of them were, subsequently, linked to a Database event that would be triggered during the night, after the BGO had run in *Click* and its updates had been diffused to the Database, in order to incorporate any alterations. These procedures included default values that could manually be altered in case some of the model assumptions were to be changed, such as:

1. A schedule defined by a default start and a default finish time that should be used in case there is missing information regarding the schedule a technician has assigned to them or the correspondent scheduled has not been extracted so far.

2. A default efficiency if this field is missing for a specific technician. Both of these two initial points were defined as a fail-safe alternative in case issues occur in the data extractions.

3. The upper bounds of the intervals shown in Table 4.2 after which dislocation and execution times are considered outlier values.

4. For each of the procedures, the number of days in the past that were being recalculated. This way, any changes made to the model can be easily incorporated in its historical records.

With these three procedures functioning, then, the calculations executed can easily be used to compile the information and presenting, for each day, the available capacity and the used capacity. Combining the two, utilization rate is given by:

$$\%\ Utlization = \frac{Used\ Capacity\ (hours)}{Available\ Capacity\ (hours)} \times 100 \tag{4.7}$$

Because this calculation can be done using many possible groupings, depending on the aforementioned WO categories, the goal, with the integration decisions, was to provide a flexible system so that, in case any of the described assumptions change or any new reporting objectives arise, the model is adaptable to them. For this reason, the utilization rate results were chosen to be presented by storing Views, rather than Tables in the Database, so that any additions or alterations to the calculated fields could be easily implemented by simply changing the associated query. The current presentation of the results obtained will, nonetheless, be exhibited in Chapter 5.

4.2 Predictive Model

The next step of the project was to incorporate a predictive component that will allow to estimate the future capacity used. The goal is that, with the previously described model shedding light on the geographical distribution of the available workforce, by knowing the number of expected WOs in each municipality, it is possible to make workforce dimensioning decisions ahead so that both responsiveness and utilization KPIs remain under control.

The output of the model was initially debated between two possibilities: estimating the demand in number of WOs or directly the amount of time used performing them. The demand in **Number of WOs** ended up being the factor chosen, however, as relevant predictors for the duration of a WO such as technology related attributes of the WO were outside the scope of this study.

The model here described was implemented through Python programming, using the scikit-learn package.

4.2.1 Data Selection

Having analyzed the business and data context, the first step of this phase was to select what predictors would be used.

It is hypothesized that yearly seasonality will have a strong impact on the output, as the effect was described by the clients of the model: for example, highly touristic areas of the country tend to see an increase in WO requests during the summer, while other areas may have a particular increase in Maintenance requests during the winter, due to weather complications. Having this in mind, **Week of the Year** together with **Year** were the time variables chosen as predictors.

From there the other three variables chosen are categorical ones and constitute the general characterization of a WO:

1. **Type of Intervention** that is briefly characterized in Table 3.5, and, as shown there, will adopt the values of "Service Alteration", "Installation + Reestablishment" or "Maintenance".

2. **Technology**, also described in Table 3.6, and divided in the five categories there presented: "DTH", "NOS Fiber", "DST Fiber", "SWAP Fiber" and "HFC".

3. **Municipality** which has a much larger set of possible categories but, as before concluded, will be relevant has patterns vary geographically.

An advantage of the case study here used is that, unlike with most demand forecasting applications, here censored demand is not an issue, as both the date when the request for an intervention was placed and the date the intervention was executed are known. Since there may be a discrepancy between the two dates, due to capacity shortages, and the intention is to capture demand as it occurs, the WO creation date will be the one used here.

4.2.2 Data Preparation

Data collection from the DW to the model was performed using the *read_sql* method within the Python *pandas* package, saving the result as a Dataframe. This query performed the grouping of WO records by the aforementioned predictors (counting how many records occurred as the output variable). It is, then, necessary that the algorithm first adds all null records as they do have a meaning that should be interpreted by the model.

All date formats used in this project were handled in accordance with the ISO 8601 norm. In this case, it means that the conversion from date to Week of the Year was done in such a way that the week that is considered to be the first of the year is the one with four or more days of the week in the new year, or in other words, the week that includes the 4^{th} of January.

As for the other variables, Type of Intervention and Technology do not require any data cleaning, except for filtering out the Types of Intervention deemed out of scope. Although it was not a frequent occurrence, data points with no Municipality associated were removed from the data set. Records corresponding to Cancelled WOs were also discarded so that the model would not overestimate the future amount of relevant WOs, which could result in an oversized workforce.

Dummy encoding was performed for all three categorical variables as there is not an intention to assume correlations between categories. It is also known that the model chosen is able to support a large number of predictors while maintaining accuracy (Cutler et al. 2012).

4.2.3 Model Building

For the above cited business reasons, the model should take in five independent variables (X), two of which characterizing the time series, while the other three are categorical ones. The dependent variable (Y) of the model is, then, the number of WOs predicted for the combination of inputs. Figure 4.3 summarizes this variable distribution.

As for the evaluation of the model, it was done through an adjusted version of the Time Series Cross Validation presented in Chapter 2. This adaptation is portrayed in Figure 4.4 and the core difference between the two is simply the fact that, while in the one previously presented, from one sample to the next, the training set increases, here its time length remains constant. The goal of this adjustment was to avoid that the first months used to train the model create a bias, due to the fact that the transition between the previous software and *Click* was still ongoing and, therefore, some data may be missing or inconsistent.

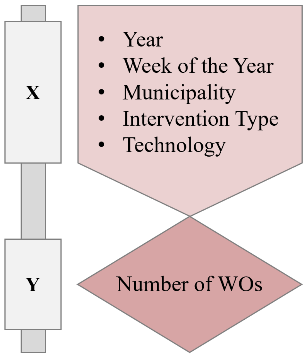

Figure 4.3: Distribution of the variables of the model.

The data set was then partitioned, following this rationale, in a training and testing set characterized by a constant time length, in this case of one year (fifty two weeks) for the training set and a month (four weeks) for the testing set. This way, it is expected that the model is able to capture any seasonal trends, while predicting the most relevant time horizon for planning decisions.

The Machine Learning algorithm chosen was the Random Forest, as it comprises some of the advantages deemed relevant for the application under study. These include, additionally to the previously mentioned high number of predictors supported, the fact that the algorithm is computationally efficient while still producing good quality results that can be easily interpreted (Lingjun et al. 2018). As explained in Chapter 2, Random Forest consists of an ensemble method with the potential to be used for either Classification or Regression – this second one being the focus in the case of this study – where each base algorithm is a decision tree build through a bootstrap sample of the original training set.

Accuracy results of this algorithm applied to the data set will be presented in Chapter 5.

Additionally, the table and database procedure were built so that the model can be in the future integrated within the DW and the below explained simulation tool, using this data to support capacity planning processes. It should also be expected that the results of this model, because it is purely based on historical results, do not, in the case of Installations and Reestablishments, overrule Commercial Department predictions, which include much more detailed business sales insights. Although it is not the scope of this project, it is advised that the two prediction sources are centralized in the DW, so both can be used to complement capacity planning decisions.

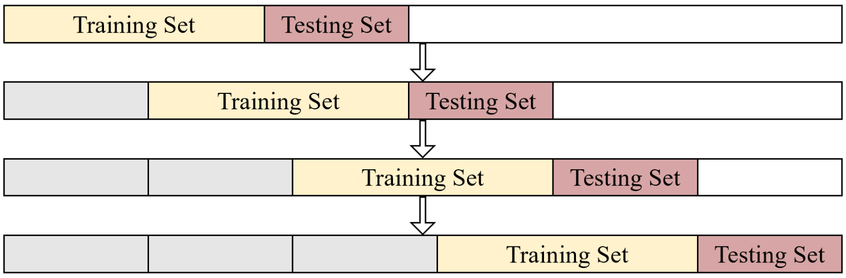

Figure 4.4: Adapted Time-Series Cross Validation.

4.3 Simulation Tool

The main objective of this final step is to have a way of communicating the current performance to the clients of the model, as well as, based on future predictions, allow them to simulate different scenarios and, thus, support them in workforce dimensioning decisions.

The main users of this interface will be the KAMs, as, from the stakeholders involved in this process, they were the ones more directly involved in the tactical planning of the workforce distribution. The tool was, then, developed bearing in mind their specific needs and iterated through the feedback received from them. Because of the amount of inputs the simulation needed to support in order to depict thoroughly the decisions it could be used to make, it was found that the platform that could fulfill those requirements in a comprehensible way for its users – as they were already familiar with it – would be a Microsoft Excel Workbook.

