SALAM AYMAN ABDEEN

FROM REPORTING TO ADVANCED ANALYTICS
A CASE STUDY ON ENERGY PRODUCTION

Orientador: Prof. Doutor Francesco Costigliola

Universidade Lusófona de Humanidades e Tecnologias

Escola de Comunicação, Arquitetura, Artes e Tecnologias da Informação

Engenharia Informática e Sistemas de Informação

Lisboa

2019

SALAM AYMAN ABDEEN

FROM REPORTING TO ADVANCED ANALYTICS
A CASE STUDY ON ENERGY PRODUCTION

Tese defendida em provas públicas na Universidade Lusófona da Humanidades e Tecnologias no dia 5 de Junho de 2019, perante o júri, nomeado pelo Despacho de Nomeação n.º 44/2019, de 14 de Fevereiro, com a seguinte composição:

- Presidente: Prof. Doutor José Luís Azevedo Quintino Rogado (ULHT)
- Arguente: Prof. Doutor Carlos Luis Menendez Gonzalez (EDP Espanha)
- Orientador: Prof. Doutor Francesco Costigliola (ULHT)

Universidade Lusófona de Humanidades e Tecnologias

Escola de Comunicação, Arquitetura, Artes e Tecnologias da Informação

Lisboa

2019

Dedication

I would like to dedicate this work to the souls of my father and mother, Ayman and Etedal, who have never left my heart, and whose words of encouragement and push for tenacity never left my mind. I hope you both are proud of me.

Salam

Acknowledgment

This master thesis would not have been possible without the assistance of several people.

First, I would like to express my deepest appreciation to those who gave me the possibility to complete my study and get the Master degree (Global Platform for Syrian Students Scholarship) and very special gratitude to Dra. Helena Barroco for her great efforts with us.

I would like to thank my thesis advisor professor Francesco Costigliola. For him having his door always open whenever I ran into a trouble spot or had a research question or doubt. With his professional guidance, I was able to finish this journey and develop this work.

I would also like to thank professor José Rogado for his professional support during my years of study.

And I would like to thank Mr. Armando Soares, from EDP/SKIPPER, who provided me with all technical support that I needed to finish this work.

I would like to thank Mohammad, my brother, for being always there for me and standing strong by my side.

I would like also to thank Khaldoun, my soul mate and best friend, for his unlimited love, patience and support.

Finally, I would like to thank all, friends, family, and university members for their assistance and support.

Resumo

No mundo atual, a energia está ligada a todos os aspetos e profissões da vida. O seu consumo está a aumentar exponencialmente, o que faz com que um sistema de gestão de energia, em qualquer país, seja um fator crucial para uma economia estável, um ambiente e um futuro seguros. Manter todos os sistemas em funcionamento de forma eficiente exige que um grande esforço na recolha e na análise de grandes quantidades de dados. Estas análises permitem extrair informações que auxiliam todos os processos como, por exemplo: avaliações de risco, tomada de decisões e definição de estratégias. Em suma, a análise avançada de dados precisa de ser parte integrante dos sistemas de gestão de energia.

Este trabalho analisa os processos e o nível de maturidade analítico de uma equipa (SKIPPER) na empresa de produção de energia em Portugal (EDP, Energias De Portugal). O trabalho estuda as plataformas de Business Intelligence (BI) existentes, a arquitetura de dados na plataforma SKIPPER e os requisitos dos clientes para aprimorar o processo de BI e integrar o conceito de advanced analytics implementando diversos modelos de previsão de longo prazo, comparando as suas diferenças e recomendando as ações necessárias.

Com foco na produção de energia térmica, este trabalho apresenta o SAS como uma nova plataforma de Analytics e BI comparando as suas funcionalidades com a plataforma de BI existente. Utilizando 27 meses de dados históricos, o SAS também foi apresentado como uma plataforma de previsão de séries temporais, oferecendo dois níveis de análise diferentes: análise de cenário e busca de metas. O R Studio é testado como uma segunda opção de advanced analytics, onde diversos modelos de previsão de produção de energia são implementados. Uma das vantagens está na possibilidade de estruturar os dados como uma série temporal hierárquica. Das diversas técnicas de reconciliação, a Optimal Reconciliation revelou ser a abordagem mais precisa. Adicionalmente, foi implementando um modelo de regressão onde foi possível extrair alguns dos fatores relevantes no que diz respeito à afetação dos níveis de produção de energia (como por exemplo, CO2 fine e horas de trabalho).

