Spare Parts Predictive Analytics for Telecommunications Company

By

Alexandros Mamakos
MBA, BSc

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Signature of Author: ____

Department of Supply Chain Management
May 6, 2022

Certified by: ____

Dr. Maria Jesus Saenz Gil De Gomez

Executive Director, Supply Chain Management Program

Capstone Advisor

Accepted by: ____

Prof. Yossi Sheffi

Director, Center for Transportation and Logistics

Elisha Gray II Professor of Engineering Systems

Professor, Civil and Environmental Engineering

Spare Parts Predictive Analytics for Telecommunications Company

by

Alexandros Mamakos

Submitted to the Program in Supply Chain Management

on May 6, 2022 in Partial Fulfillment of the

Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

Spare parts management is the backbone of asset intensive industries such as telecommunications companies, which operate in a highly competitive environment. Network reliability is a strategic goal as it ensures high customer service level and connectivity. Although companies utilize information related to the expected life of assets and plan maintenance activities, unplanned maintenance is still driven ad hoc. This has an impact not only on the company's operations, inventory levels and cost but also on customers' satisfaction. This capstone studies how telecommunications companies can improve the prediction of site failures and introduces a proactive maintenance approach. Based on our sponsor's pilot project, we apply the MIT's digital supply chain framework to define the value proposition and use the last 3 years of data to develop predictive models for site failures. To approach this case, we start by using the k-means algorithm and cluster the sites in three groups based on variability and demand for spare parts. To predict site failures, we apply time series models (exponential smoothing, Holt Winters and ARIMA) and assess the forecast accuracy based on RMSE and MAPE. In the last stage, we use supervised machine learning classification algorithms (Naive Bayes, Decision Tree, and Random Forest) and assess the accuracy using the correlation matrix. Based on our pilot project, we found that, while time series have a high percentage of error, machine learning algorithms can predict assets failures with accuracy between 60% to 85% and drive predictive maintenance and reduction of inventory levels and ageing. Nevertheless, companies should consider high quality and real time data prerequisites for machine learning. Our findings can be useful for other asset intensive companies that currently use traditional maintenance methods and are seeking to improve their predictive capabilities

Capstone Advisor: Dr. Maria Jesus Saenz Gil De Gomez

Title: Executive Director, Supply Chain Management Program

ACKNOWLEDGMENTS

First and foremost, I want to express my gratitude to my supervisor, Dr. Maria Jesus Saenz, for her amazing guidance, continuous support, and patience during this project. Maria's immense knowledge and experience were pivotal, and this research would have been impossible without her aid. I would like also to thank Dr. Ozden Tozanli, Dr. Elenna R Dugundji and Dr. Mehdi Farahani for their valuable feedback, advice, and recommendations throughout this capstone project.

A special thanks to Pamela Siska for her support and constructive feedback.

I would like also to thank all the people from the sponsor company for their support and valuable feedback: Jonathan, Gary, Mbuso, Mihir, Mbuqe, Ashraf and Holger.

A big thank you to all my classmates, the Center for Transportation and Logistics, and the entire MIT community for the amazing experience.

I want to express my gratitude to my father, Panagiotis, for always supporting me and helped throughout my life.

To my wife, Elia. Without her love, patience and support, nothing would have been possible. Thank you for taking care our family and for making me a better person.

Lastly, special thanks to my son, Thanos, for being a good boy and taking care of his mama during my absence.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

1.1 Sponsor Background

The sponsor company is the largest mobile and fixed network operators in Europe and one of the largest telecommunications companies worldwide. The revenues of the company exceeded the 40 billion Euros during Fiscal Year 2020-2021 and the company has more than 300M mobile and 26M fixed connections worldwide. The company operates its own network in most European countries, Africa and Asia, while it has partner agreements with local operators in more than 40 countries. Moreover, the company, through its Global Enterprise Division, offers telecommunication services to corporate clients worldwide.

The primary services of the company are the provision of mobile network services using an extensive network of sites as well as fixed line services such as broadband, television and voice. The company also offers additional services such as IoT applications and Cloud services. As stated in company's purpose, its goal is to provide reliable and cost-effective services to the end customers and the society through a gigabit fixed and mobile network but also to ensure that commercial success does not come at a cost to the environment.

The sponsor company, in order to manage all its different markets and divisions, operates a complex supply chain with business models that differ between markets and/or divisions. Overall, the supply chain is managed globally by a central division which is leading the digital transformation of company's supply chain.

Procurement is strategically managed by the central division, which is responsible for sourcing goods and services for the company through global contracts with major suppliers. However, all markets also maintain local contracts with smaller suppliers (or local subsidiaries of multinational vendors) to manage lower scale purchases or services. Purchase orders are raised locally against a contract (global or local) and vendors deliver the goods to a specific country warehouse.

For warehousing and logistics operations, the company does not have a standard model across all markets and divisions. In most markets, the company has a mixed model using 3PLs and deployment partners (turnkey), while in some cases, for network materials, the company uses exclusively 3PLs or exclusively deployment partners. However, regardless of the business model, all warehouses for Network and Fixed materials use the company's Warehouse Management System (SAP-EWM) to perform good receipt, storage, picking and good issue.

Deployment of Network and Fixed materials (dispatch and installation) is performed by external partners who pick the items from each warehouse, and they are responsible for site installation. The external partners also manage the returns flow, and they are responsible for returning the materials back to each warehouse after

For Spare Parts Management (SPM), the company has a mixed model. Currently, most markets have outsourced SPM to an external vendor, while other markets have recently moved SPM inhouse.

1.2 Motivation and Problem Statement

The sponsor company, by operating in an asset-Intensive industry, and to maintain its leading role and competitive advantage, has made a great investment in building an extensive network and upgrade the existing sites. While the deployment of assets to the sites is critical to achieve its purpose, the maintenance of this network through SPM is even more important. For this reason, the company is revisiting its SPM strategy (outsource vs. inhouse) to identify the optimum model and as part of this project, sponsor aims to review whether and how predictive models for spare parts can improve Maintenance, Repair and Operations (MRO) process and inventory management.

Effective SPM is a key process for the company as it affects the reliable operation of the network sites and has both a direct and indirect impact on performance, customer service, costs, and profitability as well as on the environment. The purpose of SPM is to ensure that the company has the right spare parts and tools to do the right maintenance work at the right time and the right place. However, the current SPM process is mainly manually driven, complex and costly, leading to high inventory levels, high purchasing costs as well as lack of required items and delays in repairs (based on market benchmarks, it is estimated that as much as 50% of unscheduled asset downtime can be attributed to the lack of spare parts(SAP, 2019).

The company has run several projects in the past aimed at reducing the spare parts costs, optimizing inventory levels and improving inventory turns. Although these projects brought significant financial benefits after implementation, the results were temporary. The inventory levels were increasing and overall the inventory allocation between the different storage locations was not aligned with actual demand.

Following an internal assessment, the main inefficiency identified in the current SPM process is the lack of a robust "demand and supply process" for SPM. Currently the company is planning the requirements for spare parts based only historical data by using a six-months moving average approach and a preventive maintenance schedule. In addition, the decision for the storage location of the inventory is not based on actual demand, which leads to additional issues and delays as in many cases the maintenance teams need to receive the required inventory from different storage locations (or send a repair van from another region to fix the issue). Given also that the cost of a potential downtime of a site is extremely high, the organization increases the procurement of SKUs and, unavoidably, this leads to increased inventory levels and costs.

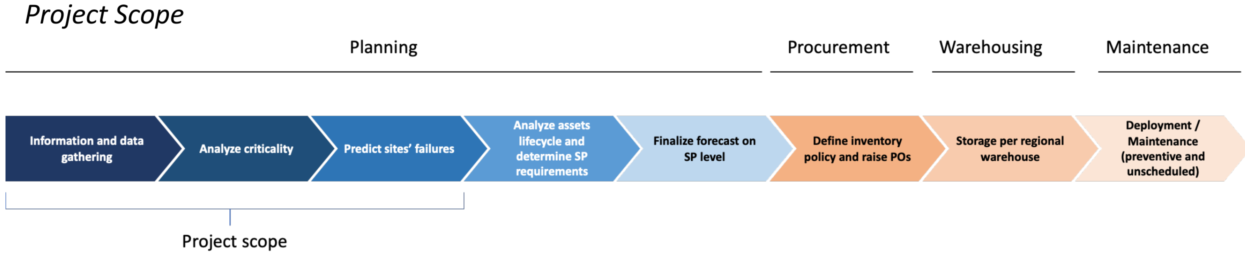
1.3 Project Scope and Research Questions

The current SPM problem affects the company's overall supply chain and has many different aspects, from prediction of site failures, to demand for specific spare parts and inventory levels and storage in different locations, to the overall supply chain strategy for network design, deliveries and service levels.

In this capstone project, given the sponsor's request and the time limitations, we will cover the prediction of site failures in one of the markets where the company operates (**Error! Reference source not found.**). The reason for this decision is that the company currently faces limitations in terms of data on SKU level but also because the purpose of the sponsor is to identify improvement opportunities in the prediction of sites failures and translate them into a benefits driven digital transformation roadmap for SPM process.

Figure 1

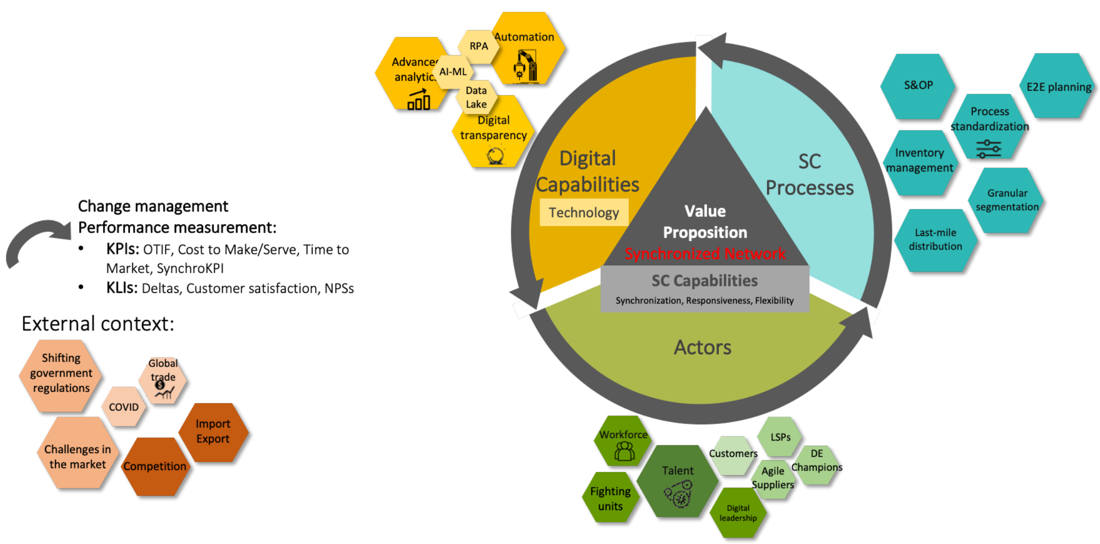
Project Scope



Given the project scope and time limitations, as well as the broader sponsor's goal of a digital transformation roadmap, we needed to define accurately the research questions and agree with the sponsor on the expected outcome. To do this, we used the framework developed by MIT's Digital Supply Chain Transformation Lab (Figure 2) (Saenz et al., 2022).

Figure 2

Digital Supply Chain Transformation framework



Note: Framework - Digital Supply Chain: Interactions between Technology, SC Processes and Organization (Saenz et al., 2022)

Per the framework, we reviewed and defined, along with the sponsor, the value proposition (vision), the supply chain capabilities and the structure levers (processes, digital capabilities) to identify with the accurate research questions.

Starting from the overall vision of the company, which is to fulfil the needs of customers by being a reliable and innovative provider, we translated this vision into a value proposition for reliable operations of network through effective and proactive spare parts management. We continued with the analysis and definition of the key supply chain capabilities to achieve our value proposition and closed our analysis by identifying the processes impacted, the digital capabilities and the performance metrics which we would use to measure the success. The results of our analysis are *summarized* in Table 1.

Table 1

Framework for site failure prediction

Level	Description
Value proposition	To ensure reliable operations of network through effective SPM
SC capabilities	Responsiveness, Transparency, Visibility
SC processes	Demand planning and forecasting Inventory management
Digital capabilities	Machine Learning for classification (Naive Bayes, Decision Tree, Random Forest) Smart Inventory control
Performance Metrics	Forecast accuracy Inventory metrics Net promoter score

For the sponsor company, to align SPM with its vision and provide reliable network to meet customers' needs, means that it should ensure that maintenance operations are driven by responsiveness in the signals and information that the company can access. Moreover, the company also needs to have visibility and transparency in the SMP operations.