The simulation proposed allows for the creation of scenarios, in specific, user-defined, municipalities, regarding the following aspects:

1. **Variations in the number of expected WOs**, which will allow an evaluation of risk, defined as the impact caused by possible deviations from the values predicted, or the quick assessment of the impact of new predictions, for example, if new information regarding Installations arrives from the Commercial Department.

2. **Variations in the workforce**, allowing different solutions regarding resource planning and their impact on performance to be tested.

3. **The establishment of Responsiveness targets**, and in this case the simulation will do the reverse calculation and communicate the minimum number of technicians necessary for those targets to be achieved.

As will be mathematically shown, the last two of these simulation variables are mutually dependent. Thus, the scenarios created will either consist of changing just one variable or a pair of variables where one of the two is the number of expected WOs.

4.3.1 Assumptions

One of the most crucial operational objectives of this phase was that it should allow for simulations to be made in which the number of technicians available in a certain municipality could be altered, so that KAMs can easily test resource planning solutions and their impact on performance.

In order to do so in a comprehensible way, without forcing the users to actually point out which specific technician would be moved, as that is, normally, a decision of the SP and managing individually the technicians is outside the scope of accountability of KAMs, some assumptions had to be made:

- Variations in the workforce apply exclusively to the municipality they are input for, so, for example, if a technician is removed from a municipality it is assumed that they were only assigned to that one and no impact will be registered in neighbor municipalities.

- Technicians all work the same number of hours daily and no unavailability is applied to them. The number of hours is a parameter the user can define.
- Technicians that are placed in a certain municipality, all have the skills corresponding to the technologies in use in that municipality.

Furthermore, in order to calculate the amount of time necessary to perform all WOs predicted for a municipality, an estimator is necessary for the amount of time each individual WO will require. It is assumed that for each specific WO type and municipality, the historical AID is an accurate estimator of the duration of future WOs.

Because AIT, although the most common KPI for service level, cannot be perfectly predicted, as it depends on factors that are not mathematically computable, such as the WO types *Click* prioritizes when scheduling or client availability, it is replaced by Responsiveness in the simulation scenarios. It is assumed that, because both indicators represent the ability of the existing workforce to flow off pending WOs, improvements in the two metrics will be proportional.

4.3.2 Data Integration and Calculations

Two sets of data were retrieved from the database: one reporting on the current geographical positioning of the technicians and a second one reporting on the WO history. Ideally, a third one would have been used to integrate predictions, however, as the second and third stages of this process were executed simultaneously, and, due to time constraints and a necessary prioritization, deployment of the predictive model was not concluded – the predicted number of WOs here used will be replaced with the average number of WOs of each type for the past N weeks, where N is a parameter that can be manually set by the user.

The geographical positioning of the technicians is retrieved from a query associated with the ACTION_TECHNICIAN_MUNICIPALITIES table, shown in Table A.2 of Appendix A, and the used number of technicians corresponds to the daily average within the present week, as there may be temporary areas to be considered.

As for the WO history, it results from a query associated with the WOs table (Table A.4 of Appendix A), where these are aggregated by Week, Municipality, Type of Intervention and Technology and data is fetched for the past N weeks. For that aggregation level, it is also retrieved the number of WOs that occurred and the AID and AIT of those interventions.

Below, the formulas that correlate these variables with the required simulation outputs will be explained, using the following notation:

- $i \in [1, M]$ as the set of WO types, resulting from a combination of Type of Intervention and Technology;
- $n \in [1, N]$ as the set of weeks used in the calculation;
- $Q_{n,i}$ as the number of WOs of type i occurring in week n;
- $AID_{n,i}$ as the AID, expressed in hours, for i WO type in week n;

- T as the number of technicians available;
- h as the manually established number of daily work hours of every technician;
- U as the average utilization rate of the technicians available in the municipality.

All calculations will take place using a municipality aggregation and, thus, all mentioned variables, except for h, will assume different values for each municipality.

AID, as explained above, refers to the estimator for WO duration. This value should then be transformed from a week average into a total weighted average.

$$AID_i = \frac{\sum_{n=1}^{N}(AID_{n,i} \times Q_{n,i})}{\sum_{n=1}^{N} Q_{n,i}} \text{ (hours)} \tag{4.8}$$

From this point, having the information regarding how much time each individual WO takes, one can estimate, for a $n+1$ week, how much time the total amount of WOs occurring in each municipality will take, by doing:

$$Total\ Duration = \sum_{i=1}^{M}(AID_i \times Q_{n+1,i}) \text{ (hours)} \tag{4.9}$$

Also important to note that, as mentioned above, $Q_{n+1,i}$ is one of the parameters that should be subject of simulation. This means that a ΔQ_i may be defined by the user, and in that event, such parameter must be added to the average number of WOs, here used as the default number of predicted WOs. So, in Equation (4.9), $Q_{n+1,i}$ would be given by:

$$Q_{n+1,i} = \frac{1}{N}(\sum_{n=1}^{N} Q_{n,i}) + \Delta Q_i \tag{4.10}$$

From here, the relationship between Number of Technicians available and Responsiveness can be established. In the case of a simulation of workforce alterations, a parameter ΔT, representing this variation in number of technicians is set. The result is:

$$Responsiveness = \frac{Total\ Duration}{(T + \Delta T) \times h \times U} \text{ (days)} \tag{4.11}$$

Alternatively, if the goal is to establish a target Responsiveness – that, as mentioned before, can result from the establishment of a target improvement in AIT – one can transform Equation (4.11) and, from there, obtain the total necessary amount of technicians (T) for that Service Level.

$$T = \frac{Total\ Duration}{Responsiveness \times h \times U} \tag{4.12}$$

Chapter 5

Results Discussion

In this chapter, the main takeaways of each of the previously described phases will be explained: the end result of the database utilization rate model, the forecasting accuracy results, and the final visualizations developed for the simulation tool, as well as the feedback received on them.

5.1 Utilization Rate Model

As explained in Chapter 4, the end goal of this phase was simultaneously to serve as a basis for the ensuing project stages and to create the data infrastructure necessary to report on this information.

The connection with the Simulation Tool proved to be crucial to the validation of this point as it allowed to test the Data Accuracy directly with the people that were the most familiar with it. This point will be further explained in the section relating to the Tool itself.

In this section, the focus will then be on presenting the end result as far as data organization is concerned, as well as some of the assumptions made and the potential errors associated with them.

5.1.1 Geographical Positioning

The resulting table from the previously explained calculations can be partially observed in Table A.2 from Appendix A.

Because of the existence of temporary areas, that, as described in Table 3.3, consist of a different set of coordinates and radius specified, for some technicians, for a specific day of the week, the attribution of technicians to municipalities would, in theory, have to be repeated with new daily records. To optimize storage space, however, the resulting information was organized in such a way that the relationship between technician and municipality would be established only once for its default configuration and then once for every weekday where the technician had a different location. From there, validity time intervals had to be set in case the technicians configuration changes over time.

This can be observed in the sample presented in Appendix A, where, column weekday is set with numbers from 0 to 6 corresponding to weekdays from Monday to Sunday, respectively. When a technician does not have a particular configuration for a certain weekday the default (7) applies.

As far as assumptions are concerned, it is also important to highlight that even with a threshold percentage having been defined, it was still observed that some technicians had more than one significant municipality under their reach. In this case, the assumption made was that the number of work hours of the technician should count as full capacity for all of those municipalities, as any assumptions regarding how this resource capacity would be split would not be representative of the potential flexibility existent. Picking up on the example displayed in Figure 3.2, the technician had three municipalities under their reach but, if, during one particular day, the WO distribution required it, they could dedicate all their work hours to a single one of those municipalities.

This, however, represents an approximation with associated error, that is expected to be reflected in large available capacity numbers, and, thus, lower utilization rates. To study and report on historical data, this can be prevented by increasing the used capacity for all municipalities of the technician regardless of where a WO occurred. However, when the study is done using forecasts, the problem will remain.

5.1.2 Work Schedule

Having performed the associations necessary to know the number of daily work hours of every technician, the result, as organized in the database, is sampled in Table A.3 from Appendix A.

The tabular organization presented for this calculation, unlike in the geographical positioning, had to be daily as technician schedules are established in a way that often means they have different start or finish times every weekday and the general timetable also changes depending on the week. This means that not many aggregations could be made that would actually simplify the problem.