Os resultados do trabalho de tese mostram que utilizar o SAS como a principal plataforma de relatórios do SKIPPER e integrar o R Studio na arquitetura estatística pode melhorar as práticas de relatório do SKIPPER e elevar o nível de maturidade de análise de dados.

Palavras-chave: Business Intelligence, Big Data, Advanced Analytics, produção de energia, previsão.

Abstract

In today's world, energy is connected to all life aspects and professions. Consumption is increasing exponentially which makes having an energy management system, in any country, a crucial factor for having a stable economy, a secure environment, and a safe future. Keeping these systems running in an efficient way, requires a large amount of data to be gathered and analysed in order to extract information that helps in different processes such as risk assessments, decision making, and defining strategies. In short, data advanced analytics needs to be integrated into the energy management systems.

This work examines the reporting process and the analytics level of a team (SKIPPER) in an energy production company in Portugal (EDP, Energias De Portugal). The work analyses the existing reporting platforms, data architecture in SKIPPER and clients' requirements in order to enhance the reporting process and integrate data advanced analytics concept by implementing long term forecasting models, comparing their differences and recommending the necessary actions to be made.

Focusing on thermal energy production, this work presents SAS as a new reporting platform and compares its functionalities with the existing reporting platform. Taking into consideration 27 months of historical data, SAS is also presented as a time series forecasting platform where a dashboard is created offering two options (scenario analysis and goal seeking). R Studio is tested as a second advanced analytics option where an energy production forecasting model is implemented shaping the data as a hierarchical time series having the Optimal Reconciliation as the most accurate approach, and a regression model is implemented defining raw material prices, CO_2 fine and working hours as the most important factors affecting energy production.

The results of the thesis work show that utilizing SAS as the main reporting platform in SKIPPER, and integrating R Studio in the statistical architecture can improve the reporting practices of SKIPPER and raise its level of data analytics maturity.

Keywords: Business Intelligence, Big Data, Advanced Analytics, energy production, forecasting.

Abbreviations

BI	Business Intelligence;
EDP	Energias de Portugal;
EDPP	EDP Produção / EDP Production;
EDPR	EDP Renováveis / EDP Renewable;
GWh	Gigawatt per hour;
MIBEL	Mercado Ibérico da Energia Eléctrica / Iberian Electricity Market;
EU	European Union;
LDW	Logical Data Warehouse;
IoT	Internet of Things;
EDW	Enterprise Data Warehouse;
OSI-PI	OSI Process Information;
SCADA	Supervisory Control and Data Acquisition;
DCS	Distributed Control Systems;
KKS	Kraftwerk-Kennzeichensystem;
UNGE	Unidade de Negócios de Gestão de Energia / Business Unit Management Energy;
OPC	Open Platform Communications;
DOT	Diração Otimização Térmica / Thermal optimization direction;
Power BI	Power Business Intelligence;
SAP	Systems Application and Products in Data Processing;
SAP BO	SAP Business Objects;
SAS	Statistical Analysis System;
SAS EG	SAS Enterprise Guide;
SAS VA	SAS Visual Analytics;
DGU	Digital Global Unit;
USDA	US Department of Agriculture;
SAS EG	Statistical Analysis System Enterprise Guide;
SAS VA	Statistical Analysis System Visual Analytics;
DEC	Direção de Eficiência / Efficiency Directorate;
LDW	Logical Data Warehouse;

EDW	Enterprise Data Warehouse;
AR	Autoregressive;
MA	Moving Average;
ARMA	Autoregressive Moving Average;
ARIMA	Autoregressive Integrated Moving Average;
ACF	Autocorrelation Function;
HTS	Hierarchical Time Series;
MAPE	Mean Absolute Percentage Error;
MASE	Mean Absolute Scaled Error;

Table of Contents

List of Figures

List of Tables

Introduction

Background and Motivation

Energy has always been an important affecting factor in the development of any nation, no matter for the location or the type. Since that the consumption of energy is increasing globally, due to different reasons such as population and society growing’s, energy management is becoming more crucial and with the rapid technology development the researchers or companies are motivated to gather more and more data, analyze it and extract meaningful information out of it. This all leads to big data advanced analytics where a lot of algorithms can be useful to gather insights from the data, perform different analytics, such as forecasting, and improve the business in the energy company. Accurate predictions of energy production, based on internal or external underlying factors, can raise awareness of future trends and thus companies are more prepared for them with more efficient decision-making systems. Generally, regardless of the company main business field, and in order to build and maintain a high-level analytics maturity model in the company system it is required to establish a robust foundation of the data in a single source so that high-quality functions are performed to structure the tires of the analytics model in a way that serves the companies needs and interests.