To achieve this, the company decided to review the demand planning and foresting process for site failures. To identify areas of improvement which will enable higher level of responsiveness and visibility, the company wants also to investigate how digital capabilities such as machine learning and data analytics can improve forecast accuracy and provide the opportunity for proactive maintenance through predictions of site failures.

The company's goal on operating level is to improve the forecast accuracy of site failures which will then give the opportunity to optimize the inventory and reduce the ageing. Overall, the success of the digital transformation should be reflected on customer (internal and external) satisfaction metrics.

Based on the above framework, this capstone project will provide the answer to the following research question: How can the sponsor improve the prediction of site failures and introduce a proactive maintenance approach?

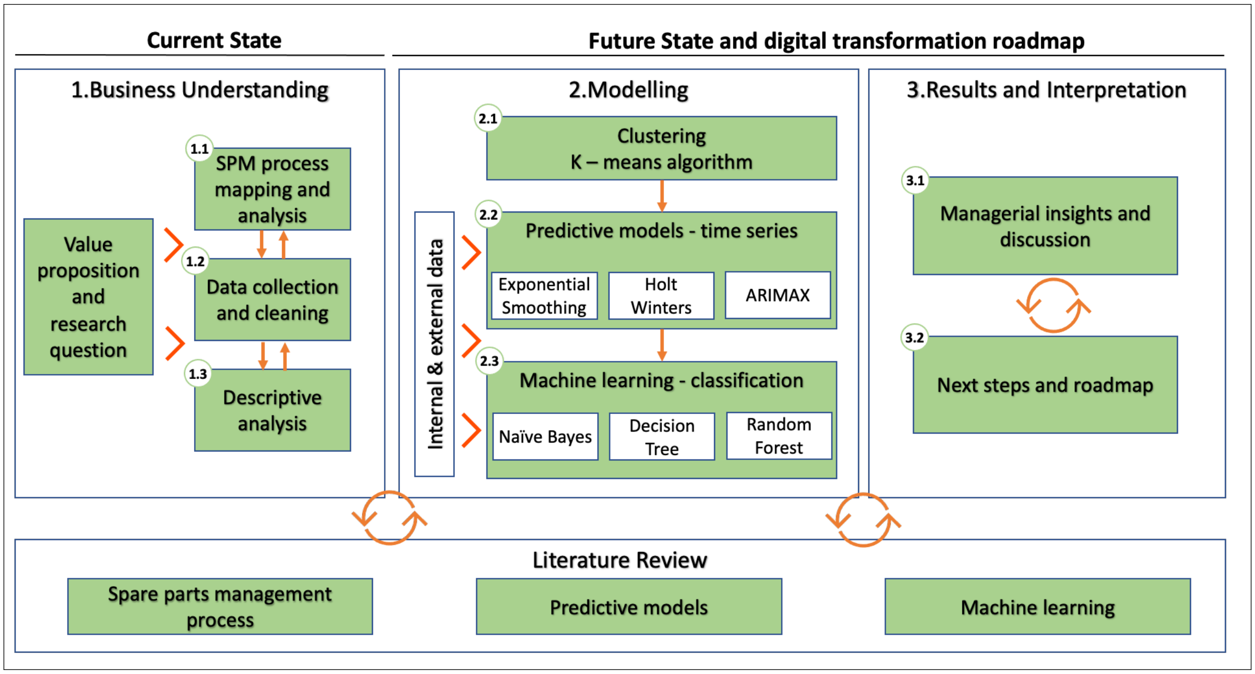
1.4 Approach and Methodology

To answer to the above question, we followed a structured approach involving 3 main areas of research: (1) Map the current state of sponsor's Spare Parts Management process, (2) Apply machine learning and develop predictive models for future demand (3) Summarize the results for the sponsor and propose the next steps required for the digital transformation of SPM.

Overall, our approach was supported by a literature review per project phase.

Figure 3

Project approach and methodology



The purpose of Phase 1, Business Understanding, was to map the current state and understand how the company manages spare parts. To achieve this end, we performed interviews with the key stakeholders of the sponsor using questionnaires to collect information about the current business model, processes, organizational structure, metrics and technology used in SPM. During Phase 1 we also removed the outliers and the cases not covered by the current scope. Finally, we performed the first round of our analysis by applying descriptive analytics to get a better understanding of current state and patterns of demand (in terms of defective network sites).

In Phase 2, Modeling, we focused on quantitative analysis and applied machine learning and predictive analytics to assess how the forecast for site failures can be improved. Phase 2 started by applying to our dataset the k-means algorithm to cluster the network sites in meaningful groups. Following the clustering of data, we developed different predictive models for each cluster and compared their accuracy by calculating the Root Mean Square Error (RMSE) and the Mean Absolute Percentage Error (MAPE). During the modeling phase, apart from the company's data coming from the ERP and WMS, we obtained information from additional resources(Statistics South Africa, n.d.) to enhanced our model.

The final step of the methodology was to translate the research findings into a business insight for the company, as well as to formulate recommendations for the next steps that the company should follow.

2 Literature Review

Maintenance, Repair and Operations (MRO) process is the backbone of asset intensive industries such as telecommunications as it ensures the reliability of their operations. The effective spare parts management (SPM) is one of the key success factors of MRO process as, based on recent studies, "as much as 50% of unscheduled asset downtime can be attributed to the lack of spare parts" (SAP, 2019,p3).

To address the challenges of Spare Parts Management and to optimize the process, many researches have been conducted the last years focusing on the definition of replacements and spare parts inventory planning models using reliability distributions analysis (Venkatesan, 1984; Aronis et al., 2004; Vaughan, 2005; Armstrong & Atkins, 1996).There are also papers and case studies which are approaching the topic from the process improvements angle, either through lean manufacturing practices, using Value Stream Mapping to identify the waste in the process (Mostafa et al., 2015; Pérez-Pucheta et al., 2019) or by trying to link demand and supply aspects through the categorization of the parts to improve inventory management (Jouni et al., 2011). The most recent studies and articles, are focusing either on how technologies and applications can use real time data and analytics to improve the predictive models (Elwany & Gebraeel, 2008; Gebraeel, 2006) or they provide a more theoretical aspect by presenting a prescriptive strategy for spare parts management through IoT (SAP, 2021). As the sponsor wants to assess how to improve SPM process and transform the business model for predictive maintenance and repair, the literature review will cover the areas of predictive models. As the inventory policies of the company are centrally defined taking into consideration multiple factors, this area will not be reviewed.

The purpose of this report is to define how the process for sponsor's Spare Parts Management can be improved by using predictive analytics and real time data through machine learning applications. To achieve this, we will have to assess the current state and how the process is currently structured, to review different approaches for demand forecasting techniques and finally to explore how machine learning can further improve the process. We will focus on different papers for alternative models and how we can predict the future demand. In the last section, we will use available studies and articles to identify the benefits of real time data and how these could enhance our predictive model. Literature review will close with the summary / conclusion of our review and the integration of the areas that we have analysed.

2.1 Predictive models for Spare Parts Replacement

One of the main challenges that our sponsor should manage is the prediction of the spare parts demand. The demand of spare parts mainly depends in two different factors, the unexpected failure of the assets and the degradation. The existing literature covers extensively both factors and many researchers developed predictive models for spare parts replacement. In most of the cases, the models are dealing separately with replacement and inventory models whereas in some cases they approach the topic from a joint optimization angle (Armstrong & Atkins, 1996). Given that the inventory policies are defined centrally by our sponsor and the purpose of the company is the eliminate the unexpected failures, only the literature for replacement models due degradation will be reviewed for this capstone project.

The degradation factor was initially reviewed by Venkatesan who modelled, "a single product single-equipment production-inventory system with infinite storage and production capacities which is reviewed periodically and production and replacement decisions are made based on the inventory level in storage and the level of deterioration of the equipment" (Venkatesan, 1984). The deterioration process was modelled as a finite state Markov process. Delia Montoro-Cazorla and Rafael Pérez-Ocón also developed a Markov model to define the replacements times and cost in a degrading system where there are different types of failures (Montoro-Cazorla & Pérez-Ocón, 2006). The degradation problem was also studied by Akturk and Gurel using a different approach where they linked the operating conditions with preventive maintenance (Selim & Gurel, 2007). In the paper, the authors developed a preventive maintenance index for the condition of the asset and maintenance needs. Their model is also taking into consideration the preventive maintenance cost to define the maintenance strategy and decisions.

Although the above studies are approaching the spare parts replacement topic from the degradation point of view, they have strong correlation with the inventory levels. For our analysis, the Akturk and Gurel models seems that most relevant as it uses Condition Based

Maintenance and proposed approaches and policies for unknown conditions which reflects the environment of network sites around a country (unknown conditions can impact the asset). To enhance our approach, we will proceed with the development of a ARIMAX model and evaluate different scenarios to assess the correlation between different independent variables which affect the degradation of assets.

2.2 Real time data in Spare Parts Management process

Given that applications for Supply Chain and Spare Parts Management have been developed the last years, the available literature does not cover extensively the topic. However, there are several researches and articles which highlight the benefits and improvements that IoT can bring to SPM.

As Gary Forger (2018) highlights in his article, improvement to SPM process can be achieved through preventive maintenance but with sensors, big data and advanced analytics, the organizations will be able to apply predictive maintenance which in reality is real-time maintenance. “The basis of predictive maintenance is collecting data (sensors) about the operating condition of individual components and pieces of equipment. That data dump (big data) is then analysed (data analytics) to determine in real time if action is required. Over time, machine learning (artificial intelligence) can not only anticipate equipment condition but help determine who should be working on each maintenance work order based on skills and wages of technicians, optimizing all aspects of maintenance” (Gary Forger, 2018). The benefits coming from predictive maintenance include increased efficiency, reduction of unplanned downtimes and overheads by eliminating the unnecessary maintenance as well as reduction of operating costs more than 3%.

SAP (2019) also presents in a recent study how the sensors can act as the ‘voice of the asset” providing valuable information for maintenance and replacement This approach treats every asset as a “customer” that communicates specific and ever-changing requirements or maintenance, and spare parts needs and links the real demand with supply. The benefits delivered from sensors are significant for the companies as they can achieve reduced meantime to repair, increased mean time between failures, reduced inventory levels and lower unplanned outages up to 18% versus the organizations who are using reactive maintenance.

The improvements of IoT (sensors) to spare part management and maintenance have also been modelled by Gebraeel (2006) in his study for sensor driven degradation model. In his paper, Gebraeel developed a stochastic degradation model and evaluated the accuracy of failure predictions where was shown that the average prediction error was around 8% compared to an average of 22% for conventional degradation models without the sensor-driven updating methodology.

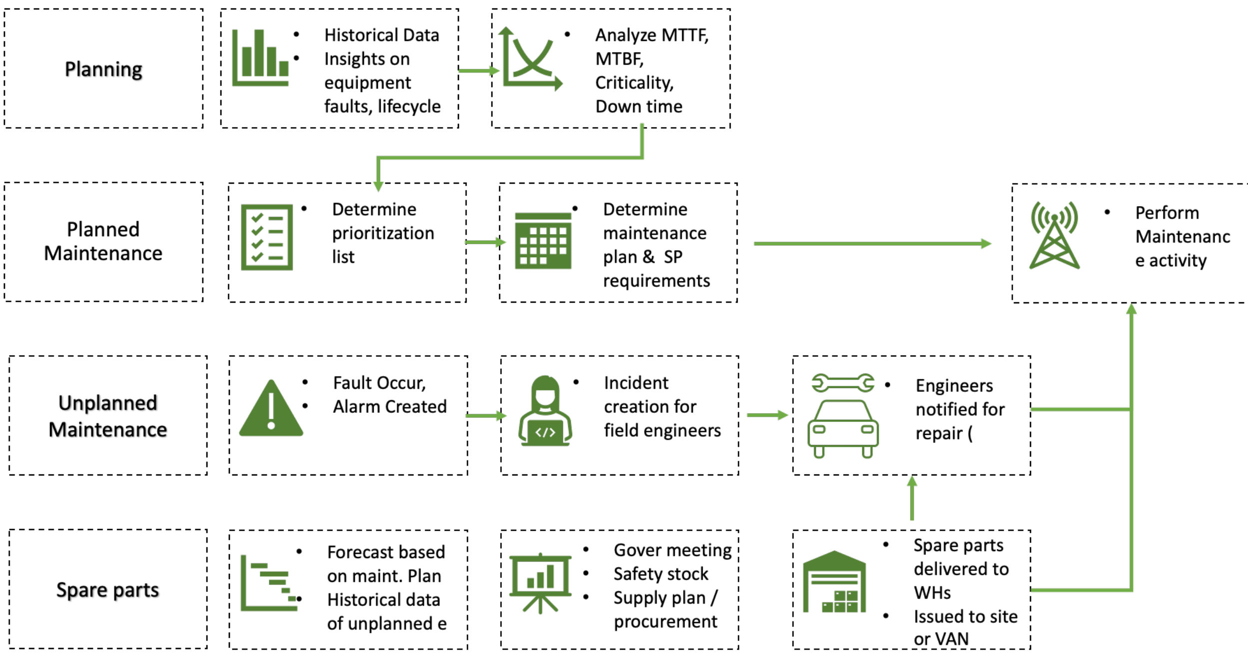
Given the sponsor’s purpose to investigate ways to improve the Spare Parts Management process, the application of real time will provide significant benefits to the company as it will be able to develop a predictive business model for Spare Parts using real time data.