As far as error driven by assumptions made is concerned, it is relevant of being mentioned here the fact that it was found some particular cases where, despite the integration between schedules and unavailability, still some technicians showed to be fully available and, looking at their WO history, ended up not performing any interventions during a full day. Although it may be the case that there were simply no WOs attributed to them, it was found that it was more frequently the case that this scenario resulted from an unforeseen technician absence, for which, therefore, no unavailability had been registered. Under this assumption, it was then, included in the algorithm a function that changed Δt (variable representing the number of work hours) to 0 in days when WO history showed no interventions had been performed by the technician in question.

It is, however, important to take into account that this refinement, similar to the one mentioned for geographical position, can, too, only be incorporated posteriorly and will, therefore, produce a discrepancy between the accuracy of the model when applied to historical data or to the outputs of a forecast.

5.1.3 Occupied Capacity

The occupied capacity, calculated as introduced in Chapter 4, was compiled into a WOs table from which a sample is presented in Table A.4 from Appendix A. Here, as can be seen, three DATETIME (dt) variables are used to calculate dislocation and execution times through Equations

(4.5) and (4.6), respectively. These variables are (displayed in seconds): $start\_dislocation\_dt$, $start\_execution\_dt$ and $end\_execution\_dt$.

5.1.4 Utilization Rate Calculation

As for the final integration of calculations, as mentioned before, it is meant to be adaptable depending on new groupings or assumptions intended, hence the View display.

Another way found of increasing the potential adaptability of the model to future applications was to reduce the utilization rate calculation to the technician as a basis before introducing the geographical component. The presentation of these results simply required the creation of a View where SCHEDULE_TECHNICIAN was joined with the WOs table – grouped by day and technician – and the application of Equation (4.7). A sample of the result of running this View for a specific day is displayed in Table 5.1.

Table 5.1: Sample of the results obtained for the daily technician utilization rate.

DAY	TECHNICIAN_ID	total_hours	real_hours	total_execution_duration	total_duration	utilization_rate
01/06/2021	987611136	8	8	3,17	5,49	0,68625
01/06/2021	987611139	8	8	1,67	2,06	0,2575
01/06/2021	987611140	8	8	4,2	5,44	0,68
01/06/2021	987619329	8	8	2,57	2,99	0,37375
01/06/2021	987619331	8	8	3,13	5,95	0,74375
01/06/2021	987619332	8	8	4,75	5,93	0,74125
01/06/2021	987619333	8	8	1	1,51	0,18875
01/06/2021	987619334	8	8	2,05	2,98	0,3725
01/06/2021	987627522	8	8	4,91	7,13	0,89125
01/06/2021	987676672	8	8	5,53	7,24	0,905

Although other values are shown, the columns currently used for the utilization rate calculation are the ones correspondent to the total work hours ($total\_hours$) – meaning the one where efficiency is not incorporated – for the available capacity, and the sum of execution and dislocation times ($total\_duration$) for the used capacity.

When integrating the geographical component, different groupings were applied so that it would be easily visible the available and used capacity depending on the categories WOs performed and technicians available fit into. As it has been referred, however, these categories are easily adaptable by simply adding or adjusting fields in the associated query.

Nonetheless, the model as it was here implemented allowed for visibility on the following values, by day and municipality:

- The **total available capacity** and the total available capacity after removing NOS non-outsourced workforce, only including the capacity provided by subcontracted SPs. This value could then be displayed by:
  - Providing SP
  - Technology – meaning, from the total amount of hours available, how many were worked by technicians who had the skills to work each kind of infrastructure

- The **total used capacity**, as well as, again, the used capacity in WOs only performed by the subcontracted workforce. This value was also then divided by:
  - Providing SP
  - Technology in which the WOs occurred
  - Type of Intervention
- The **utilization rate** that could then calculated both for the total subcontracted workforce as well as for each SP individually.

Furthermore, the number of technicians available daily by municipality was also included in the display. This value is a good indicator to confirm if the model is leading to a realistic reflection of what is happening in the field, as the amount of technicians operating in a certain area tends to be a value that KAMs, who are responsible for negotiating that number with SPs, are aware of. This way, assumptions made for geographical positioning could be confirmed.

5.2 Predictive Model

This section will focus on presenting the errors obtained for the Machine Learning model, which aimed at forecasting by Week, Municipality, Type of Intervention and Technology, the number of WOs that should be expected.

As explained in Chapter 4, because the model focused on a time series forecast, the approach to error estimation was a Time Series Cross-Validation as represented in Figure 4.4. Due to the transition between platforms, the data was estimated to start having some decent accuracy levels starting at the beginning of 2020. However, it is important to consider that the platform migration, for all SPs, was only fully concluded in September 2020 and, thus, since this analysis based itself on a data set collected up to May 2021, it is evidently early to conclude anything significant regarding the ability of the model to capture yearly seasonality.

Nonetheless, it was defined that the first week of January 2020 would be where the first training set would start, with the matching testing set being the first four weeks of that month the following year, and so on. The five folds performed are represented in Table 5.2.

Table 5.2: Time-Series Cross Validation: Summary of the folds tested.

Fold	Train Period	Train Records	Test Period	Test Records
1	2020-01-04 - 2021-01-02	53617	2021-01-02 - 2021-01-30	6219
2	2020-02-01 - 2021-01-30	59399	2021-01-30 - 2021-02-27	6330
3	2020-02-29 - 2021-02-27	65085	2021-02-27 - 2021-03-27	6194
4	2020-03-28 - 2021-03-27	70477	2021-03-27 - 2021-04-24	6013
5	2020-04-25 - 2021-04-24	74609	2020-04-25 - 2021-04-24	5981

From here two main metrics were studied: MAE and MAPE, with a stronger emphasis placed on the second one as it allows to understand the error regardless of fluctuations in the number of

WOs being estimated. Figure 5.1 displays graphically the results obtained for each fold and its mean value, representing the CVA.

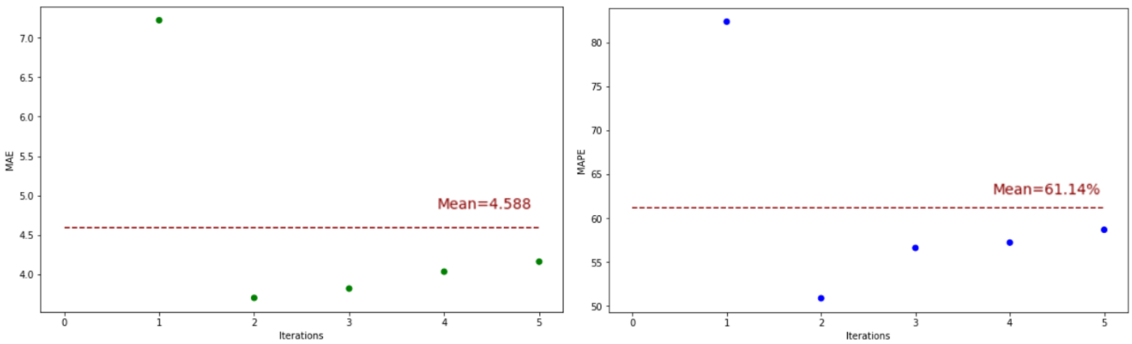

Figure 5.1: MAE and MAPE obtained for each fold and its average value.

As can be observed, the first fold clearly resulted in worse error metrics, which can be justified by the low amount of values in its training set. Hence, this first month was removed from the ensuing analysis, as it can be considered that the training set used was still not representative enough. The new MAE and MAPE values, without this point, were then 3,93 and 55,91%, respectively. These values can be seen in a similar graphic representation to the one above presented, but following this removal, which is shown in Figure B.1 of Appendix B.

The subsequent analysis then focused on identifying, from these aggregated error values, how they were divided through the municipalities. This analysis was deemed a priority because it can be intuitively understood that, due to the strong difference in population density between geographical regions and the resulting strong difference in the amount of weekly intervention requests, some regions may have a stronger impact on the error observed. Figure 5.2, then, provides the analysis of the relationship between the average number of WOs per week and the MAPE CVA value obtained for each municipality.