EDP is the main company in Portugal which produces electricity out of several resources and based on the types of resources the energy is divided into three main categories, hydraulic, thermal, and renewable. Usually, most of the studies focus on the hydraulic or the renewable energy in Portugal, however, this research is focusing on the thermal production, which is based on coal and natural gas, with the goal of developing an advanced analytics system inside the company to predict thermal electricity production with future objective to involve other types of energy resources into the system.

Literature Review

During the last recent years, on a global level, the interest in forecasting energy demand and production has increased due to the importance of having a management system that guarantees the balance between the two aspects. Most of the researches done regarding energy forecasting are interested in demand forecasting or renewable energy production, and various forecasting approaches are applied in this objective.

In (Ghalehkhondabi, Ardjmand, Weckman, & Young, 2016a), exists a review of the energy demand forecasting methods that have been published as articles between 2005 and 2015, dividing these methods into two main categories, causal and historical data based methods, where the most used forecasting methods are artificial neural network and regression models, and the most common examples of methods that use past values to forecast the future ones are time series, grey prediction and autoregressive models. The work in(Gvaladze, 2015) is a comparison between statistical and machine learning forecasting methods on a dataset from the one power station in Hvaler – Norway to perform consumption energy forecasting, and the research found that the machine learning models performed better than statistical models and the best performance was using nonlinear autoregressive neural network. In (As'ad, 2012), four ARIMA models based on three, six, nine and twelve months of data are applied in order to predict the peak electricity demand for the first week of June 2011 starting from 31st of May 2011 in New South Wales, Australia.

The work in (Martín et al., 2010) presents a comparisons of statistical models based on time series applied to predict half daily values of global solar irradiance with a temporal horizon of 3 days. The used dataset in the study is from the stations of Spanish National Weather Service (AEMet), and the tested models are autoregressive, neural networks and fuzzy logic models.

> "The best approach to forecast half daily values of solar irradiance is neural network models with lost component as input, except Lerida station where models based on clearness index have less uncertainty because this magnitude has a linear behaviour and it is easier to simulate by models." (Martín et al., 2010)

The work in (Rodrigues et al., 2007) presents a proposed a wind power forecast system, EPREV, which has a human-machine-interface and a power forecasting model that uses several statistical models, power curve model (PCM), auto regressive (AR) and neural network assembling model (NNAM). based on historical data, the project has a purpose of forecasting up to 72 hours using mesoscale models and different boundary conditions. Three Portuguese wind farms were used for testing cases and validating the developed tools.

It is important to understand that the forecasting approach can be expanded and used for different types of energy production and sources.

Research Questions

As a multinational company, it is important for EDP to have a forecasting system, included in its business planning system, that takes full advantage of the gathered data since this will lead to better strategies formulation and policies definitions. However, and going through the methods and models used in similar situations, it would be fair to say that the efficiency of a method depends on the nature of the used data set and the required horizon of the forecasting. This research concentrates on SKIPPER, a team in EDP Production in one of the companies of EDP Group, going through its reporting system, the used tools and the role that SKIPPER is taking as the main data source of the energy generation operation. Taking all of the previous points into consideration, the following questions arise in the research:

- Is it possible to improve the existing reporting system in SKIPPER? And what are the suitable platforms to achieve the improvement?
- Based on the results of the first question, what is the best way to integrate advanced data analytics in SKIPPER?
- Focusing on forecasting analytics, what are the suitable frameworks to integrate in SKIPPER?

Research Objectives and Methodology

While this research was growing, two main objectives raised. The first aim of this research was developing the reporting system in SKIPPER to become more flexible and able to adapt to any change. The second aim was involving advanced data analytics techniques in the reporting system with the objective of fully utilize of the data. The existing reporting system in SKIPPER is based on high-quality data and used to support business decisions based on historical Key Performance Indicators (KPIs) to lead to the desired outcome. Even though this system forms a stable and good base, it is no longer enough to face the quick change in the market and stay ahead in the competition.

This research suggests, in order to achieve the desired objectives and answer the raised questions, implementing two improvements starting with applying a reporting platform, different from the already existed one in SKIPPER, with the objective of comparing both of the platforms, making the necessary alterations and choose the fittest one, and the suggested platform is one product of SAS software suit.