3 Current Spare Parts Management Process Mapping

The two main key areas of the sponsor's SPM process are the demand forecast and the supply planning. The overall purpose of the process is to ensure that the correct SKUs will be available to support planned and unplanned maintenance activities.

Figure 4

SPM process



Demand planning is performed by the control tower and the two key elements which feed the planning process are the maintenance plan on the site level (planned maintenance) and the historical data for spare parts demand on the SKU level. The annual maintenance plan is developed by the technology team and is mainly driven by the criticality of each site (based on the users), the mean time to failure as defined by the vendor, and the mean time between failures. Based on this information, the technology team finalizes the maintenance plan, splits it per operating region and hands over the plan to the control tower to proceed with the demand planning on the SKU level.

The purpose of the SKU demand plan for spare parts is to ensure that the required materials will be available to cover the requirements of the planned maintenance activities but also unplanned incidents. The SKU demand plan is also performed on operating region level (storage locations) and driven by historical data. The methodology used by control tower team is a six-months moving average spare parts demand model (run rate) to predict the requirements. In particular, the team extracts the purchase orders raised during the last six months with the quantities per SKU and uses spreadsheet models with built SQL queries to define the average quantity per SKU.

Overall, the process is manually driven as the planners must extract information from the ERP tables, collect input from technology team and then perform the required analysis on spreadsheets.

Following the definition of the demand plan, the control tower team performs the supply planning along with the procurement team. The control tower team starts by classifying the available inventory by age, the pipeline stock, and the already raised purchase orders and creates the inventory position for the company. The next step for the team is to compare the future demand with the inventory position and the required safety stock (between 5% - 15% per SKU depending on the material category, vendor and area) to define the supply requirements. The supply requirements are linked with the delivery lead times per vendor and then the control tower team finalizes the supply plan with the purchase orders which need to be raised on a weekly basis.

The overall process and the status of demand and supply are reviewed in a biweekly spare part governance and planning meeting. The purpose of the meeting is to refine demand requirements, assess safety stock and inventory of spares, review procurement process, lead times and perform corrective actions.

The supply planning process is also manually driven. The required information is extracted from the ERP and the supplier management platform and then the team uses, as in demand planning, spreadsheets with SQL queries to perform the supply planning. The overall effectiveness of the process is assessed by the inventory KPIs and stock availability metrics for the planned maintenance projects. In particular, the team measures three metrics: the inventory level, the inventory ageing, and the allocated stock per maintenance project.

Good receipt of the orders is performed based on POs. In most cases, good receipt is performed at the central warehouse and then the company distributes the spare parts to the different regional warehouses where they are stored. However, there are cases where the orders are delivered directly to the regional warehouses to cover local requirements for spares.

The maintenance activities are performed by two different contractors who have regional responsibility. The contractors pick the required spare parts from the regional warehouses and store them in their storage locations, which can be either physical storage locations or vehicles (vans). For the planned maintenance activities, the company raises work orders for the contractors based on the maintenance schedule. The contractors visit the sites to perform the maintenance activity, complete the work orders with correct material, and return the defective spares to the warehouse. In case of network failures, the company is informed by an alarm system, creates incident note for an unplanned maintenance activity and assigns it to a contractor. The contractor assesses the incident and determines the spare part requirement. If the SKU is available on the vehicle, he visits the site the performs the maintenance. If there is no available stock, the contactor requests the material from warehouse.

Overall, the process covers end to end the planning and maintenance requirements, but it is manually driven without automation mainly on the prediction side. Demand and supply planning is based on internal data coming from fragmented systems and does not take advantage of all the available internal (e.g. traffic per site) and external sources (e.g. weather forecasts). Finally, the effectiveness of the SPM process is measured by the stock availability per project, leading to high stock levels or misalignment in the SKUs requirements per operating region.

4 Data and Descriptive Analysis

As a first step and in order to assess if and how advanced analytics and machine learning can improve the forecast accuracy for spare parts, we had to developed different forecasting models using time series approach and compare the accuracy versus the existing model. To do this, we had to collect different data sets either from internal sources or external. Following the data collection, we analysed and cleaned the initial data set by removing the outliers and the "out of scope" entries. Finally, we mapped the current state by performing descriptive analytics before moving to the development of the models.

4.1 Data Collection

The first step for the analysis was the definition and collection of the required data. The sponsor provided different data sets covering the spare parts and maintenance process. The time horizon of the data sets is covering movements of spare parts from 2019 till the end of 2021. The data sets provided by the sponsor were:

Material Master for Spare Parts: This file is the central record of spares parts for the organization and contains all the required information per material code. In particular, the key features of the file that I also used in my analysis are presented in Table 2.

Table 2

Material master attributes

Attribute	Description
Material Number	The unique identifier of each material / code
supplier_vendor_id	The unique identifier for the vendor of each material / code of the vendor
en_description	Description of each material
base_unit_of_measure	Units of measure per material. We have 3 categories:

	EA: Each M: meters KM: Kilometers
Item price	The value / price of each material in Euros (€)
Material Group	The category of each material (70 different material groups)

The material master file contains additional attributes which were not used in our analysis and for this reason we excluded them from the above table (eg. Weight, purchasing group and others)

Sites Master (List of Sites): The site master is the central records of all the sites of the sponsor and contains information for each site / location. The key features of this file are:

Table 3

Sites master attributes

Attribute	**Description**
erp_location_key	The unique identifier of each site in the company's global ERP system.
Service_Level	Categorization of sites based on the required service level. Sponsor has 4 categories which are defined by the importance of the sites (Platinum, Gold, Silver and Bronze)
Country Code	The code of the country
Province Code	The code of the province
Operating Region Code	The code of each operating region which is defined by the logistics set up of the Sponsor (9 operating regions)
City	The city where the site is located
Subplace	The area in the city where the site is located

Longitude	The longitude of the site
Latitude	The latitude of the site

Spare Parts Run Rate: The file contains the existing forecasting of the sponsor and it consists of 2 attributes, the material code and the forecasted volume per material

Spare Parts MRNs (Materials Requisition Notes): This is the key file as it contains all the movements of spare parts from the storage locations to each site on a specific date. The coverage period is 3 years (2019 – 2021) and the key attributes of the file are:

Table 4

Spare parts MRNs

Attribute	Description
Document Number	Unique code per MRN issued a specific date for a specific site
Service Level	Categorization of sites based on the required service level. Sponsor has 4 categories which are defined by the importance of the sites (Platinum, Gold, Silver and Bronze)
Country Code	The code of the country
Province Code	The code of the province
Operating Region Code	The code of each operating region which is defined by the logistics set up of the Sponsor (9 operating regions)
City	The city where the site is located
Subplace	The area in the city where the site is located
Longitude	The longitude of the site
Latitude	The latitude of the site

Traffic Per Site: the file has the average *number* of users per site location. The two key attributes are the site id and the average number of users.

4.2 Data Cleansing

Our key data set for the analysis was the “Spare Parts MRNs (Materials Requisition Notes)”. The initial file contained 10,265 entries of spare parts movements from the storage locations to the destinations (sites). Primary key of the file was the document number. Given that we normally have many different materials in the same MRN (as multiple materials were sent at the same date to the same site), we developed a new Primary Key which is the combination of the document number and the material number.

Starting with our analysis, we tried to produce our initial descriptive analytics and noticed that our data set had several errors or missing values which should be fixed or removed. In particular:

- Entries with missing values: 2586 entries with missing values (cost per material). In order to find the missing information and update the file, we asked from the sponsor company to share with us the current catalogues and prices and managed to find the missing data for 1000 entries. For the remaining entries, we sent the file to the sponsor and confirmed that these entries should be excluded from the analysis since:
    - They cover services (not actual materials) which were added by mistake in the data set
    - The materials were old and were decommissioned (pricing info were not available)

- Movements to dummy sites. We identified 461 entries with extremely high values. Following a meeting with the sponsor company, we discovered that these were not actual movements to the sites. Sponsor company has created in the ERP storage locations (physical or VANs) for regional partners, configured as dummy sites. The locations maintain inventory and, as per sponsor's recommendation, these entries were excluded from the analysis
- Following the exclusion of all above cases, we performed an outliers test using the Grubbs' method. More specifically, we identified outliers by calculating the difference between the value and the mean, and then dividing that difference by the standard deviation. For the cases where that ratio was too large, the value defined as an outlier. Again, these cases where shared with the company and it was explained that, for low-cost materials, the users record in the ERP bulk movements to a specific site.

The final usable data set contained 6725 entries which represented movements of Spare Parts to locations (sites) between 2019 – 2021

4.3 Descriptive Statistics

4.3.1 Spare Parts demand on operating region level

As already mentioned in chapter 1.1 "Sponsor's Background", Supply Chain and Network Operations are organized in 9 operating regions and storage locations. From our initial analysis we found that between 2019 and 2021, the company deployed 26790 spare parts in 2058 sites in total. The average demand for spares per site is 13.02 materials but we have noticed that we have a very high standard deviation (3 times higher). We also got similar outcome when we segmented the data by operating region (Table 5).

Table 5

Descriptive statistics per operating region

Ops Regions	Num of Sites	Total Q	Avg Q	StdDev Q
Op Region 1	162	1323	8.17	24.03
Op Region 2	298	8308	27.88	54.21
Op Region 3	253	2528	9.99	23.98
Op Region 4	252	1503	5.96	7.55
Op Region 5	315	3496	11.10	33.03
Op Region 6	286	2840	9.93	36.33
Op Region 7	217	2582	11.90	41.88
Op Region 8	101	550	5.45	9.39
Op Region 9	174	3660	21.03	85.24
Grand Total	**2058**	**26790**	**13.02**	**41.80**

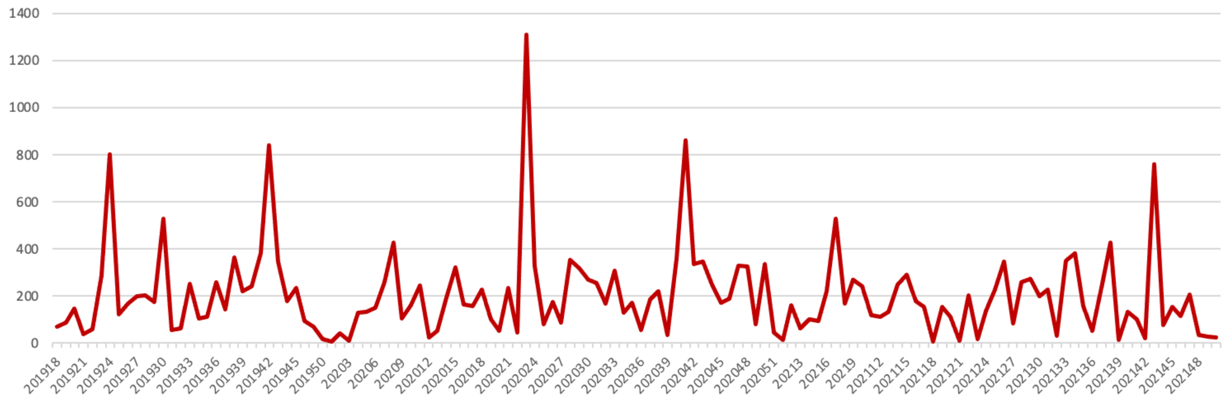
The above outcome can be explained by the fact that although the spare parts demand in most sites is low, we have cases where the demand was extremely high (fast moving) mainly due to vandalisms or extended issues. Taking into consideration this, we concluded that we would have to cluster our data before developing our forecasting model.

Further to the initial descriptive analysis, we also reviewed the seasonality (total quantity deployed per week) and we have found that the average weekly demand is 203 materials with a Standard Deviation of 186 materials. When we plot the total demand on a weekly level, we noticed that we have demand peaks in week 42 (October). We also saw that during 2019 and 2020 we had peaks in week 23 (June) but this pattern was not repeated in 2021 and we will have to review during our analysis if there is any specific factor which affects the seasonality patterns.

The above findings also confirmed by the sponsor as the months with the highest demand for spare parts.