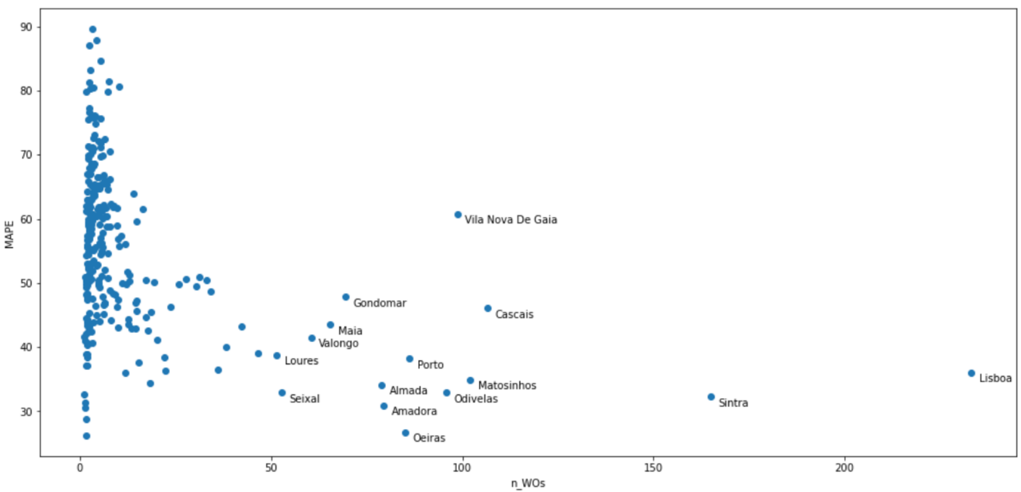

Figure 5.2: Relationship between the average number of WOs per week and the MAPE.

The labeled municipalities here are the ones where the average number of WOs per week surpasses 50, a number that was set through observation of where the points in the graph decreased in density. These municipalities belong to the biggest Portuguese metropolitan areas and, with the exception of Vila Nova de Gaia, resulted in relatively stable error values, with an average MAPE among them of 38,5%. Their individual values can also be observed in Table B.1 of Appendix B.

Figure 5.2 already provides an idea that the variability of results will be larger in smaller municipalities. Nonetheless, to analyze them more in-depth, the previously highlighted points were discarded in the subsequent analysis, which can be seen in Figure 5.3.

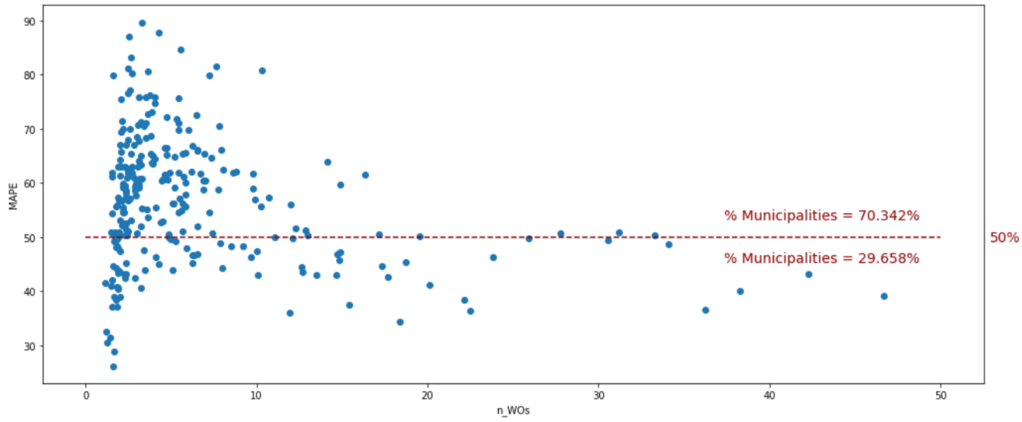

Figure 5.3: Relationship between the average number of WOs per week and the MAPE, for smaller municipalities.

As can be visualized here, the vast majority of the municipalities obtained a MAPE of over 50%, a result that is not very promising as far as these municipalities are concerned. A big concentration of these points with larger errors, is, however, located, in the graph, in the area corresponding to very small municipalities, with around ten or less WOs per week. This can mean that the high variability here visible, although high in percentage, originates from a considerably small absolute error. This discrepancy between different sized municipalities can be studied through the aggregation of small (with 10 or less WOs per week) and medium sized municipalities (with between 10 and 50 WOs per week) into average points, so they can be directly compared to larger ones. This analysis, shown in Figure B.2 of Appendix B, proves that, in general, medium sized municipalities obtained fairly decent results when compared to smaller ones.

The error calculation can, then, be complemented using WMAPE, to understand the overall performance of the model when given weighted importance to municipalities depending on their magnitude. This value, given the data display used here, can be approximated by:

$$WMAPE \approx \frac{\sum_{m=1}^{M} (MAPE_m \times n\_WOs_m)}{\sum_{m=1}^{M} n\_WOs_m} , \tag{5.1}$$

with $m \in [1, M]$ as the set of municipalities

The application of this formula lead to a total WMAPE of 46,12% for the general country-wide predictions, as opposed to 38,09% when applied only to the set of urban municipalities above mentioned.

Overall, it can be concluded that the model still needs some improvements and a further analysis done once there is enough fully reliable data, which should be expected by the end of 2021. Nonetheless, it did show promising results for more populated areas, although the same cannot be said for smaller ones, where it is likely that additional predictors should be included in order to decrease the variability in performance. Regarding this topic, a hypothesis worth testing in future work is that, these smaller areas, being mostly rural municipalities and, thus, more predominant in DTH Technology interventions, may require that the model uses weather information as a predictor, due to the technological implications of a satellite solution.

5.3 Simulation Tool

Finally, this section aims at showing the final visualizations made available in the simulation tool. Feedback received by its end users is also included here.

5.3.1 Visualization

In addition to an introductory sheet that explains briefly the capabilities of the model, this tool, at its core, encompasses five sheets: two for the user to insert complementary information necessary for the displays and three of which are returning results. Table 5.3 summarizes their content.

Table 5.3: Summary of the Sheets presented in the simulation tool.

Category	Sheet Name	Description
Input Sheets	TARGET AIT	For every WO Type, defined by a Type of Intervention and Technology, which SLA is the goal.
	WO PREDICTIONS	For every municipality and WO Type, the user can change the number of predicted WOs by setting a ΔQ_i
Output Sheets	AIT REPORT	AIT records, weighted against the SLA defined, so that performance in each municipality is visible.
	TECHNICIAN SIMULATION	User can set, for specific municipalities, the ΔT to be simulated and Responsiveness is, then, calculated.
	RESPONSIVENESS SIMULATION	User can set, for specific municipalities, Target Responsiveness and the necessary workforce is, then, calculated.

The three output sheets will be represented here, while the input ones can be found Appendix C. The first one, corresponding to AIT REPORT, as mentioned, corresponds solely to a report on past performance, compared against the inputs inserted in TARGET AIT (Figure C.1 of Appendix C) and an example of it, filtered for two municipalities, can be seen in Figure 5.4.

Semanas do Cálculo
Histórico
Semana Atual

SP
FUTURCABO
GIZZ
MMCI
PDT
SINALCABO
(blank)

TECN
FIBRA DST
SATELITE
CABO
FIBRA NOS
FIBRA SWAP

MUNICIPIO
Almeida
Almeirim
Almodôvar
Alpiarça
Alter Do Chão
Alvaiázere
Alvito
Amadora
Amarante
Amares
Anadia
Ansião
Arcos De Valdevez

MUNICIPIO	TECN	TIPO_INTERVENCAO	N_Técnicos	Target TMI_calculado	TMI_calculado_
Arcos De Valdevez	FIBRA DST	AS	4	3,000	4,437
Arcos De Valdevez	FIBRA DST	I+R	4	5,000	8,882
Arcos De Valdevez	FIBRA DST	M	4	3,000	2,555
Arcos De Valdevez	FIBRA DST Total		4	4,182	6,550
Arcos De Valdevez	SATELITE	AS	4	4,000	4,501
Arcos De Valdevez	SATELITE	I+R	4	7,000	1,380
Arcos De Valdevez	SATELITE	M	4	4,000	1,058
Arcos De Valdevez	SATELITE Total		4	4,750	1,999
Arcos De Valdevez Total			4	4,333	5,337
Ponte De Lima	FIBRA DST	AS	5	3,000	5,390
Ponte De Lima	FIBRA DST	I+R	5	5,000	20,111
Ponte De Lima	FIBRA DST	M	5	3,000	3,024
Ponte De Lima	FIBRA DST Total		5	3,458	7,137
Ponte De Lima	SATELITE	I+R	5	7,000	2,319
Ponte De Lima	SATELITE	M	5	4,000	1,310
Ponte De Lima	SATELITE Total		5	5,714	1,887
Ponte De Lima Total			5	3,745	6,469

Calcular

Figure 5.4: Filtered AIT REPORT sheet.

Although this is not the main focus of the tool, this view is necessary as an introduction both because it allows the user to identify improvement targets and because it is the only way of checking disaggregated performance by Type of Intervention and Technology. Because there is no computable way of establishing prioritization between WO types, simulations can only be done for the aggregated total of each municipality.