The second proposed improvement, to answer the second and the third research questions, is building up the right analytics model in SKIPPER, focusing on the energy production data, integrating a new BI system and implementing newly advanced analytics pilots to gather more insights from the data and use these insights in different forecasting techniques, and finally identifying the appropriate ones. And to build the suggested analytics model, the research suggests using another product of SAS as well, besides implementing statistical methods for time series data processing in the R Studio software.

The work eventually aimed to migrate SKIPPER reporting process, be closer to the end users, work with them within a feedback loops system that helps to assess performance improvement, drive innovation and define future plans and strategies.

Research Outline

The remaining part of the thesis is structured as follows. Chapter1 presents a business overview introducing the concept of data analytics, explaining the stages of its development, the fields where it can be implemented, and its categories. It also presents the case study organization, EPD Group, its structure, companies, and in particular EDP Production – team SKIPPER where this internship research was conducted. Chapter 2 introduces the concept of analytical maturity inside an organization and explains the changes that took place in the organization's structure and analytics behavior facing the changes of business processes adapting to the fact that data analytics has played a leading role shaping organization's processes and decision-making systems. Furthermore, it describes the current technological structure in the case company, EDP, and the current reporting process in the team SKIPPER. Chapter 3 explains the proposed methodology for improving the reporting process by proposing the implementation of a new reporting platform and integrating a forecasting model to support decision-making system. It also presents a theoretical background about forecasting in general and deepens in the forecasting techniques and specifically in the time series models. Chapter 4 presents analyzing collected data, the actual work of implementing a new reporting framework, and developing the proposed forecasting model, with the final results of both proposed steps. Chapter 5 concludes the thesis work, commenting on the overall success, and proposes potential improvements as future work.

CHAPTER 1

Business Overview

1.1. Theoretical Principles

The concept of advanced data analysis has been around for years, organizations have more data than ever at their disposal and they understand that, in order to take the full advantage, they need to capture all the data running in their businesses and apply advanced analytics to solve business problems. The use of advanced analytics allows companies to get the most of their data and use the insights to improve their businesses.

Advanced analytics goes beyond simple mathematical calculations or operations such as filtering and sorting, is based on statistical formulas and algorithms that enable optimization and improvement of the existing business, finding patterns and correlations, supporting organizations to be more successful and competitive, and provide new opportunities to satisfy the customer needs by giving the power to know what customers want which, in turn, leads to a smarter and data-driven business. In addition to easing the decision-making process, advanced data analytics is considered a cost-effective process. In fact, the cost of data storage is reducing, “big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data”(SAS, n.d.).

However, advanced analytics is divided into four categories, (1) Descriptive Analytics that analyses the data from the past as it answers the “What happened?” question, (2) Diagnostics Analytics tries to answer the “Why something happened?” question, then comes the (3) Predictive Analytics that answers the “What is likely to happen?” question, and finally the (4) Prescriptive Analytics which answers the “What should be done about it?” to offer guidance on the best course of action (Kumar, 2017)

The four categories complete each other in a progressive order to accomplish the objective of the analytics dividing the tasks of the categories in two steps: the first step is going through the data collection, and having a sense of what the company wants to improve in the business and defining the objectives regarding either signing employees to the teams or getting products into the market more quickly, the company also need to define the risks and the protective actions.

Once this information is available, the management team will be able to decide what kind of data resources will be needed, what kind of investments are recommended and what is the most suitable team structure, this step is called “data to insights” which is learning about the business and figuring out what is new (Manual, Wiseman, & Zecha, 2016), and the next step is using these insights to change the business with it, and this is called “insights to actions”, which means making the necessary information available for the persons who need it in real time in order to make change with it.

Advanced data analytics is being applied to various types of organizations as the main technology to achieve the objective of satisfying customer needs. However, the approach that each organization follows to apply and implement analytics is different, as the customers of each sector have different needs and the time that the organization should satisfy these needs is also different. The most obvious example is sales and marketing, where the most successful marketing enterprises are the ones that are integrating advanced analytics in their campaigns to measure the effectiveness of their plans and models and customers’ satisfaction of newly offered products. Moving to the travel industry, data analytics helps to, immediately, collect the data, analyze it and identify the potential problems and the appropriate solutions. “By analyzing large amounts of information – both structured and unstructured – quickly, health care providers can provide lifesaving diagnoses or treatment options almost immediately” (SAS, n.d.)