Figure 5

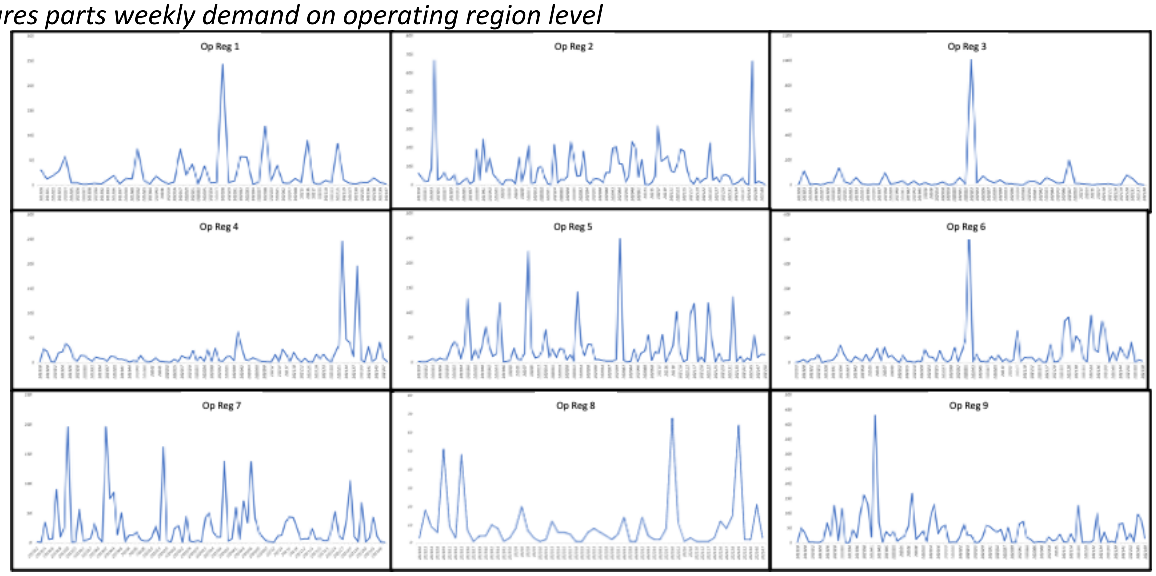
Spare parts weekly demand on country level



However, when we reviewed the seasonality on operating region level, we noticed that each region has different seasonality patterns. As it is shown in the below charts, there are regions with relatively stable demand and one high peak (ie. Op Reg3 and Op Reg5) while we have other areas where the demand fluctuates through the review period (ie Op Reg6 and 7).

Figure 6

Spares parts weekly demand on operating region level



Given the above patterns, we concluded that demand is affected by regional factors so we decided to aggregate our data on Operational regions level and model the spares parts demand as this will affect the compamy's decision about the storage location where the sponsor company should locate spare parts Inventory.

4.3.2 Sites demand for spare parts due to operating issues

The next step was to investigate how many sites have issues and require maintenance and spare parts (demand on site level). As we explained in the beginning, the key objective of the sponsor company is to ensure that the sites are running properly and provide network coverage to the end customers. From our analysis, we saw that between 2019 and 2021 we had to deploy spare parts to 2926 sites which represent the 14% of all active sites (some sites had demand for spares more than once) and most cases were in Op Reg4, Op Reg2, Op Reg5 and Op Reg6.

Table 6

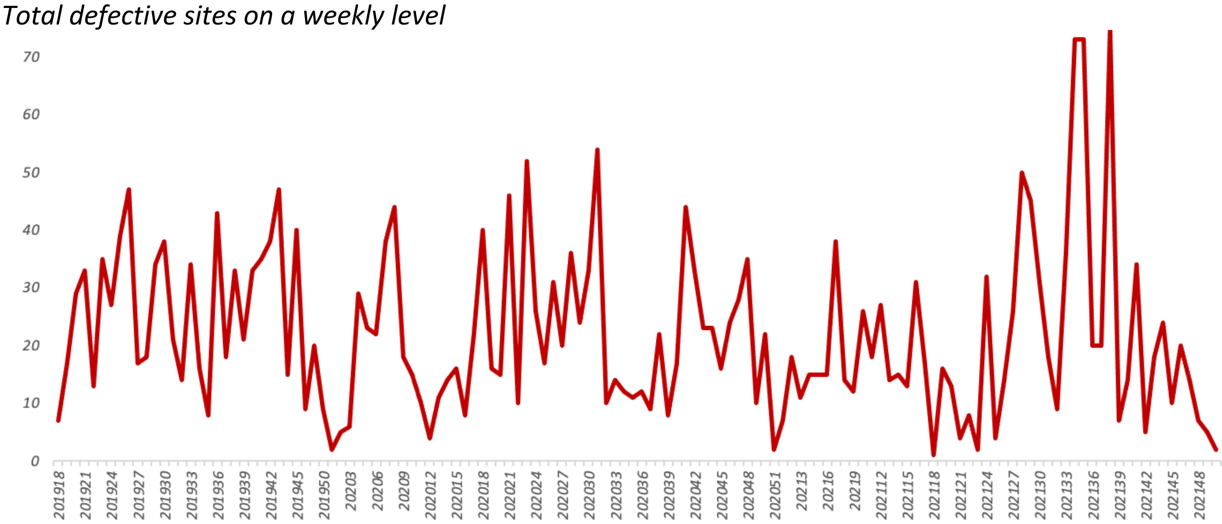
Defective sites per region

Ops Region	Defective sites	Active Sites	Percentage of defective sites
Op Region 1	201	1732	12%
Op Region 2	456	1894	24%
Op Region 3	290	2930	10%
Op Region 4	479	1821	26%
Op Region 5	430	1633	26%
Op Region 6	397	2729	15%
Op Region 7	305	2952	10%
Op Region 8	120	2297	5%
Op Region 9	248	2293	11%
Grand Total	**2926**	**20281**	**14%**

On a seasonal level (aggregated on a weekly level), we have a weekly average demand of 22.16 sites, but the standard deviation is high (14.7 sites per week), meaning that we have a high variability of demand on site level (defective sites)

Figure 7

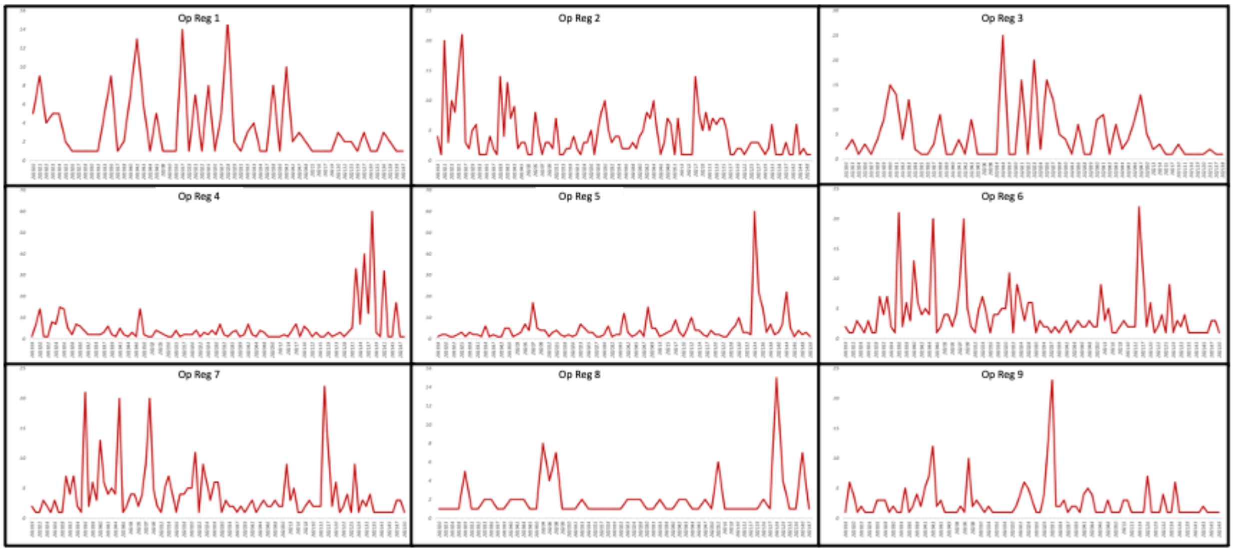
Total defective sites on a weekly level



Following the seasonality analysis on country level, we also analysed and plotted the seasonality on operating region level. Similarly with the country analysis, we have some peaks in the demand but there is no clear seasonality pattern in any of the 9 regions.

Figure 8

Weekly demand - defective sites per operating region



Following the above analysis and outcome, we decided to maintain the regional segmentation due to the current structure of Logistics Operations and Network (Warehouses maintain inventory and service operating regions) but before proceeding to the development of our model we will attempt to develop cluster of sites using additional feature.

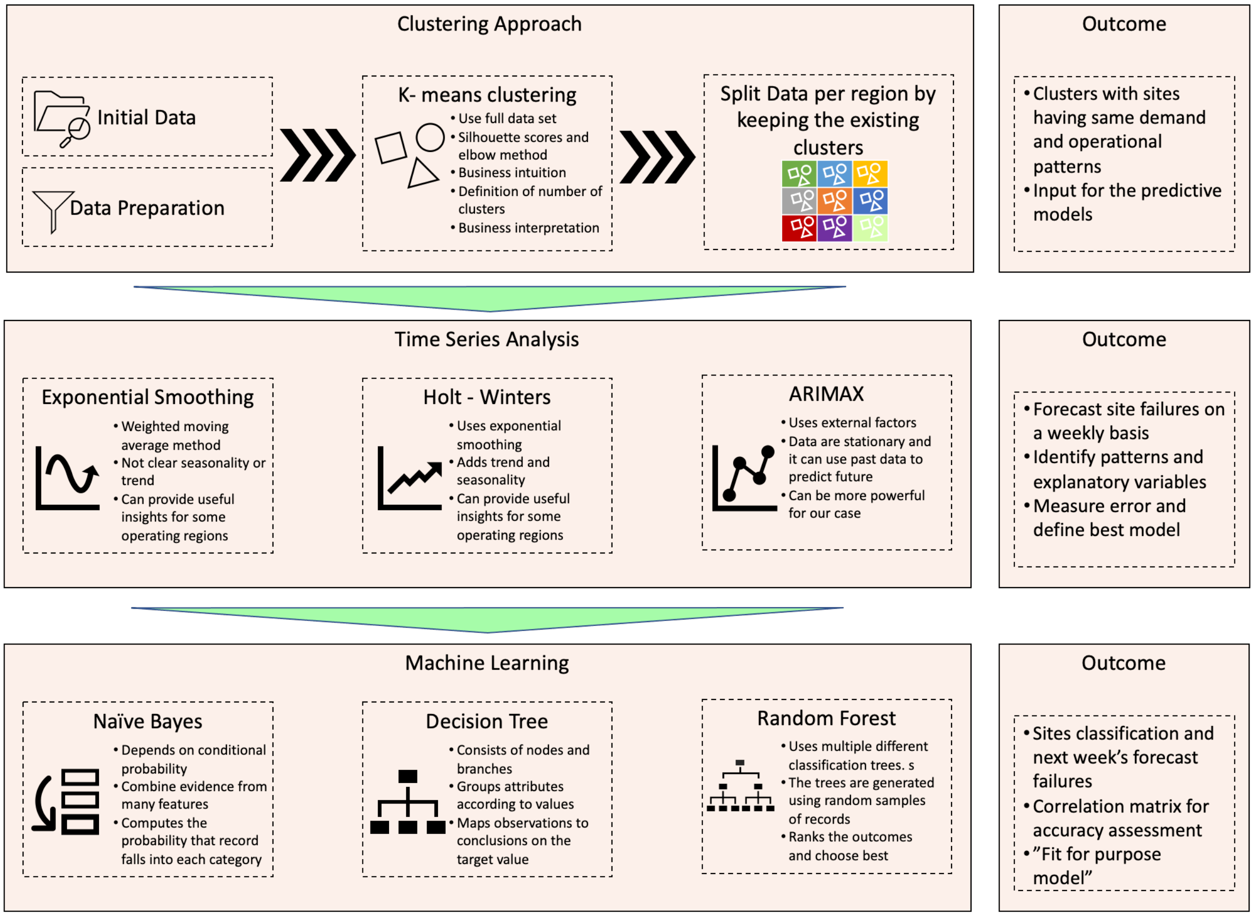
5 Modeling and Results

To support our sponsor company in developing a model for spare parts management demand, we used time series (section 5.2Predictive models using time series analysis) and machine learning models (section 5.3Machine learning algorithms) to determine the model with the most accurate outcome.

Before starting the predictive analysis, we performed cluster analysis (section 5.1) to group network sites with similar characteristics and demand patterns. Although we already performed an initial segmentation on operating region level, we decided to perform a further segmentation as network sites have different characteristics in terms of demand within the same regions.

Figure 9

Modeling approach



5.1 Cluster Analysis

To segment our data, we performed cluster analysis using a k-means algorithm. Given that the business objective of the sponsor company was to predict failures on site level, we proceed in our clustering by using two key features which would group our target variable by demand characteristics:

a. Number of material codes replaced in sites with defects: This feature will provide us with information on the state of the defective sites. Sites with a high number of material codes are likely to have extensive issues, while sites with a low number of material codes probably will have only a specific issue.

b. Variability of demanded spare parts quantity. With this second feature, we aim to segment the sites based on the predictability of demanded quantity for spare sparts. Low variability means that the actual quantity required can be predicted with high accuracy and maintain the required inventory, while high variability means that the company should follow a more proactive approach in maintenance.