As for the simulations, as previously mentioned, both of them use any inputs inserted in the WO PREDICTIONS sheet, shown in Figure C.2 of Appendix C, in addition to the inputs in the simulation sheets.

From there, TECHNICIAN SIMULATION, shown in Figure 5.5, displays the current workforce and allows users to plan any variations in it, using Equation (4.11) to calculate Responsiveness and the marginal improvement of this KPI.

MUNICIPIO	Nº TÉCNICOS	VARIAÇÃO
Abrantes	7	5
Aveiro	49	2
Amadora	55	-20
Ponte de Lima	5	-3
	0	
	0	
	0	
	0	
	0	
	0	
	0	
	0	
	0	

Calcular

Número de horas diárias de trabalho p/ técnico
8

MUNICIPIO	Duração Total	Nº Total OTs Esperadas	DMI Esperado	Nº Técnicos	Ocupação dos Técnicos	Capacidade Diária (horas)	Capacidade de Resposta Original	Nova Capacidade de Resposta (dias)	Melhoria
Abrantes	113,80	83,50	1,36	12,00	0,81	96,00	1,74	1,19	32%
Albufeira	146,18	116,75	1,25	0,00	0,93	0,00	1,54	0,00	100%
Amadora	647,78	472,75	1,37	35,00	0,81	280,00	1,36	2,31	-71%
Aveiro	296,95	193,50	1,53	51,00	0,79	408,00	0,76	0,73	4%
Cascais	858,28	610,75	1,41	0,00	0,80	0,00	2,67	0,00	100%
Faro	220,19	144,50	1,52	0,00	0,98	0,00	1,60	0,00	100%
Ovar	148,10	85,25	1,74	0,00	0,87	0,00	1,26	0,00	100%
Ponte De Lima	92,58	56,75	1,63	2,00	0,99	16,00	2,31	5,79	-150%
0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0%
0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0%
0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0%
0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0%
0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0%

Figure 5.5: TECHNICIAN SIMULATION sheet.

The visualization is color-coded in such a way that blue columns correspond to user-defined parameters, while white ones represent model calculated fields and the red corresponds to the final results that are the most crucial for the clients of the model to analyze.

The last relevant output sheet consists of the RESPONSIVENESS SIMULATION, which, as previously explained, basically does the reverse process: assuming a target responsiveness, it returns the minimum amount of technicians necessary in order to achieve that goal, calculated through Equation (4.12). Its visual presentation is shown in Figure 5.6.

MUNICIPIO	Capacidade de Resposta Desejada (dias)
Abrantes	1
Aveiro	1
Ponte de Lima	3
Mafra	2

Capacidade de Resposta Desejada Default (dias)	Número de horas de trabalho p/ técnico
5	8

Calcular

MUNICIPIO	Duração Total	Nº Total OTs Esperadas	DMI Esperado (horas)	Ocupação dos técnicos	Capacidade de resposta desejada (dias)	Capacidade necessária (horas/dia)	Nº Técnicos Atual	N Técnicos Necessários
Abrantes	113,80	83,50	1,36	0,81	1,00	113,80	7,00	15
Albufeira	146,18	116,75	1,25	0,93	5,00	29,24	11,86	4
Amadora	647,78	472,75	1,37	0,81	5,00	129,56	55,00	17
Aveiro	296,95	193,50	1,53	0,79	1,00	296,95	49,00	38
Cascais	858,28	610,75	1,41	0,80	5,00	171,66	40,00	22
Faro	220,19	144,50	1,52	0,98	5,00	44,04	10,00	6
Mafra	576,14	296,25	1,94	0,73	2,00	288,07	17,00	37
Ovar	148,10	85,25	1,74	0,87	5,00	29,62	12,00	4
Ponte De Lima	92,58	56,75	1,63	0,99	3,00	30,86	5,00	4
0	0,00	0,00	0,00	0,00	5,00	0,00	0,00	0
0	0,00	0,00	0,00	0,00	5,00	0,00	0,00	0
0	0,00	0,00	0,00	0,00	5,00	0,00	0,00	0
0	0,00	0,00	0,00	0,00	5,00	0,00	0,00	0
0	0,00	0,00	0,00	0,00	5,00	0,00	0,00	0
0	0,00	0,00	0,00	0,00	5,00	0,00	0,00	0

Figure 5.6: RESPONSIVENESS SIMULATION sheet.

As previously stated, this second simulation sheet can be used based on the assumption that there is a proportional relationship between AIT and Responsiveness. Therefore, if, in the AIT REPORT, a need for a certain marginal improvement is identified for a certain aggregated municipality, that can be transposed into target responsiveness, by applying the same marginal improvement to this metric.

5.3.2 User Feedback

Interviews and discussions were held with the KAMs – the end users of this tool – in order to understand the success of this end result of the project.

The feedback process was directed in such a way that would cover three key areas to be assessed: accuracy of the data model, simulation capabilities and user experience provided by the tool.

The first area is not solely connected with this stage of the project but also with what was done in the utilization rate model. It was, nonetheless, necessary to be included here, as this initial phase had no values to be tested against and, therefore, to evaluate if calculations were correctly performed, the best approach was to check the values obtained directly with the people with field knowledge on what was the situation in the municipalities of their accountability.

For each of these feedback components the main takeaways worth highlighting were the following:

1. **Data Accuracy:**

   Initially, some problems were detected with the values presented that were originating from the utilization rate model. Particularly, it was identified that some cells were not being associated with any municipality and, thus, KAMs could detect that the historical number of WOs was lower than the reality of each of these geographical areas. The number of technicians associated with each municipality also had to be corrected due to a problem in the View that presented it.

   These problems were iteratively fixed, however, and in the end, both WO records and technician geographical positioning data were deemed reliable and, consequently, the assumptions made for each of them could be validated.

2. **Simulation Capabilities:**

   The feedback here was generally positive, as the main objective of simulating impacts in response time caused by alterations in the regional workforce was achieved.

   There was a general understanding that it would have been beneficial, for the operations of the KAMs, that it was possible to both visualize calculated Responsiveness and establish its Target value disaggregated by WO type. However, as previously explained, that could not be mathematically achieved unless assumptions were made regarding how *Click* establishes prioritization between interventions, which could have led to big estimation errors. It can, nonetheless, be deemed as an improvement point as long as investigative work is done to establish a basis for these prioritization assumptions.

3. **User Experience:**

   Between the three dimensions, this was the one where the model showed a bigger improvement margin as it was determined that the understanding of its capabilities required some level of explanation. The tool should, then, continue to be improved, to the point where it is manageable regardless of previous introduction to its functionalities.

   Particularly, a specific improvement point identified was the relationship mentioned between AIT and Responsiveness. A more intuitive connection between the AIT REPORT and the RESPONSIVENESS SIMULATION should be built, in order for new users that may arrive to intuitively understand that by utilizing the two combined they are able to plan a response that will improve the performance reported.

Chapter 6

Conclusions and Future Work

In this final chapter, a reflection on the project from the optic of the results obtained is performed. Future potential improvements are also presented, in hopes that the work done will serve as a basis for an extensive analysis on the topic.

6.1 Conclusion

The problem under study through this project was compelling, not only as a research matter but, foremost, for its importance for the business context in which the case study was performed. From the initial analysis of the capacity planning processes, it became evident that there was an issue of lack of information and, consequently, of bargaining power on this matter.

This project, therefore, focused on bringing to light information that could be relevant when making workforce capacity planning decisions. In order to do so, the work developed focused on the connection of the different dimensions that compose a successful Business Intelligence system (historical, current and predictive) and explored potential improvements to the data used in resource planning through each of these perspectives.

This way, information available was expanded through both the exploration of new reporting opportunities for complementary KPIs and the proposal of a demand forecasting Machine Learning model. The proposed interconnection between both these data sources and an interface for decision support was also presented through this work, showing how current performance and future predictions should be mutually complementary in the decision making process, and how a BI system can aid this process.

Nonetheless, a challenge encountered, within the execution of this system, was the absence of crucial data necessary to establish the current situation regarding the geographical placement and work time of field technicians. Without these two components being fully known, there is not sufficient data infrastructure to report and study available capacity. Thus, the project timeline had to be, from the beginning, adjusted to incorporate these data acquisition requirements.

This, along with the results obtained for the Machine Learning model, justify the fact that the full integration of all data obtained could not be achieved, through the deployment of the model during the extent of this project. Although this is not deemed as a failure of the project, as the core objectives were achieved and all components were introduced, it is important to note that the project was idealized as one that would set the basis for a single concise deliverable. Ideally, then, it is hoped that future versions of the Simulation Tool incorporate the forecasts created so that decision-makers can truly work with scenarios realistic to what is expected to occur on the field.