The adoption of analytics to improve businesses presents some challenges. “The biggest challenge has nothing to do with data science or mathematics or data storage, it has to do with legal and governance framework” (Manual et al., 2016). The organizations should follow a framework of rules regarding customer protection and employee protection, the framework identifies the data that they can use, the way they can process it and what they can use it for. The organizations need to provide promises of trust that they will keep the customer information secured, and the same goes for the employee information. They need to explain how they will use the information to manage and improve the performance as the data protection and confidentiality laws are increasing but in a different way in the markets and the countries which form another struggle to the organizations. Furthermore, choosing the appropriate analytics method to be applied is also challenging as there are many algorithms and methods and choosing the best one depends on the analytics team members and their abilities.

In a more general overview, analytics is an immense field that refers to team skills, technologies and applications chosen to investigate the data and drive the business planning. Analytics consists of two major areas: Business Intelligence, BI, and Advanced Analytics, where BI looks at historical information to learn what has happened by measuring and monitoring, advanced analytics helps making predictions and forecasts based on examining historical data in new ways. “Business Intelligence is focused on reporting and querying, Advanced Analytics is about optimizing, correlating, and predicting the next best action or the next most likely action” (RapidMiner, n.d.). However, this classification doesn’t mean that organizations can skip the work associated with BI to do the advanced analytics immediately because it is not possible to create prediction models without having a historical data to base the models on it. Taking a closer look at the mutual and different points, both BI and Advanced Analytics work on big data, either structured or unstructured data, and the difference lies in the orientation of the work, as BI’s orientation is rearview and the initiatives are reactive while the advanced analytics has a future orientation and the initiatives are proactive.

Furthermore, BI and advanced analytics are different with the methods they depend on, BI depends on reporting, dashboards, scorecards and automated alerting, while advanced analytics use predictive modeling, data mining, and descriptive modeling besides the statistical and quantitative analytics.

1.2.EDP:

EDP is a multinational, vertically integrated company that operates in 14 countries and four continents. Through a history of 40 years, EDP has been present in both sectors of electricity production value chain and gas commercialization activity, and has been considered the largest electricity generation company in Portugal, one of the largest gas distributions in the Iberian Peninsula, the fifth largest private group in Brazil, and the third largest electricity production company in the world where almost 70% of its energy is produced from renewable resources. EDP Group has around 12000 employees and provides electricity to almost 10 million customers (EDP, n.d.-a).

The electricity production chain is composed of four large areas of application, Generation, Transport, Distribution, and Commercialization, and EDP is present in three of these activities.

Generation is the first activity in the electricity sector's value chain, energy production is categorized depending on the sources which are either renewable sources such as water, wind and sun, or non-renewable sources such as coal, natural gas, nuclear and cogeneration.
In addition to the previous categorization, it is important to know that the production is also categorized into an ordinary regime and a special regime, the ordinary regime is typically related to electricity production centralized in power plants which are large hydric power and thermal power plants where each power plant might be composed of several groups, while the production in the special regime is achieved through mini-hydro power plants.

The activity of transporting electricity is handled by the national electrical network that links the production centers to the distribution network that handles the flow of energy into electricity supply points. EDP operates in three electricity distribution markets, Portugal, Spain, and Brazil. "The distribution network consists of high, medium and low voltage cables and lines, the distribution strategy focused on implementation of smart networks and related services, to meet future challenges and become an electricity distribution benchmark" (EDP, n.d.-b).

EDP Commercialization handles the activity of selling the energy as it gets to the supply points. EDP is present in the electricity supply activity in Portugal, Spain, and Brazil, while in the gas sector it is only present in Iberia. Commercialization is also responsible for the relationship between EDP and the end customers, and this relationship has a big matter for EDP as it is trying to focus on the clients and provide them with high quality services to establish itself in the market as a trusted and related-to company.

The previous four companies, which are involved in the electricity production value chain, are only part of the EDP Group. EDP Innovation forms also a part of EDP Group, and is responsible of looking for entrepreneurial creativity spirits, finding new and disrupting ideas, prototyping them and transforming them into reality by providing the resources and the technology. EDP Group also includes EDPR, which is the company that invests in the renewable energy sector, is currently present in 13 countries and stands as the world's fourth-largest wind energy producer.

In Europe, EDP is present in Spain in the activity of generation, distribution, and supply of energy, besides that EDPR has it headquarter in Madrid. In Portugal, EDP is also present in the activity of energy production, distribution, and supply chain, EDPR, and EDP Innovations are also present in Portugal. In Belgium, France, Italy, Poland, Romania, and the UK, EDP is present through EDPR in the renewable energy segment.