The reason for this approach is that the sites in each cluster would have similar demand characteristics, and our expectation was that the predictions would thus be more accurate.

5.1.1 Clustering methodology

The first step is to identify the number of clusters. The number and characteristics of the clusters will drive the predictive models but also help us define the overall business approach for managing spare parts. For this purpose, we run the k-means algorithm for the whole data set and not per operating region subset.

To find the optimal number of clusters, we use the silhouette method to assess the quality of our clusters. Silhouette will compute coefficients of each point and define its similarity to its own cluster compared to other clusters. As shown in Table 7, the optimal number of clusters is 2.

Table 7

Evolution of silhouette scores

Cluster	2	3	4	5

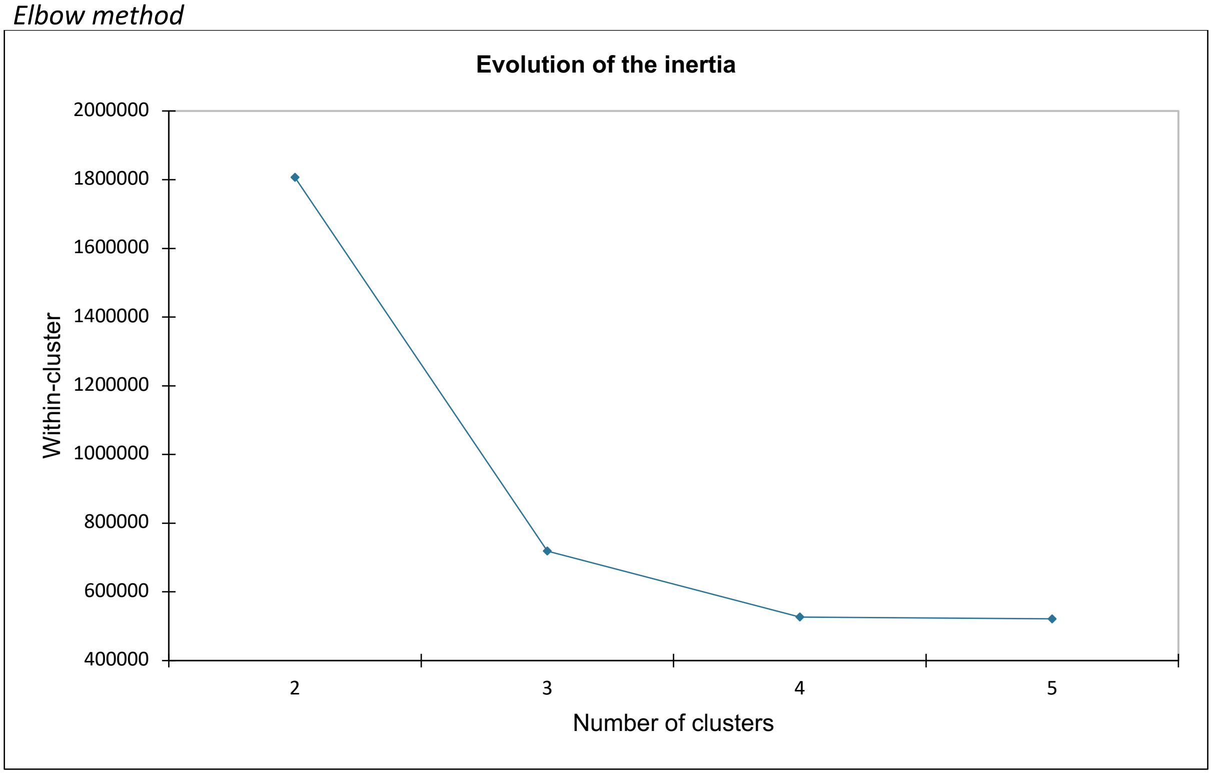
Silhouette scores	0.880	0.695	0.673	0.681

Note: Silhouette scores represent the density of the clusters. Scores near 1 represent very dense and nicely separated clusters. In our case, we see that clustering in 2 groups provides the highest quality.

In contrast, using the elbow method (Figure 10), we find that the optimal number of clusters is 3, meaning that adding an additional cluster will not provide much better results.

Figure 10

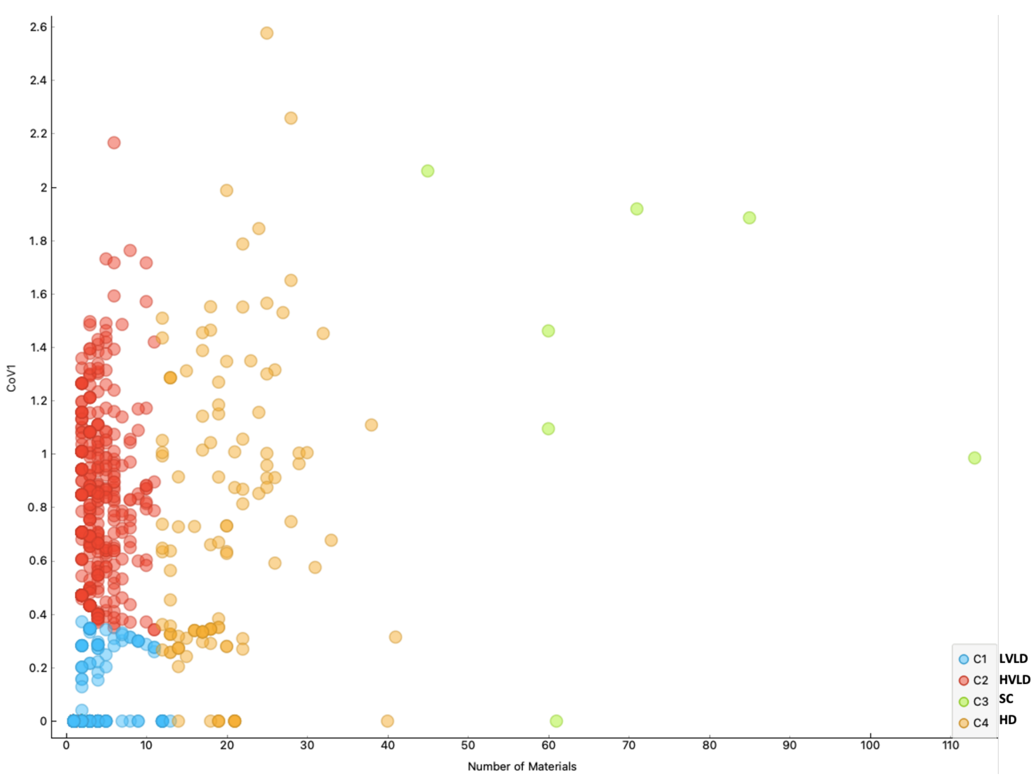
Elbow method



However, after plotting the results and discussing the outcome with the sponsor company, we decided that neither 2 nor 3 clusters would represent accurately the sites' status in terms of failures and demand. The main reason for this decision was that the 2-cluster approach will not give the sponsor a clear segmentation of sites based on the magnitude of damage (extensive and complex issues where multiple different spare parts are required versus more simple issues). The 3-cluster approach seemed more aligned with the business model however, using this approach we would miss a segment for special cases. In particular, with 3 clusters, we would manage the special cases where variability and number of materials are extremely high, as part of the cluster where we mainly see medium to high variability. The outcome of the above business decision, the outcome was to cluster the sites into 4 main categories.

Figure 11

Site Clusters



5.1.2 Business Interpretation of defined clusters

Based on our analysis and as shown in Figure 11, we have 4 main clusters.

- **C1. Low variability and low demand cluster** (LVLD - blue): This cluster contains 1362 network sites (66% of total sites with issue) with variability lower than 0.38, meaning that the prediction accuracy can be high. The demand for unique material codes is also low (less than 12). By deep diving in the data, we see in this cluster network sites where the issues were not extensive as a few material codes had to be replaced due to defects and the required quantity for maintenance was also low.
- **C2. High variability and low demand cluster** (HVLD - red). In this cluster we have 572 network sites, which represent the 28.3% of our data. These sites have low demand in terms of unique material codes, meaning that the defects were not extensive but comparing to C1, the variability of demand is high (larger than 0.38). This means that the prediction of the actual quantity needed cannot be predicted with high accuracy. Based on our initial dataset and following a discussion with the company, in this category we see sites where we had damage either in multiple items of the same material code or cases where the field engineers discovered additional issues when the visited the site to replace a specific material.
- **C3. Special Cases cluster** (SC - green). In this cluster we only have 7 cases (0.3%). The main characteristic of this cluster is that the total demand for different the material codes was extremely high (higher than 50) and from our initial data set we understood that these were cases where the company had to run extensive

maintenance (although data were not available, we assumed that this happened due to the age of the sites or vandalism)

- **C4. High Demand cluster** (HD - orange). The remaining 117 sites fall in this category (5.6%). The main characteristic of this cluster is the high demand in terms of material codes (more than 11 but less than 50). Normally in this category we find sites either with extensive issues due to different reasons (e.g. vandalism) or sites where several materials had to be replaced mainly due to age.

Our cluster analysis of the full data set yielded the above four clusters, which are driven by the variability of demand for spares and the total demand for spare parts SKUs. With this clustering we also developed a framework driven by the number and demand of SKUs which will be used by our sponsor to define the inventory policy of the company.

5.1.3 Clustering per Operating Region

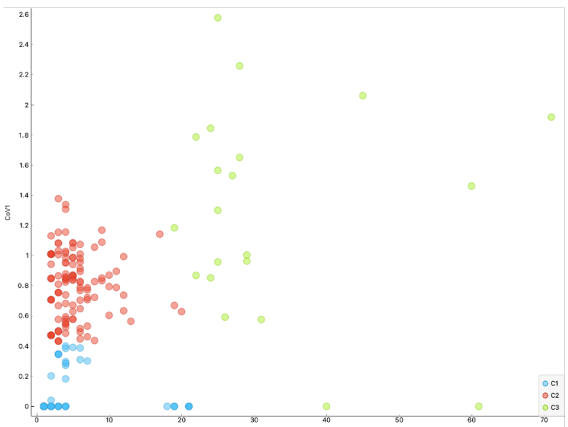
During the descriptive analysis phase, we found that the demand has regional drivers and patterns. Moreover, as already mentioned, the company is organized operationally on a regional level with separate storage locations. Given these findings, the last step before running the predictive models was to perform the clustering analysis on a regional level.

For the clustering per operating region, we followed the same approach and applied the k-means algorithm. However, the different datasets per region provided different centroids (mean) and when we ran the k-means algorithm we received a different clustering per region and in most cases. For instance, as we can see in Figure 12, for operating Region 2 the k-means algorithm gives us different clusters. The variability of demand between C1 and C2 is now 0.41 (comparing to 0.38 that we have defined) but most importantly the number of SKUs that defines how extensive is the damage and the demand increased from 12 to almost 20. Moreover, when we compared this with the proposed clusters in other regions, we found that the results were not constant in terms of business rules (variability higher than 0.38 and number of materials higher than 12). For instance, in region operating region 9 (Figure 12) we see that the variability of demand increased in 0,45 and the SKUs demand changed to 14.

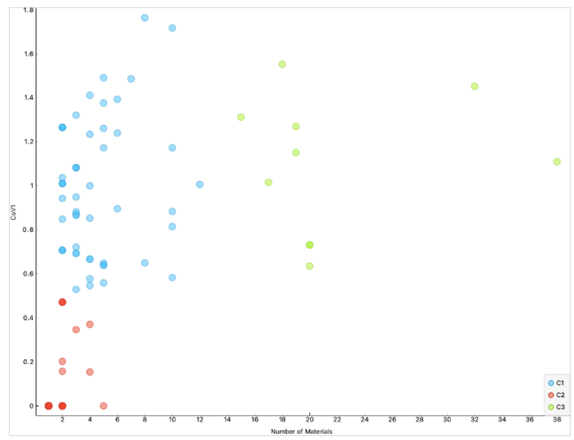
Figure 12

Clusters using k-means for OP Regions 2 and 9

Op Region 2



Op Region 9



The findings for the k-means clustering per operating region led us to the final approach to segmenting our data and proceeding with the forecasting modelling. In particular, we discussed the results with the sponsor company and decided to maintain the k-means clustering from the full data set and then split the network sites in each of the nine operating regions by maintaining the initial clustering.

The reason for this clustering approach was the additional business complexity that a more detailed model would create. The sponsor company sought a model that is not complex and could be adopted by the planners and the supply chain team. A multi cluster approach per operating region would increase complexity significantly and there would be high risk of low adoption.

Based on our chosen approach to clustering, we obtained 4 clusters using a k-means algorithm and then split the data into the 9 operating regions.