As for each of the phases of the project, individually, starting with the Utilization Rate Modelling: it can be stated that this step was successful as the main goal was achieved – the company has, now, the database structures required to report and measure this performance metric, which is fundamental has a complementary KPI that should be weighed against the service level ones presented in Chapter 3. From what was gathered from the Feedback interviews, the model is reporting the necessary information and within data intervals that are realistic to the field reality.

Regarding the Machine Learning Forecasting Model, the results leave a need for strong improvements before data deployed can be fully trusted to be used as a support for decision making. Some of the inaccuracy found may, however, be profoundly connected with data requirements not having been fully met, as reliable historical records of over a year could not yet be incorporated in the analysis due to the platform migration of all SPs only having been concluded in September 2020 and the tests having been performed using data up to May 2021, so the ability of the model to capture potential effects of yearly seasonality was limited.

Nonetheless, with an average MAPE of 38,5% and a WMAPE of 38,1%, the potentialities of the model are especially encouraging for the prediction of interventions to be executed in urban areas, where the error metrics showed more stable results. As for more rural areas, a broader range of predictors may be necessary to be used in order to decrease variability.

The final deliverable of the project was a crucial point to achieve its essential goal of empowering stakeholders involved in capacity planning decisions with the necessary information to do so. For the development of this Simulation Tool, there was a strong trade-off between usability and accuracy faced, however, as the first version of a platform with these capabilities, usability had to be prioritized. For this reason, the tool took the form of an Excel Workbook, where a few assumptions had to be made that could lead to some rough approximations. It was, nonetheless, well received among its users who validated the capabilities incorporated in the tool. Indeed, they, now, have a structured way of analyzing the technician distribution and simulating reallocation scenarios for these resources, throughout the municipalities and SPs under their accountability, something which they deemed of major importance for the operations impacted.

6.2 Future Improvements

Finally, throughout the course of this project, some improvement points were identified, that could be valuable for the business and are, thus, worth highlighting here.

Firstly, it can be stated that direct improvements in the reporting accuracy of the Utilization Rate model could be achieved by cooperating with the IT Department to ensure that all data requirements are diffused whenever alterations occur to them, as opposed to the Data Extractions solution found. Although some assumptions are also made through the calculations of this stage, this is believed to be the most crucial point, since it could not only create stronger accuracy for the model built, but also open other valuable reporting opportunities for the company.

Still regarding the Utilization Rate model, this project leaves open the possibility of extending its use for reporting applications. A possibility that should be explored is the integration of the data obtained within a Data Visualization report, specifically using PowerBI software, which is already widely in usage within the company.

As for the Machine Learning Forecasting Tool, specific improvements that may constitute strong contributions to achieving better accuracy, and should, therefore, be attempted in the future, can be enumerated as:

1. Incorporating weather information as a predictor and testing its impact on accuracy, particularly on the prediction of DTH interventions.
2. Studying the possibility of instead of using the number of WOs as the output variable, applying the model to the prediction of total WO duration, since that value will more directly impact capacity requirements. In order to obtain accurate predictions, it may be required a further understanding of the technological implications of interventions than what was under the scope for this study.

It is, also, hoped that this forecasting model will remain under use and is revised once a broader set of historical records is available, and, if deemed reliable at that point, it is deployed following the recommendations given in Chapter 4 – with the centralization, in DW, of the information regarding these predictions and the ones developed in the Commercial context.

Furthermore, for all stages of this project, the assumption that records corresponding to Cancelled interventions should be eliminated from the analysis, and only Closed interventions should be considered, was constant. This assumption, nevertheless, raises some questions as to its impact on the model accuracy, as before acquiring this status, Cancelled interventions highly influence the assignment process and technician availability. A further study on this matter may be an interesting topic of research.

Although some adjustments could be made as feedback was collected, it is important to note that the incorporation of a tool in the work processes, should always be an iterative process, where its users, after integrating it within their routine, can gradually realize and suggest improvement points. This is a process that should be carried out and evaluated in the long term and that, therefore, could only be initiated with the study here presented.

Bibliography

Akkermans, H. and Vos, B. (2003). Amplification in service supply chains: An exploratory case study from the telecom industry. *Production and Operations Management*, 12(2):204–223.

APRITEL (2019). Análise de preços das comunicações eletrónicas na Europa. Technical report.

Arnott, D., Lizama, F., and Song, Y. (2017). Patterns of business intelligence systems use in organizations. *Decision Support Systems*, 97:58–68.

Ashraf, S. and Khan, S. A. (2015). Visualizations-based analysis of Telco data for business intelligence. In *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, volume 2015-Novem, pages 242–246. IEEE Computer Society.

Azevedo, A. and Santos, M. F. (2008). KDD, semma and CRISP-DM: A parallel overview. In *IADIS European Conference on Data Mining*, Amsterdam.

Bose, I. and Mahapatra, R. K. (2001). Business data mining - A machine learning perspective. *Information and Management*, 39(3):211–225.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brentan, B. M., Meirelles, G., Herrera, M., Luvizotto, E., and Izquierdo, J. (2017). Correlation Analysis of Water Demand and Predictive Variables for Short-Term Forecasting Models. *Mathematical Problems in Engineering*, 2017.

Bughin, J. (2016). Reaping the benefits of big data in telecom. *Journal of Big Data*, 3(1):1–17.

Castillo-Salazar, J. A., Landa-Silva, D., and Qu, R. (2016). Workforce scheduling and routing problems: literature survey and computational study. *Annals of Operations Research*, 239(1):39–67.

Chen, C. M. (2016). Use cases and challenges in telecom big data analytics. *APSIPA Transactions on Signal and Information Processing*, 5:1–7.

Chen, H., Chiang, R. H., and Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems*, 36(4):1165–1188.

Cordeau, J. F., Laporte, G., Pasin, F., and Ropke, S. (2010). Scheduling technicians and tasks in a telecommunications company. *Journal of Scheduling*, 13(4):393–409.

Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). Random Forests. In *Ensemble Machine Learning*, pages 157–175. Springer US, Boston, MA.

Davenport, T. H. (2005). Competing on Analytics. *hosteddocs.ittoolbox.com*.

Decerle, J., Grunder, O., Hajjam El Hassani, A., and Barakat, O. (2017). A general model for the home health care routing and scheduling problem with route balancing. *IFAC-PapersOnLine*, 50(1):14662–14667.

DGT (2020). Mapas Online. May 19, 2021. http://mapas.dgterritorio.pt/viewer/index.html.

Dutot, P.-F., Laugier, A., and Bustos, A.-M. (2006). Technicians and Interventions Scheduling for Telecommunications. Technical report, France Telecom R&D.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3):37–37.

Froger, A., Gendreau, M., Mendoza, J. E., Pinson, E., and Rousseau, L. M. (2018). Solving a wind turbine maintenance scheduling problem. *Journal of Scheduling*, 21(1):53–76.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier Inc.

Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., Dai, W., Yang, Q., Zeng, J., Noah, H., Lab, A., and Kong, H. (2015). Telco Churn Prediction with Big Data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 607–618, New York, NY, USA. ACM.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Jain, P., Gyanchandani, M., and Khare, N. (2018). Differential privacy: its technological prescriptive using big data. *Journal of Big Data*, 5(1):1–24.

Kassem, S., Hagras, H., Owusu, G., and Shakya, S. (2012). A type2 fuzzy logic system for workforce management in the telecommunications domain. In *IEEE International Conference on Fuzzy Systems*.

Kern, M., Shakya, S., and Owusu, G. (2009). Integrated resource planning for diverse workforces. In *2009 International Conference on Computers and Industrial Engineering, CIE 2009*, pages 1169–1173. IEEE Computer Society.

L. van Eck, M., Firat, M., P. M. Nuijten, W., Sidorova, N., and M. P. van der Aalst, W. (2017). Human Performance-Aware Scheduling and Routing of a Multi-Skilled Workforce. *Complex Systems Informatics and Modeling Quarterly*, (12):1–21.

Lingjun, H., Levine, R. A., Fan, J., Beemer, J., and Stronach, J. (2018). Random Forest as a Predictive Analytics Alternative to Regression in Institutional Research. *Practical Assessment, Research, and Evaluation*, 23.

Mastorocostas, P. and Hilas, C. (2012). A computational intelligence-based forecasting system for telecommunications time series. *Engineering Applications of Artificial Intelligence*, 25(1):200–206.