In South America, EDP is present in Brazil in the production, distribution and supply of electricity, while in North America, EDP is present through EDPR in Canada, Mexico and USA. EDP Group has institutional representation in China and Angola.

1.2.1. EDP in Portugal

EDP Group is a private Foundation, present in Portugal as the largest corporate institution, and since its creation in 2004, many social, cultural, scientific and volunteering activities were implemented as a part of EDP contribution in the social development. EDP is present, as the main investor participating in the development of the national economy being the largest producer, distributor and supplier of electricity in the country.

The energy production is handled by EDP Produção / EDP Production (water, coal or natural gas) and EDP Renováveis / EDP Renewable (wind energy). EDP Distribuição / EDP Distribution is the Distribution Network Operator in the mainland of Portugal. EDP Serviço Universal / EDP Universal Service and EDP Comercial / EDP Commercial ensure the Group's energy supply.

In numbers, EDP has around **7000** employees in Portugal, and since the inception of EDP Volunteer Program, in **2011**, the program has more than **30000** entries (EDP, n.d.-d). In **2017**, as shown in Figure 1, EDP Group had an installed capacity of **11.335** MW, distributed **44**,**753** GWh of electricity to the network, and commercialized **21**,**489** GWh of electricity and **3**,**890** GWh of gas to the final customer in Portugal.

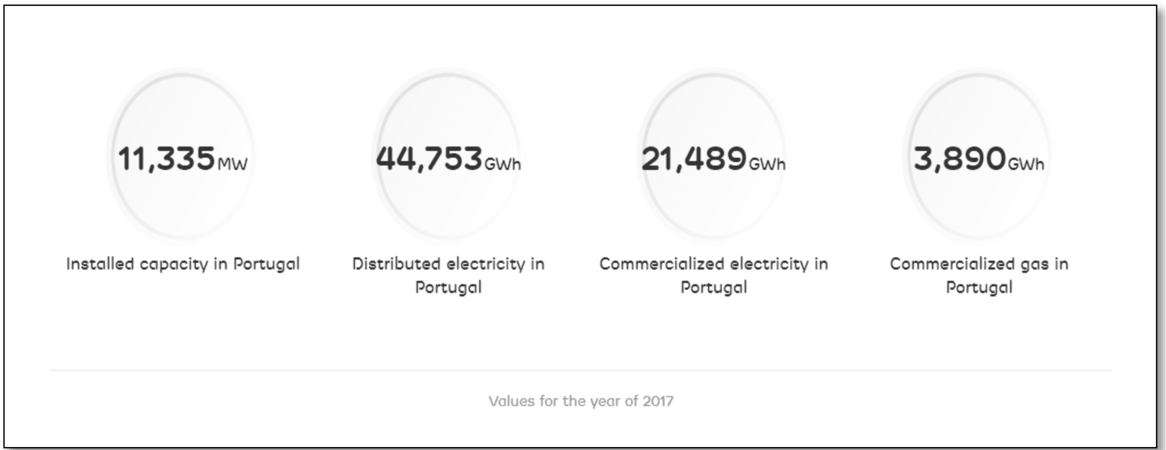


Figure 1 EDP Operational Indicators in Portugal

Image available at:

https://portugal.edp.com/en/main-indicators-portugal

1.2.2. EDP Production in Portugal (EDP Produção):

EDP Produção, is present in the activity of generation in the electricity sector's value chain of EDP, with focusing particularly on hydroelectric energy, coal energy and natural gas combined cycle.

Electricity production in the ordinary regime is based on traditional and non-renewable sources and on large hydro-plants. "The ordinary regime has been operating on a competition basis since 2007, after the implementation of MIBEL" (EDP, n.d.-c).Since then, the power plants started to share their energy on a common, integrated Iberian energy platform. After the implementation of MIBEL, making the plants' operation decisions was moved to the operators themselves to achieve a decentralized order system.

The special regime production in Portugal includes electricity production processes through mini-hydropower plants, cogeneration and biomass.

> "Special regime generation has been encouraged by EU policies, with the definition of the technical conditions for distribution network connections and purchase guarantees for all energy transferred into the network, in compliance with the remuneration processes set out in various legal documents" (EDP, n.d.-c).

1.2.3. Team SKIPPER

SKIPPER is an initiative of EDP Production in the Generation area, started as a small idea of having a "lake of information" with several objectives that can be summarized in providing the companies of EDP Group with an integrated information system to support management and monitoring of production assets, optimizing efficiency of existing asset management and supporting the internationalization of the EDP Group.