Figure 13

Sites clusters per operating region

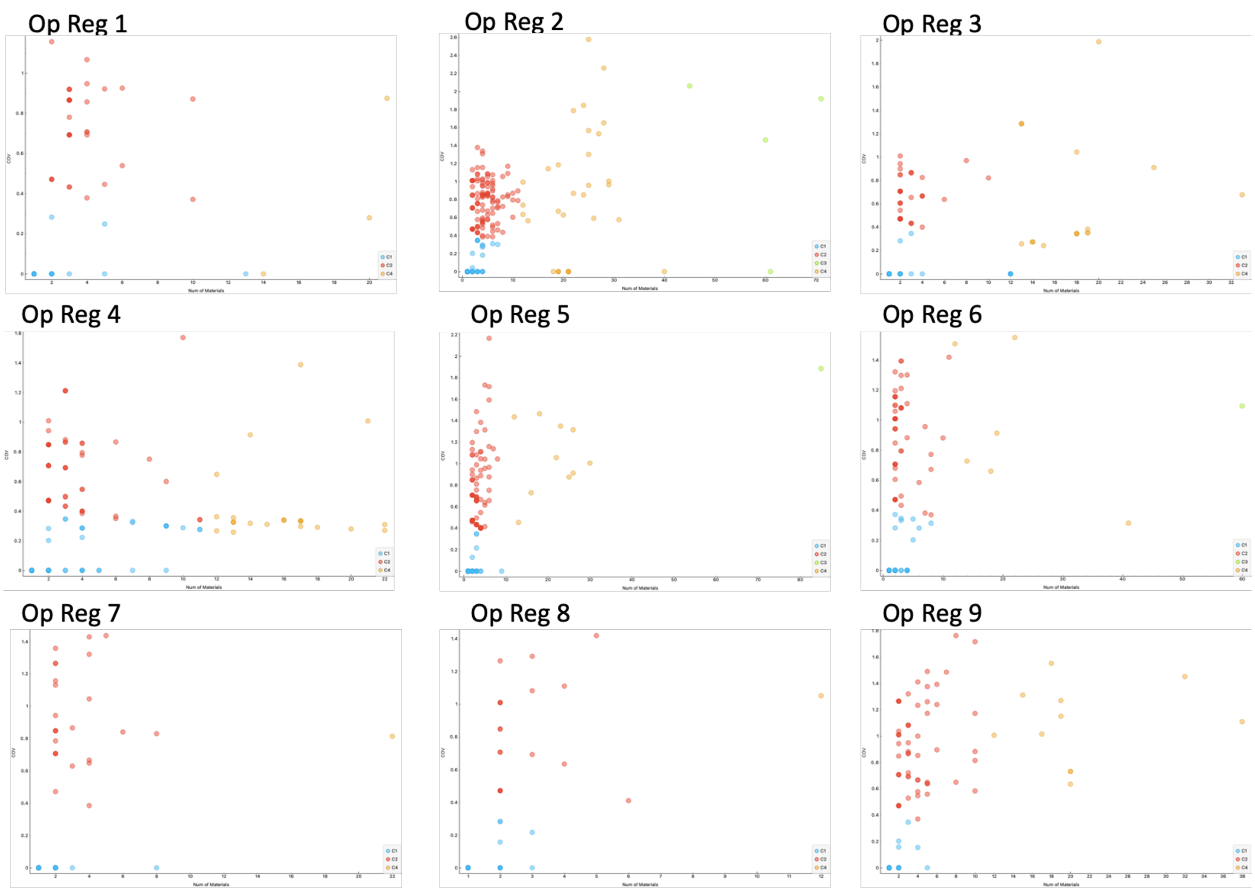


Table 8

Number of sites per cluster and per operating region

Clusters	Op Reg1	Op Reg2	Op Reg3	Op Reg4	Op Reg5	Op Reg6	Op Reg7	Op Reg8	Op Reg9
C1	123	134	172	168	180	220	120	80	96
C2	36	121	58	56	124	59	25	19	57
C3		4			1	1			
C4	3	33	19	26	10	6	1	1	11

Overall, 66% of the defective sites fall into Cluster 1, where we have low variability and low demand of SKUs, which means that the sponsor company had no extensive defects in these sites and could also predict with accuracy the required demand and define the appropriate inventory policy. Cluster 2 is the second largest cluster, with more than 28% of the cases. Almost 6% of the cases fall into Cluster 4, and a small percentage of cases are in Cluster 3 (extreme cases).

Most of the regions follow the same pattern. As we can see in Table 8, in 6 regions (Op Reg1, Op Reg3, Op Reg4, Op Reg6, Op Reg7, Op Reg8), most defective sites fall in cluster 1. Cluster 2 in these regions is also the second largest group but with a lower percentage comparing to the overall Cluster 2 percentage mainly since Cluster 1 has much higher percentage than 66%. Cluster 4 (high demand) is the third largest, with a few cases per region (only in Op Reg4 we see a high number of sites in C4) while Cluster 3 does not contain any sites (only 1 case in Op Reg6).

However, we also have 3 cases where the above findings are not confirmed. In Op Reg9 and Op Reg5 the percentage of Cluster 1 sites are 58.5% and 57.1% respectively and the percentage of C2 is higher than 35%. More interesting is the case of Op Reg2 where most sites fall in Cluster 1 and 2 (46% and 41% respectively) and we also see highest number of sites in Cluster 4 (high demand cluster) and Cluster 3 (extreme cases). In the next section, where we will explain our predictive models, we need to analyse these cases separately, identify the reason for them, and take this reason into consideration during our predictions.

5.2 Predictive models using time series analysis

To forecast the number of defective sites per operating region, we will develop predictive models using time series analysis and compare the results in terms of errors per model. Our analysis will be driven by the clusters that we defined per operating region in section 5.1, and we will use the simple exponential smoothing, the Holt-Winters and the ARIMA method.

Simple exponential smoothing is an extension of the moving average method, and it uses a weighted moving averaging to forecast future demand. The method takes the forecast for the previous period, adjusts it using the forecast error and makes the next forecast period(Chase, 2013). The reason for choosing single exponential smoothing as one of our methods is that it is used for data where we do not have clear seasonality or trend which is the case for some of our cases as we noticed during the descriptive analysis (e.g. Op Reg3, Op Reg4)

Although exponential smoothing might provide us with good predictions in some cases, we have other regions where we noticed some indications of seasonality (e.g. Op Reg2). For this reason we decided also to test the prediction accuracy using Holt Winters method which takes into consideration trend, cycle and seasonality (Chase, 2013).

The last model that we used is the ARIMA (autoregressive integrated moving average) which considers that data are stationary, meaning that there is no clear trend over time in our data set and it can understand past data to predict future data in time series. Given that there are no relations between the different network sites in terms of installation and date of defects, ARIMA method provide very good results for our analysis. Moreover, we will add the explanatory variables that affect the site failures of the sites and try to create a more powerful model which senses signals to predict future demand (often called ARIMAX model).

5.2.1.1 Methods application and forecast performance

Before applying the methods to our dataset, we had to consider the specific characteristics of our project and data set and decided to aggregate the information on a weekly level. The reason for doing this was that defects in sites cannot be modeled as “daily demand” given that there are not defective sites on a daily level. A time series model structured on a daily level would give us very poor prediction due to the smoothing factor and our purpose was to build a robust model where the previous week’s demand could provide useful information for the predictions.

Moreover, for the ARIMA method, as mentioned, we agreed with the sponsor and added explanatory factors which affect the performance of the sites to enhance the accuracy of the predictions (Table 9). The sources of the information were either internal (ERP system) or external (e.g. visual crossing for weather history).

Table 9

Demand explanatory variables

Attribute	Description
Maximum traffic	The maximum traffic per week for a site
Duration from last maintenance	Number of weeks from the last maintenance of the site
Maximum temperature	Maximum temperature recorder in the area where the site is located
Minimum temperature	Minimum temperature recorder in the area where the site is located
Maximum dew	Maximum dew recorder in the area where the site is located
Maximum humidity	Maximum humidity recorder in the area where the site is located

Precipitation coverage	Maximum percentage of precipitation coverage recorder in the area where the site is located
Precipitation	Maximum amount of precipitation in the area where the site is located
Wind speed	Maximum speed of wind in the area where the site is located
Wind gust	Maximum wind gust (km/h) in the area where the site is located
Snow	Maximum amount of snow in the area where the site is located

To avoid multicollinearity of our variables, we also performed a correlation test using the Pearson method which measures the linear correlation between two variables using the covariances and standard deviations of the variables (Figure 14). Per the method, a correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other while a correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. A correlation coefficient of 0 means that the two variables are not correlated.

To avoid multicollinearity, we want to keep variables where the correlation coefficiency is lower than 0.75.

Figure 14

Correlation matrix (Pearson) for the explanatory variables

Variables	Max. of tempmax	Min. of tempmin	Max. of dew	Max. of humidity	Max. of precip	Max. of precipcover	Max. of snow	Max. of windgust	Max. of windspeed	Traffic	From last maitenance
Max. of tempmax	1	0.429	0.426	0.051	0.037	-0.050		-0.025	-0.024	-0.031	0.004
Min. of tempmin	0.429	1	0.813	0.397	0.235	0.164		-0.073	-0.061	0.010	-0.029
Max. of dew	0.426	0.813	1	0.650	0.267	0.193		-0.022	0.054	-0.005	-0.041
Max. of humidity	0.051	0.397	0.650	1	0.339	0.298		-0.051	0.048	-0.002	-0.036
Max. of precip	0.037	0.235	0.267	0.339	1	0.409		-0.074	-0.049	-0.013	0.002
Max. of precipcover2	-0.050	0.164	0.193	0.298	0.409	1		-0.134	-0.167	0.022	0.017
Max. of snow											
Max. of windgust	-0.025	-0.073	-0.022	-0.051	-0.074	-0.134		1	0.717	-0.011	-0.017
Max. of windspeed	-0.024	-0.061	0.054	0.048	-0.049	-0.167		0.717	1	-0.010	-0.010
Traffic	-0.031	0.010	-0.005	-0.002	-0.013	0.022		-0.011	-0.010	1	0.016
From last maitenance	0.004	-0.029	-0.041	-0.036	0.002	0.017		-0.017	-0.010	0.016	1

As we can see in Figure 14, there is a strong correlation between the maximum amount of dew and the minimum temperature recorded. For these reasons and taking into consideration that we already have a temperature related attribute in our analysis, we decided to remove the Min temperature attribute from the models. In addition, we see that there is a correlation very close to 0.75 between windspeed and wind gust. However, as the correlation is not higher than 0.75 and since wind gust can be a significant phenomenon in only one area, we decided to keep both variables.

Following the above adjustments to our dataset, we applied the three time series methods to each cluster of each region and assess their performance. The purpose of the assessment was not only to measure how good the prediction was but also to compare the different statistical models and define which fits best our goal.

To assess the performance, we used the mean absolute percentage error (MAPE). MAPE is a commonly used metric which measures the average absolute percent error for each period. However, with the MAPE is biased toward estimates that are below the actual value and penalizes negative errors. For this reason, we assessed the models using the root mean squared error (RMSE) which measures the error.

The results of our assessment are presented in Figure 15 where we can see that the ARIMA method outperforms the other two methods in terms of RMSE. This was expected as ARIMA(X) takes into consideration additional factors that affect the demand (site failures). However, we have cases where the simple exponential smoothing method is better in terms of MAPE. This mainly happens due to the nature of the demand, as we have many weeks where the demand (defective sites) is 0 and the exponential smoothing method provide a prediction higher than the actual (positive error). As result, exponential smoothing has a much better MAPE than the ARIMA model which is not biased towards positive errors as it calculates the RMSE.