Misir, M., Lau, H. C., Mısır, M., and Hoong (2014). Diversity-Oriented Bi-Objective Hyper-heuristics for Patrol Scheduling. Technical report.

Mohamed, A., Hagras, H., Shakya, S., and Owusu, G. (2012). Tactical resource planner for workforce allocation in telecommunications. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7326 LNAI, pages 87–94. Springer, Berlin, Heidelberg.

Olson, D. L. and Delen, D. (2008). *Advanced data mining techniques*. Springer Berlin Heidelberg.

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. *Data Stream Mining and Processing*, 4(1):21–25.

Pekel, E. (2020). Solving technician routing and scheduling problem using improved particle swarm optimization. *Soft Computing*, 24(24):19007–19015.

Pereira, D. L., Alves, J. C., and Moreira, M. C. d. O. (2020). A multiperiod workforce scheduling and routing problem with dependent tasks. *Computers and Operations Research*, 118:104930.

Philips-Wren, G., Daly, M., and Burnstein, F. (2021). Reconciling business intelligence, analytics and decision support systems: More data, deeper insight. *Decision Support Systems*, page 113560.

Romero, Á., Dorronsoro, J. R., and Díaz, J. (2019). Day-Ahead Price Forecasting for the Spanish Electricity Market. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(4):42.

Snyder, J. P. (1987). Map Projections: A Working Manual. Technical report, U.S. Government Printing Office, Washington, D.C.

Thelen, S., Mottner, S., and Berman, B. (2004). Data mining: On the trail to marketing gold. *Business Horizons*, 47(6):25–32.

Vairagade, N., Logofatu, D., Leon, F., and Muharemi, F. (2019). Demand forecasting using random forest and artificial neural network for supply chain management. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11683 LNAI, pages 328–339. Springer Verlag.

Xu, Y., Zhou, Y., Sekula, P., and Ding, L. (2021). Machine learning in construction: From shallow to deep learning. *Developments in the Built Environment*, 6:100045.

Appendix A

Utilization Rate Modeling

For Geographical Positioning, it was necessary to establish, first the equivalence between cells as the smallest geographical unit used, its coordinates and the municipality they belong to.

Table A.1: Sample of INFO_CELLS table, containing the information regarding the geographical positioning of the cells.

Cell_name	AVG_LATITUDE	AVG_LONGITUDE	NoHUs	MUNICIPALITY	dt_calculation	dt_loading
LIS00	38,73903	-9,149156	628635	Lisboa	05/03/2021 13:10	05/03/2021 13:40
SIN00	38,787466	-9,317492	317570	Sintra	05/03/2021 13:10	05/03/2021 13:40
POT00	41,160122	-8,619056	289369	Porto	05/03/2021 13:10	05/03/2021 13:40
GA00	41,099003	-8,607023	274928	Vila Nova De Gaia	05/03/2021 13:10	05/03/2021 13:40
ALM00	38,651974	-9,180542	250720	Almada	05/03/2021 13:10	05/03/2021 13:40
CAI00	38,710638	-9,383084	218955	Cascais	05/03/2021 13:10	05/03/2021 13:40
LOU00	38,820731	-9,135429	198829	Loures	05/03/2021 13:10	05/03/2021 13:40
C00	40,208918	-8,430239	179924	Coimbra	05/03/2021 13:10	05/03/2021 13:40
LR00	39,762918	-8,807335	169994	Leiria	05/03/2021 13:10	05/03/2021 13:40
BG00	41,548494	-8,4217	166249	Braga	05/03/2021 13:10	05/03/2021 13:40

The compilation of the information regarding which technicians cover each municipalities in which days can be seen summarized below. In the weekday column, numbers 0 to 6 represent days from Monday to Sunday respectively, while 7 is the default configuration of the technician, for days when they do not work temporary areas.

Table A.2: Sample of ACTION_TECHNICIAN_MUNICIPALITIES table, containing the information regarding the geographical distribution of technicians.

TECHNICIAN_ID	weekday	MUNICIPALITY	NoHUs_Municipality	NoHUs_Total	percentHUs_Municipality	valido_de	valido_ate	dt_calculation
1182375960	0	Lisboa	2231351	2317386	96%	13/04/2021 15:54	31/12/2900 00:00	03/06/2021 01:00
1182375960	1	Lisboa	221341	1061638	21%	13/04/2021 15:54	31/12/2900 00:00	03/06/2021 01:00
1182375960	1	Loures	516212	1061638	49%	13/04/2021 15:54	31/12/2900 00:00	03/06/2021 01:00
1182375960	1	Odivelas	324085	1061638	31%	13/04/2021 15:54	31/12/2900 00:00	03/06/2021 01:00
1182375960	2	Lisboa	1708781	2232585	77%	13/04/2021 15:54	31/12/2900 00:00	03/06/2021 01:00
1182375960	3	Lisboa	420905	1340866	31%	13/04/2021 15:54	31/12/2900 00:00	03/06/2021 01:00
1182375960	3	Loures	653512	1340866	49%	13/04/2021 15:54	31/12/2900 00:00	03/06/2021 01:00
1182375960	3	Odivelas	266449	1340866	20%	13/04/2021 15:54	31/12/2900 00:00	03/06/2021 01:00
1182375960	4	Loures	72316	409283	18%	13/04/2021 15:54	31/12/2900 00:00	03/06/2021 01:00
1182375960	4	Odivelas	321013	409283	78%	13/04/2021 15:54	31/12/2900 00:00	03/06/2021 01:00
1182375960	7	Loures	288869	320529	90%	04/09/2020 00:00	22/12/2020 00:00	03/06/2021 01:00
1182375960	7	Lisboa	927319	1469097	63%	23/12/2020 00:00	31/12/2900 00:00	03/06/2021 01:00
1182375960	7	Loures	339671	1469097	23%	23/12/2020 00:00	31/12/2900 00:00	03/06/2021 01:00

The next step was then the Work Schedule equivalence. The following table summarizes the number of hours worked by each technician in each day after joining together the information concerning their schedules and their planned unavailability times.

Table A.3: Sample of SCHEDULE_TECHNICIAN table, containing the information regarding the daily schedule of technicians.

DAY	TECHNICIAN	SP	efficiency	start_time	finish_time	total_hours	real_hours	index	dt_calculations
19/04/2021	1072504844	PDT	0,9	19/04/2021 09:00	19/04/2021 20:00	10	9	c21e33cfea44548a5da61e32d82dc00d	03/06/2021 01:00
19/04/2021	1072513027	PDT	0,8	19/04/2021 09:00	19/04/2021 20:00	10	8	e6b0cddb7e2687059d5eb3d505808a96	03/06/2021 01:00
19/04/2021	1415143426	PDT	0,9	19/04/2021 09:00	19/04/2021 20:00	10	9	b5c94bea7f71a9454db89f91d7ebba6e	03/06/2021 01:00
20/04/2021	1072504844	PDT	0,9	20/04/2021 09:00	20/04/2021 20:00	10	9	c4e20aad8637fc1d450e59d80fdd6829	03/06/2021 01:00
20/04/2021	1072513027	PDT	0,8	20/04/2021 09:00	20/04/2021 20:00	10	8	27129e3e6963a1034ef337b6ca533167	03/06/2021 01:00
20/04/2021	1415143426	PDT	0,9	20/04/2021 09:00	20/04/2021 20:00	10	9	d6fe647162294de06ef15b8186de6623	03/06/2021 01:00
21/04/2021	1072504844	PDT	0,9	21/04/2021 09:00	21/04/2021 20:00	10	9	9136b7c4e6007264daf5928b39d34452	03/06/2021 01:00
21/04/2021	1415143426	PDT	0,9	21/04/2021 09:00	21/04/2021 20:00	10	9	7df2924c4981df949eb010f3abd150f9	03/06/2021 01:00
22/04/2021	1072504844	PDT	0,9	22/04/2021 09:00	22/04/2021 20:00	10	9	52f4bc0041671b48c12799fbdbdd8f60	03/06/2021 01:00
22/04/2021	1415143426	PDT	0,9	22/04/2021 09:00	22/04/2021 20:00	10	9	ea0a90a16f074a2adc6a7e8f9226e7dc	03/06/2021 01:00

Finally, the WOs table incorporates a copy of the ACTIVITIES one with the columns deemed relevant for this application plus the Municipality equivalence and the dislocation and execution times calculated based on the timestamps of the respective status changes.

Table A.4: Sample of WOs table, containing the information regarding the WO records, their location and duration.