The name, SKIPPER, is an abbreviation of a group of words: System, Knowledge, Information, Plant, Performance and Environment, however, the name has an indication to the person who is going through the "information" lake seeking for "Knowledge". SKIPPER provides data (real-time, historical, raw, processed, organized and validated) from most EDP Group centers in a continuous and relatable manner, and in an environment, that provides the necessary exploration tools in accordance with the needs of the EDP units and where the results of the work of each employee can be shared.

The solution provides reports, dashboards, and graphical presentations. SKIPPER contributes to the breakdown of barriers to access information, where, at a more accurate level, EDP organization has an available series of information to support strategic planning and decision-making processes in an automatic way.

In SKIPPER, the production is divided into thermal and hydraulic energy. Thermal production is divided in its turn based on the used raw material type into gas and coal, while hydraulic powerplants are divided based on the production and income into gold and silver, where the type gold refers to hydraulic power plants that produce more comparing with others and thus the gain of selling energy from these power plants is more.

CHAPTER 2
State of the Art

2.1. Analytical Maturity Assessment

In any organization, data represents the starting point for any decision-making process, however, the data volume, variety, and velocity is increasing day by day, the sources of data are becoming more diverse including internal and external sources, data and analytics have a more active role that in not exclusively limited to support decision making; analytics are now empowering new activities, participating in shaping the interaction model with customers and leading the business processes. "In short data analytics is the brain of the enterprise, becoming proactive as well as reactive, and coordinating a host of decisions, interactions, and processes in support of business and IT outcomes" (Hagerty, 2016). These changes force organizations and their team of professional analysts to adapt in order to achieve the best outcomes. They had to start by managing and framing the data in the right structure, could be an LDW, apply advanced analytics and provide the end user with self-service data access.

Furthermore, data is rapidly shifting to the cloud and the demands to share data across business ecosystems is increasing, and to keep up with these changes, organizations are supposed to make some changes in the structure followed to store the data, also change the architecture of the professional team that supports the analytics seeking for new skills and roles so that the team can design a new data management that is more flexible, supportive to the changes in the analysis needs, and lower-cost utilizing analytics platforms that enable the organization to be successful and competitive in the digital business era. However, the new data management needs to accommodate with traditional as well as the advanced analytics techniques as the shifting process can't be immediate, even though this process is inevitable, it is recommended that it occurs gradually and incrementally (Hagerty, 2016).

Organizations might differ in the way they acquire data, some of them capture data coming in searching for insights and possible actions, while other organizations start from the desired goals to build up and manage a data architecture that supports those outcomes. Regardless of the followed approach, according to Gartner[1] (Hagerty, 2016), both types of organizations need to have a system architecture that has data, insights and actions integrated into it and includes four steps, which are shown in Figure 2,

1. Data acquisition, wherever data is coming from;
2. Data Organization, using an LDW at the core to connect relevant data across various platforms;
3. Analysis of data when and where it makes the most sense – including reporting and visualization, machine learning and everything in between;
4. Delivery of insights and data that can be used for both, supporting human activities and feeding analytics algorithms that recommend the best actions to implement.