Figure 15

Forecast performance per method, region and cluster

Method	Metric	Op Region 1			Op Region 2			Op Region 3		
		C1	C2	C4	C1	C2	C4	C1	C2	C4
Exponential Smoothing	RMSE	2.91	1.58	0.31	2.45	1.71	1.05	3.05	1.46	1.05
	MAPE	76.53	64.73	95.51	37.26	72.49	63.94	91.72	65.09	75.45
Holt Winters	RMSE	3.10	1.63	0.31	2.77	2.18	1.16	4.48	1.53	2.11
	MAPE	88.59	73.99	95.52	102.26	77.49	90.21	150.06	57.37	185.65
ARIMA	RMSE	1.17	1.11	0.23	1.82	1.25	0.78	1.98	0.93	0.82
	MAPE	72.51	101.90	51.12	95.12	62.81	76.52	110.51	76.82	63.44

Method	Metric	Op Region 4			Op Region 5			Op Region 6		
		C1	C2	C4	C1	C2	C4	C1	C2	C4
Exponential Smoothing	RMSE	2.56	1.42	3.67	2.38	5.15	0.55	3.22	1.55	0.53
	MAPE	63.49	62.26	104.36	42.09	61.92	83.61	57.54	63.76	79.50
Holt Winters	RMSE	2.78	1.42	3.90	2.35	5.42	11.03	4.27	1.68	0.57
	MAPE	89.03	63.68	128.95	58.06	64.50	288.22	111.09	92.82	77.30
ARIMA	RMSE	1.80	0.80	2.56	1.48	4.37	0.36	2.38	0.91	0.34
	MAPE	89.46	62.60	222.56	70.48	124.14	38.53	93.11	64.69	44.38

Method	Metric	Op Region 7			Op Region 8			Op Region 9		
		C1	C2	C4	C1	C2	C4	C1	C2	C4
Exponential Smoothing	RMSE	3.66	0.89	0.15	1.66	0.53	0.09	2.09	1.24	0.49
	MAPE	89.34	75.98	99.45	66.58	76.23	100.00	29.60	72.36	74.04
Holt Winters	RMSE	3.61	0.95	0.18	1.72	0.54	0.09	2.40	1.34	0.56
	MAPE	110.68	69.77	100.00	66.18	62.84	100.00	78.76	69.37	72.57
ARIMA	RMSE	2.59	0.60	0.12	1.11	0.37	0.08	1.62	0.65	0.35
	MAPE	134.51	60.28	65.03	75.97	58.18	84.28	84.29	64.02	45.55

Moreover, we see that in C4 clusters where the number of defective sites was very low, all methods have extremely low RMSE (e.g. SGC – C4, SGS – C4 where we had just one defective site per cluster) but very high MAPE which indicates that they were not able to predict accurately the failure.

Finally, we noticed that due to the low value of our target variable, as the number of defective sites per week is low (between 0 and 3 per cluster whereas in most cases is between 0 and 1), time series do not provide accurate predictions which are aligned with this model. In particular, due to the moving average approach that they follow, we get predictions in decimals, which does not represents the reality.

For the above reasons, we decided that time series is not a fit-for-purpose approach for the sponsor's case, and we decided to move to machine learning techniques and apply classification algorithms.

5.3 Machine learning algorithms

With the available information that we used for our ARIMA model, we will develop machine learning algorithms and train our models to provide an accurate prediction for upcoming failures. Given that we have historical data about the site failures, we will use supervised machine learning to train our model and come up with an accurate prediction. In particular, we will assess the applicability of Naïve Bayes, Decision Tree and Random Forest classification algorithms.

Naïve Bayes depends on conditional probability, and it is efficient due to its ability to combine evidence from many features. The algorithm considers the value of each feature independently, for each record, and computes the probability that a record falls into each category. Then, the probabilities associated with each feature are combined for each class according to determine the most likely category for each new record.

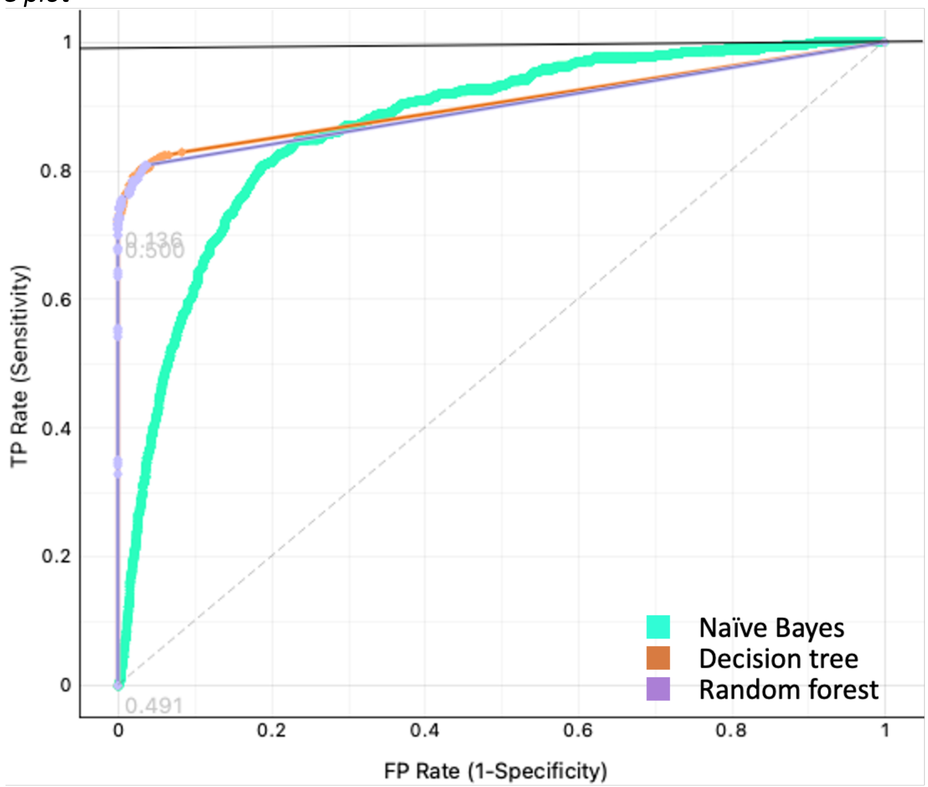
A Decision Tree groups attributes by sorting them according to their values. The Decision Tree maps observations to conclusions on the target value. It consists of nodes and branches. The nodes represent the attribute and the branches the values that our target variable can take.

Random Forest is a classifier that uses multiple different Decision Trees. The trees are generated using random samples of records of the original training set and give a classification for a new object according to attributes. Then the Random Forest collects the outcome of each tree, ranks the outcomes, and choose the predictions based on the occupancy. In general, Random Forest is similar to a Decision Tree but provide better predictions.

Before proceeding with the machine learning algorithms, we will evaluate the three different algorithms using the AUC – ROC metric (Figure 16). With ROC we will plot the true positive case against the false positive rate. The area under the curve (AUC) represents the accurate true positive predictions and the higher the area, the better the model performance.

Figure 16

AUC – ROC plot



As we can see in Figure 16, all three classifications techniques have a quite large area under the curve, which indicates that we should use them. In particular, the area of Naïve Bayes is 0.869, the area of the Random Forest is 0.900 and the area of the Decision Tree is 0.905.

For our data sampling, we will use a fixed proportion approach and split our data set in training and test data. 70% of our data set will be used in the learning process to feed the algorithm and train how to perform predictions. With the remaining 30%, we will test the accuracy of our models.

To assess the accuracy of the classifications' techniques, we will assess the accuracy between actuals and predictions where we have four possible combinations.

Predicted

Actuals	0	1
0	True positive	False negative
1	False positive	True negative

For our case, the most important error to assess is the false positive (Type I) as this error reveals how many of the actual defective sites the model was not able to predict. High percentage of false positive can cause delays in the repair and maintenance process, lack of inventory in the regional storage locations and unavailability of repair teams (VANs).

Equally important is also the Type II error (false negative), as this can cause increased inventory levels for the company on regional and central storage locations. However, for this error type we see that all three classifications perform well due to the fact that the number of true positives is very high, and the models are trained well on that combination. In Figure 17 we summarize all the results for all three classification methods per operating area and per cluster.

(Classifications for all C3 cluster and for Operating Region 8 C4 were not possible due to low number of defective sites)

Figure 17

Classification machine learning performance per region and cluster

Method	Metric	Op Region 1						Op Region 2						Op Region 3					
		C1		C2		C4		C1		C2		C4		C1		C2		C4	
		0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
Naïve Bayes	0	100.0%	0.0%	100.0%	0.0%	86.1%	13.9%	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	99.9%	0.1%	100.0%	0.0%
	1	100.0%	0.0%	100.0%	0.0%	66.7%	33.3%	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	90.0%	10.0%
Classification Tree	0	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%
	1	28.2%	71.8%	100.0%	0.0%	100.0%	0.0%	39.5%	60.5%	44.1%	55.9%	100.0%	0.0%	19.2%	80.8%	100.0%	0.0%	100.0%	0.0%
Random Forest	0	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%	99.9%	0.1%	100.0%	0.0%	100.0%	0.0%	100.0%	0.0%
	1	33.3%	66.7%	11.8%	88.2%	0.0%	100.0%	39.5%	60.5%	33.9%	66.1%	36.8%	63.2%	15.4%	84.6%	33.3%	66.7%	10.0%	90.0%

Method	Metric	Op Region 4						Op Region 5						Op Region 6					
		C1		C2		C4		C1		C2		C4		C1		C2		C4	
		0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
Naïve Bayes	0	100.00%	0.00%	100.00%	0.00%	98.39%	1.61%	100.00%	0.00%	100.00%	0.00%	97.67%	2.33%	100.00%	0.00%	100.00%	0.00%	97.35%	2.65%
	1	100.00%	0.00%	100.00%	0.00%	75.68%	24.32%	100.00%	0.00%	100.00%	0.00%	88.89%	11.11%	100.00%	0.00%	100.00%	0.00%	100.00%	0.00%
Classification Tree	0	100.00%	0.00%	99.95%	0.05%	100.00%	0.00%	100.00%	0.00%	100.00%	0.00%	100.00%	0.00%	99.99%	0.01%	100.00%	0.00%	100.00%	0.00%
	1	31.25%	68.75%	24.24%	75.76%	37.84%	62.16%	37.93%	62.07%	33.33%	66.67%	100.00%	0.00%	23.61%	76.39%	7.41%	92.59%	100.00%	0.00%
Random Forest	0	100.00%	0.00%	100.00%	0.00%	100.00%	0.00%	99.99%	0.01%	99.98%	0.02%	99.22%	0.78%	100.00%	0.00%	100.00%	0.00%	100.00%	0.00%
	1	31.25%	68.75%	18.18%	81.82%	29.73%	70.27%	36.21%	63.79%	26.32%	73.68%	55.56%	44.44%	20.83%	79.17%	3.70%	96.30%	18.18%	81.82%

Method	Metric	Op Region 7						Op Region 8						Op Region 9					
		C1		C2		C4		C1		C2		C4		C1		C2		C4	
		0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
Naïve Bayes	0	99.98%	0.02%	99.80%	0.20%	2.63%	97.37%	99.97%	0.03%	100.00%	0.00%			100.00%	0.00%	100.00%	0.00%	99.29%	0.71%
	1	100.00%	0.00%	100.00%	0.00%	0.00%	100.00%	100.00%	0.00%	100.00%	0.00%			96.67%	3.33%	100.00%	0.00%	90.00%	10.00%
Classification Tree	0	100.00%	0.00%	100.00%	0.00%	100.00%	0.00%	100.00%	0.00%	100.00%	0.00%			100.00%	0.00%	100.00%	0.00%	99.30%	0.70%
	1	30.00%	70.00%	100.00%	0.00%	100.00%	0.00%	100.00%	0.00%	100.00%	0.00%			100.00%	0.00%	100.00%	0.00%	100.00%	0.00%
Random Forest	0	99.98%	0.02%	100.00%	0.00%	100.00%	0.00%	100.00%	0.00%	100.00%	0.00%			99.97%	0.03%	100.00%	0.00%	99.30%	0.70%
	1	20.00%	80.00%	35.71%	64.29%	100.00%	0.00%	32.00%	68.00%	40.00%	60.00%			36.67%	63.33%	50.00%	50.00%	30.00%	70.00%

Note: The number of defective sites for Op Region 8 – C4 was not sufficient to perform classification

Overall, we see that the Random Forest algorithm outperforms the other two and prediction accuracy of the model is between 60% up to 85% (the cases above 90% should be revisited and check if the data sample is sufficient for safe prediction).

While Naïve Bayes is the very accurate in terms of the true positive, it performs extremely poorly on the negative part, as it cannot predict the site failures (only in a few cases was able to predict failures). A Decision Tree has much better accuracy comparing to the Naïve Bayes, but it misses some clusters completely (e.g. Operating Region 7, 8 and 9 – C2) while the Random Forest seems to perform well.

6 Discussion

Maintenance and spare parts are currently managed manually by the sponsor and driven by historical data and 6-month moving average forecast. However, as the results of our analysis revealed, the sponsor company can predict the site failures with high level of accuracy by utilizing different sources of data and introducing machine learning into the demand planning process. Even if this will be the first step of the broader digital transformation roadmap for sponsor's spare parts management process, as we will explain in section 6.1, it can bring significant benefits for the company in many areas such as efficient inventory management, reduction of stock ageing, increased responsiveness in site failures, cost optimization and increased customer satisfaction.

In any case, before applying the model, the sponsor company and readers should take into consideration the project limitations, especially in the data availability area (section 6.2).

6.1 Management Insights

6.1.1 Analysis and result insights

Based our analysis, we concluded that unsupervised and supervised machine learning techniques can significantly improve the prediction of site failures, and subsequently the prediction of spare parts, and in turn inventory management.