ACTIVITY_ID	WO_Type	TECNOLOGY	ACTIVITY_STATUS	TECHNICIAN_ID	SP	Cell_name	MUNICIPALITY	Planned_Duration
1-GT495LI	Installation	FTTH_v3	Not Fulfilled	1031200770	Sinalcabo	YSB20	Almada	9000
1-GZZJ628	Installation	HFC	Fulfilled	1002332160	Futurcabo	CAN01	Odivelas	5400
1-H097JAI	Installation	HFC	Not Fulfilled	1251155987	Sinalcabo	AMA18	Amadora	5400
1-H0K7FMU	Installation	FTTH_v3	Fulfilled	1062141967	PDT	YBJ01	Beja	9000
1-H0L54AB	Installation	FTTH_v3	Fulfilled	1221664798	Sinalcabo	YCP09	Almada	9000
1-H0RZ4S0	Reestablishment	HFC	Fulfilled	1124909106	PDT	EP17	Espinho	5400
1-H148AI8	Maintenance	HFC	Fulfilled	1084096531	PDT	FOZ06	Porto	2700
1-H17CRCI	Installation	HFC	Não realizada	1434484777	PDT	MT04	Matosinhos	5400

first_creation_dt	last_click_alteration_dt	start_dislocation_dt	start_execution_dt	end_execution_dt	dislocation_duration	execution_duration	total_duration	dt_calculation
08/04/2021 12:50	01/06/2021 15:21	28/05/2021 11:24	01/06/2021 15:13	01/06/2021 15:21	2098	480	2578	03/06/2021 01:40
21/05/2021 20:59	01/06/2021 10:07	31/05/2021 16:38	01/06/2021 09:17	01/06/2021 10:07	563	3000	3563	03/06/2021 01:40
24/05/2021 16:50	01/06/2021 14:02	31/05/2021 14:11	01/06/2021 13:42	01/06/2021 14:02	1403	1200	2603	03/06/2021 01:40
26/05/2021 15:10	01/06/2021 13:11	29/05/2021 13:52	01/06/2021 10:01	01/06/2021 13:11	2334	11400	13734	03/06/2021 01:40
26/05/2021 16:48	01/06/2021 16:46	31/05/2021 13:18	01/06/2021 13:38	01/06/2021 16:46	2592	11280	13872	03/06/2021 01:40
27/05/2021 17:12	01/06/2021 16:03	31/05/2021 14:33	31/05/2021 15:25	01/06/2021 16:03	3067	5400	8467	03/06/2021 01:40
30/05/2021 15:30	01/06/2021 10:26	31/05/2021 08:48	01/06/2021 10:00	01/06/2021 10:26	1272	1560	2832	03/06/2021 01:40
31/05/2021 14:11	01/06/2021 08:12	31/05/2021 15:43	31/05/2021 15:52	01/06/2021 08:12	497	5400	5897	03/06/2021 01:40

Appendix B

Forecasting Accuracy Results

Having removed the first fold, the MAE and MAPE were recalculated as the following:

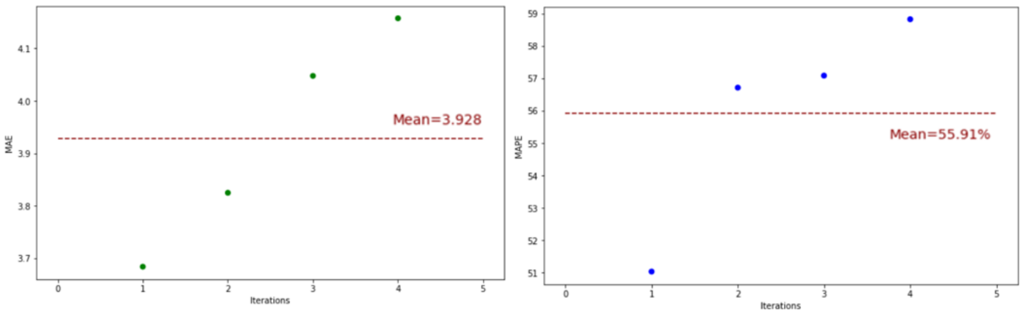

Figure B.1: MAE and MAPE obtained after removal of the first fold.

Additionally, one can see here in detail, for the bigger municipalities – where number of WOs greater than 50 – their individual MAPE results obtained.

Table B.1: Summary of the accuracy results obtained for large municipalities.

Municipality	MAPE	# WOs/Week
Almada	34.10%	78.98
Amadora	30.88%	79.52
Cascais	46.07%	106.72
Gondomar	47.81%	69.41
Lisboa	36.07%	233.15
Loures	38.79%	51.54
Maia	43.48%	65.51
Matosinhos	34.95%	102.09
Odivelas	32.91%	95.72
Oeiras	26.74%	85.18
Porto	38.21%	86.18
Seixal	32.92%	52.69
Sintra	32.25%	164.98
Valongo	41.50%	60.47
Vila Nova De Gaia	60.79%	98.74

In order to see more generally how the small and medium sized municipalities, characterized by 10 or less WOs per week and between 10 and 50 WOs per week respectively behaved when compared to the larger ones that could be isolated in Figure 5.2, these were aggregated into average points for each of the groups. The results show that medium sized municipalities have a performance comparable to the larger ones and the main source of variability is, indeed, the smaller ones.

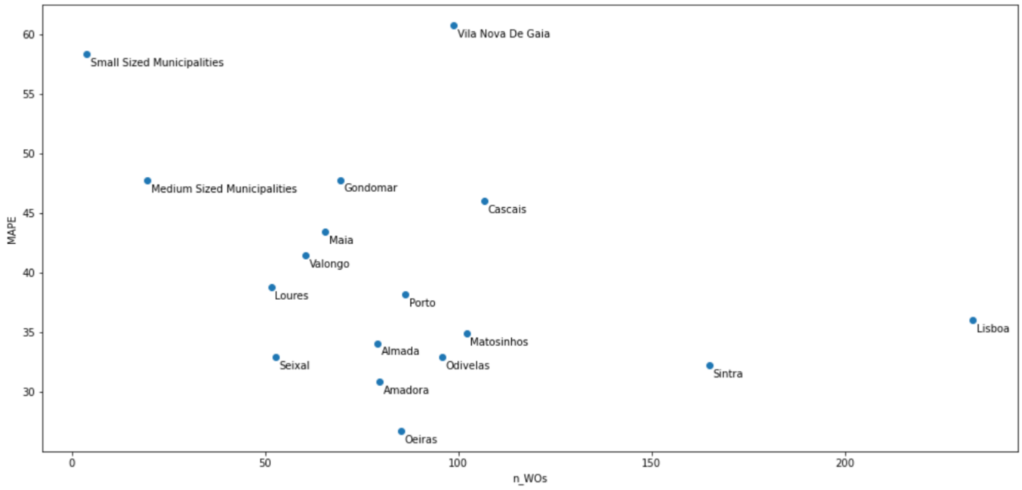

Figure B.2: Relationship between the average number of WOs per week and the MAPE with Small and Medium Sized municipalities aggregated into their average points.

Appendix C

User Inputs of the Simulation Tool

The TARGET AIT sheet consists of a single table with all existing WO types – all combinations of Type of Intervention and Technology – where the user can define the business objective for each of them so that then the report on the performance presented having this limit as a reference.

Tecnologia	Tipo de OT	Target TMI
CABO	AS	2,5
FIBRA DST	AS	3
FIBRA NOS	AS	1
FIBRA SWAP	AS	1
SATELITE	AS	4
CABO	M	3
FIBRA DST	M	3
FIBRA NOS	M	
FIBRA SWAP	M	
SATELITE	M	4
CABO	I+R	
FIBRA DST	I+R	5
FIBRA NOS	I+R	
FIBRA SWAP	I+R	
SATELITE	I+R	7

Figure C.1: TARGET AIT sheet.

As for the WO PREDICTIONS it is composed of a table where the user can, for any municipality and WO type, define a variation in the number of weekly WOs that are expected to occur and feed this information to the two simulation scenarios.

				Previsão semanal
MUNICIPIO	TECNOLOGIA	TIPO DE INTERVENÇÃO	Nº OTS MÉDIO	VARIAÇÃO ESPERADA
Abrantes	FIBRA SWAP	I+R	1	10
Faro	CABO	I+R	20,5	50
Albufeira	SATELITE	AS	3	4
Aveiro	CABO	I+R	25,75	-5
Ovar	FIBRA NOS	AS	3,5	4
Ovar	FIBRA DST	I+R	0	3
Ovar	CABO	M	12,5	10
Amadora	CABO	AS	104	40
Cascais	FIBRA NOS	M	0	4
			0	
			0	
			0	
			0	

Figure C.2: WO PREDICTIONS sheet.