Clustering using machine learning: Unsupervised techniques such as k-means can cluster the network sites in a meaningful way to support the value proposition as defined by the sponsor (section 1.3). Instead of more traditional methods such as ABC analysis, the sponsor can use different parameters and cluster the sites in a way that better fits the purpose of each project and strategy. For our analysis, we wanted to segment the sites in a way that would provide insights about the expected demand (prediction of failures) and how the sponsor should manage inventory per case. Following several attempts using traditional techniques like ABC analysis, we managed to achieve this aim using k-means. We identified the optimal number of clusters based on the number of defective materials and the variability of demand.

Predictive analytics for site failures and spare parts management: Machine learning can support the prediction of failures. During our analysis we explored various methods to address the research question and determine whether the sponsor could develop a predictive maintenance approach. Although the simple or more complex time series methodologies have good application in other areas of demand planning, when it comes to the prediction of failures, the results are poor. Nevertheless, the options offered through machine learning techniques provide a great outcome. During our analysis, we explored three different techniques (Naïve Bayes, Decision Tree, and Random Forest). We concluded that the Random Forest provides the most accurate results (between 60% and 85%) as it leverages information from different sources and creates patterns of demand. This can reduce operating cost and, more importantly, reduce inventory levels and ageing. Based on internal research by the company, a more accurate prediction of failures can bring significant results in inventory reduction (up to 15%), ageing reduction (more than 10 days) and overall operating cost through the optimization of the process (up to 10% due to better planning and maintenance).

However, the sponsor should consider that the quality and accuracy of information could provide misleading results and develop a biased model. In our model, we used many different features, but it seems that there is strong correlation between site failures and the traffic, the duration from the last maintenance and extreme weather conditions. The model can be enhanced in the future with additional factors such as human errors during configuration and vandalism.

6.1.2 Project insights and lessons learned

Apart from the results of our analysis, we believe that it is useful to share with our sponsor some best practises that we applied when we were approaching the project.

Define the value proposition. What are the current capabilities and how can digital capabilities support them? It is essential for a successful digital transformation project to define what we want to achieve and how we can do it. For the spare parts topic, we used MIT's Digital Supply Chain framework to approach the topic and define a clear research question.

Frame the problem and define a feasible outcome. Understanding of the current state and processes is extremely important to define the appropriate approach. During our analysis, we conducted many interviews with all the key stakeholders to ensure that we had a good understanding of the current process for maintenance and spare parts demand. Descriptive analytics had a key role in this phase as we were able to understand not only the current way of working but also the limitations in terms of feasible outcome.

Define the strategy and methodology to approach the problem and performance metrics. Resolving a complex problem such as prediction of failures, where data are not available, can be chaotic. To avoid this problem, we defined, just after the descriptive analytics phase, how we should approach the topic and what would be the best possible outcome. In our case, limitations in data did not allow us to deep dive at the SKU level. However, predictions for site failures were feasible and equally important. The sponsor will be able to predict failure and define a proactive maintenance.

Apart from these practices, we should also highlight the main challenge, which should be considered if the sponsor decides to implement and scale up the project.

Machine learning techniques require a high level of integration with internal and external sources to provide accurate outcome. They key advantage of ML is that the algorithm is learning constantly by new sets of data. To maintain a ML algorithm effectively, the sponsor should feed it with real time data from legacy systems and external sources. Data should have the required quality to ensure that the predictions will reflect the reality. In our research, we assumed a static system and used extractions of all required information. Moreover, we cleaned manually all data to get them into the required format. This would not be an actual scenario if the sponsor applies ML: the access to and the quality of data should be ensured.

6.1.3 The road ahead

The current project covered just a small part in terms of the processes impacted and also in terms of geography. As mentioned in the introduction chapter, the sponsor company should review how it can improve end to end the SPM process by applying predictive analytics and machine learning.

Model validation and application: The next step for the company should be to validate the assumptions and link the model of sources or real time data (e.g., sensors in the sites, weather predictions). The outcome will be a much more accurate forecast of site failures, which will give the opportunity to apply the appropriate inventory strategies for spare parts and optimize the inventory and maintenance costs.

Expand the model on SKU level: Prediction of failures on the SKU level is the company's goal. Taking the current model as a baseline, the company can enhance it with additional information on the SKU level (currently not available to the research team) and develop predictions on SKU level per site location.

Definition of inventory strategy: Demand for spare parts is the key input for the company's inventory strategy. Having accurate predictions on the SKU level and following a similar approach, the company can segment the SKUs in different clusters (using volume and variability of demand) and define the best strategy (push vs. pull) between the different storage locations (central warehouse, regional warehouses, and VANs). This will improve the responsiveness of the company, reduce the inventory ageing, and improve the services provided to the customer (reduction of downtimes and issues)

Supply planning: Supply planning and procurement decisions can be also optimized based on the predictions for site and SKU failures. The company can define more accurately the supply decisions and ensure the resilience for the organisation by comparing the predicted requirements with the suppliers' lead times for delivering the materials. On that basis, supply planning can be much more accurate, and the company will reduce significantly the "last minute orders" which can impact the procurement cost.

Global roll out – scalability: The current project was focused on one operating market of the sponsor. The company should scale up the model and include data and information from all the markets. This will give the opportunity to further enhance the model and identify additional benefits (e.g., how SPM and inventory management can be improved through cross market collaboration).

End to End Control Tower: The current capabilities of the control tower are limited mainly due to the fact that the data are not integrated. As we discussed in Chapter 3, the control tower mainly collects information from different sources and analyses in spreadsheets to define the procurement decisions and inventory levels. The company, through the integration of information in a machine learning system, will be able to get real time visibility, connect demand with supply, and design an Integrated Business Planning system managed by the control tower.

6.2 Project Limitations

1. **Scope limitations**: The scope agreed upon with the sponsor company was limited to the prediction of failures on the site level in just one market to explore the opportunity for a predictive maintenance approach. The project did not cover inventory management between the different storage locations; hence we could not identify opportunities from reduction of inventory levels. Moreover, the scope of the project was not to design a new process for predictive maintenance and / or define algorithms, which will be implemented in the demand planning system of the company
2. **Data limitations:** While we managed to collect data from different sources, which gave us the opportunity to explore improvements on the demand planning side, we were not able to capture the required information on the SKU level. This happened for the following reasons:
    - The quality of SKU data maintained in the ERP system. While the sponsor maintains SKU data, the accuracy of the information is not high mainly due to human intervention and manual processes (e.g., users do not record the correct material on the correct site and in many cases, they perform mass upload movements to the ERP in order to complete faster a task).

- Data availability. A big part of the sponsor's operations is managed by contractors. The sponsor's visibility ends at the contractors' storage location and there are no available accurate data for many sites. In addition, some required information was not available, so we had to proceed with assumptions. In particular:
    - Traffic information could not be shared due to confidentiality reasons, so we had to assume the potential traffic by using the population of the area, the sponsor's market share, and the service category of each site
    - Mean time to failure and mean time between failures as defined by the vendors were not available. Moreover, the installation date of each component was not available, so we had to exclude this parameter from the analysis
    - Replacement or maintenance of sites performed and recorded as normal deployment activities could not be captured.
- Accuracy of movements. As pointed out in the descriptive analysis (section 4.3), the users perform mass movements to dummy sites. These sites are either the storage locations of the contractors or the vans. As a result, we had to remove a significant amount of data from the analysis to avoid inaccurate outcome
- Other data limitations. Due to the nature of the process (MRO), the volume of data was not high (between 0 and 2 sites per day). As a result, we had to aggregate on a weekly level to perform time series analysis.

7 Conclusion

In this capstone project, we analysed how a telecommunications company can develop a proactive approach to sites maintenance by applying predictive analytics. We explored various methodologies and compared their results to show that machine learning techniques can bring significant improvements in demand planning by achieving forecast accuracy for site failures between 60% to 85%. This improvement can lead to further benefits for the company. Our sponsor expects that a more accurate prediction of failures can bring significant results in inventory reduction (up to 15%), ageing reduction (more than 10 days), and overall operating cost through the optimization of the process (up to 10% due to better planning and maintenance). Time series may have application in sectors where we have high volumes in the demand, but it seems that it does not provide the required accuracy in the MRO process. Nevertheless, time series should be explored in the future by the sponsor as an alternative method to predict the demand of spare parts on the SKU level.

Although the pilot project was limited in scope, the learnings can be applied to our sponsor’s businesses and used as a lever for more detailed research in order to define the business opportunities in other areas as well (e.g. inventory management). In addition, they can be useful for other businesses in asset intensive sector where MRO process has a key role in Supply Chain.

Future Research

Although our research was limited by data, it highlighted some great opportunities for the sponsor. This can be considered as a “quick win” for the sponsor and used as a starting point for a further in spare parts management process through machine learning.

Future research can expand the model and the machine learning algorithms to the SKU level. One critical point is to include additional features and review how the machine learning models interact. More specifically, human intervention and error is considered as one of the major reasons for assets failures. In our research, human error could not be included due to data privacy issues. Moreover, while our model takes into consideration weather predictions, it assumes that the traffic data are known. Given that supply and inventory strategies cannot be changed on a tactical level, it would be useful to feed the model with traffic predictions and develop scenarios based on the probabilities.

In general, it seems that machine learning can provide significant capabilities for spare parts demand prediction and enable a digital transformation for the operations of a company. However, to enable machine learning and receive accurate outcome, companies should ensure that the data quality and availability are very high.

References

Armstrong, M. J., & Atkins, D. R. (1996). Joint optimization of maintenance and inventory policies for a simple system. *IIE Transactions*, *28*(5), 415–424. https://doi.org/10.1080/07408179608966288

Aronis, K.-P., Magou, I., Dekker, R., & Tagaras, G. (2004). Inventory control of spare parts using a Bayesian approach: A case study. *European Journal of Operational Research*, *154*(3), 730–739. https://doi.org/10.1016/S0377-2217(02)00837-8

Chase, C. (2013). *Demand-driven forecasting a structured approach to forecasting*., 103 - 123, 125 - 158, 203 - 237 https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118691861

Elwany, A. H., & Gebraeel, N. Z. (2008). Sensor-driven prognostic models for equipment replacement and spare parts inventory. *IIE Transactions*, *40*(7), 629–639. https://doi.org/10.1080/07408170701730818

Gary Forger. (2018, February). *NextGen Supply Chain: NextGen Technologies Are Already Here*. Supply Chain Management Review, February 14, 2018 https://www.scmr.com/article/nextgen_supply_chain_nextgen_technologies_are_already_here

Gebraeel, N. (2006). Sensory-Updated Residual Life Distributions for Components With Exponential Degradation Patterns. *IEEE Transactions on Automation Science and Engineering*, *3*(4), 382–393. https://doi.org/10.1109/TASE.2006.876609

Jouni, P., Huiskonen, J., & Pirttilä, T. (2011). Improving global spare parts distribution chain performance through part categorization: A case study. *International Journal of Production Economics*, *133*(1), 164–171. https://doi.org/10.1016/j.ijpe.2010.12.025

Montoro-Cazorla, D., & Pérez-Ocón, R. (2006). Replacement times and costs in a degrading system with several types of failure: The case of phase-type holding times. *European Journal of Operational Research*, *175*(2), 1193–1209. https://doi.org/10.1016/j.ejor.2005.06.027

Mostafa, S., Dumrak, J., & Soltan, H. (2015). Lean Maintenance Roadmap. *Procedia Manufacturing*, *2*, 434–444. https://doi.org/10.1016/j.promfg.2015.07.076

Pérez-Pucheta, C. E., Olivares-Benitez, E., Minor-Popocatl, H., Pacheco-García, P. F., & Pérez-Pucheta, M. F. (2019). Implementation of Lean Manufacturing to Reduce the Delivery Time of a Replacement Part to Dealers: A Case Study. *Applied Sciences*, *9*(18), 3932. http://dx.doi.org/10.3390/app9183932

Saenz, M. J., Borella, I., & Revilla, E. (2022, April). DIGITAL SUPPLY CHAIN TRANSFORMATION: ALIGNING OPERATIONS AND STRATEGY. *Supply Chain Management Review*.

SAP. (2019). *The Intelligent Supply Chain for Assets*. SAP, 1 - 12 https://www.sap.com/documents/2019/09/a2f0412f-657d-0010-87a3-c30de2ffd8ff.html

SAP. (2021, June 9). *Industry 4.0: Intelligent Assets*. Supply Chain Digital. SAP, 1 -10, https://supplychaindigital.com/company/sap-industry-40-intelligent-assets

Selim, M., & Gurel, A. (2007). Machining conditions-based preventive maintenance. *International Journal of Production Research*, *45*(8), 1725–1743. https://doi.org/10.1080/00207540600703587

Statistics South Africa. (n.d.). *Department of Statistics South Africa*. Statistics South Africa. https://www.statssa.gov.za/?page_id=964

Vaughan, T. S. (2005). Failure replacement and preventive maintenance spare parts ordering policy. *European Journal of Operational Research*, *161*(1), 183–190. https://doi.org/10.1016/j.ejor.2003.06.026

Venkatesan, M. (1984). Production-Inventory with Equipment Replacement-PIER. *Operations Research*, *32*(6), 1286–1295.