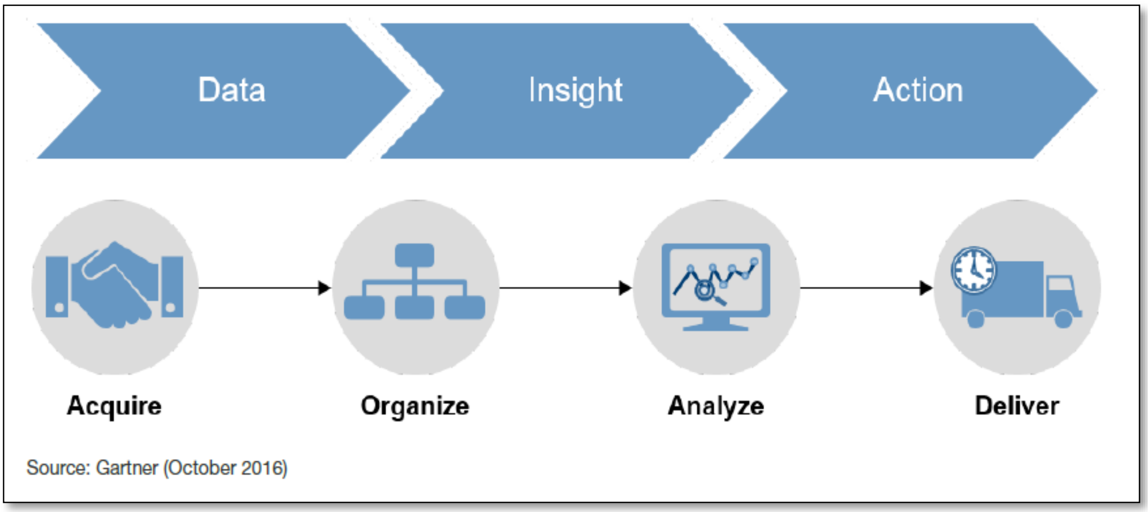


Figure 2 The Revitalized Data and Analytics Continuum

Image available at:
https://www.gartner.com/binaries/content/assets/events/keywords/catalyst/catus8/2017_planning_guide_for_data_analytics.pdf

[1] Gartner is a global research and advisory firm providing insights, advice, and tools for leaders helping them to make the right decisions.

2.1.1. End-to-End Architecture:

The previous explained system architecture is referred to as an end-to-end architecture. The process of building this system starts with "gathering" data, as it is already being explained, data sources are becoming more diverse, the thing that might require more attention from the analytics team, for example, IoT (Internet of Things) is emerging as an external source of data which means the need of integrating a real-time data analysis technique, internal log data needs to be inspected in real time to make sure it is healthy and protected.

This step shouldn't be only about collecting all the data that comes in into the organization hoping to find a way to use it. This process should always include an assessment of the data coming in. This could be done using some machine learning algorithms, and using the result to understand to understand weather keeping or discarding the analyzed data. This data filtering step is considered one of key aspects to achieve the required end-to-end architecture.

Another key aspect is taking advantage of the content available from third parties that could add much value to the already collected data in an organization, a third party could be a traffic information, weather data or risk management data for insurance. "Enabling the data and analytics architecture to embrace these new forms of data in a more dynamic manner is essential to provide contextual information needed to better support data-driven, digital businesses" (Hagerty, 2016).

Organizing the acquired data is not an easy task as the structure needs to support the emerging data revolution. EDW which is the usual traditional way that organizations used to structure their data in, was no longer able to both the explosion of interactive and sensor generated data as well as the more traditional transactional data, and a new structure was needed, that is when the environment of LDW appeared in the end-to-end architecture. "It's an agile architecture for developing BI systems, in which data consumers and data stores are decoupled from each other" (Lans, 2016). LDW can be seen as a collection of query-able data models that execute queries to return resulting data to the calling application. When these models are invoked, data accessed directly from the sources, sources are logically federated. In an LDW, only a little hardware required as the data is not moved to an intermediate store and processed immediately.

Building an LDW requires a combination of different technologies to get the full advantages. Existing data stores need to be understood very well by the technical team besides many data integration techniques, and it is always a good practice to choose the techniques that are most suitable for the organization's requirements (Hagerty, 2016). LDW provides a data architecture that is scalable, fast to implement and access, regardless of the source, supporting incremental development approach the thing that prevents the failure that might come out of an immediate change and gives the organization the opportunity to leverage an existing enterprise data warehouse.

Analyzing data can range from being simple to increasingly multifaceted as the demands for predictions and real-time reactions grow. Since most of the organizations' efforts are focused on descriptive and diagnostic analytics and a considerable part of this work is done by the users in a self-service manner, organizations need to focus more on the predictive and prescriptive analytics, provide more support for advanced analytics capabilities, invest in automation-business techniques such as machine learning, data sciences and cognitive computing to be able to cover all the analytics demands starting from traditional data reporting reaching he advanced prediction demands. The last phase of the end-to-end architecture is delivering data to the users, the actions of this phase involve human-to-data interfaces. Information delivery is handled in different methods, users can receive information on their mobile devices where they have the right access to the information that satisfies their needs, or they can use applications, with in-context analytics results embedded in them, to support their activities with just-in-time information.

Another way of delivering information is done from the inside of an organization, usually the output data of one activity is used as an input data to another one, however, the organization can store part of its resulting insights in another data store that is accessed by third parties who need these insights to make business decisions.

2.1.2. Data Analytics Maturity

Gartner defines five levels of data analytics maturity in an organization. Low level of maturity might form an obstruct in the way of modernizing BI, and if an organization wants to maximize the value of its data assets, the analytics team needs to improve the maturity level first (Moore, 2018).

An organization can be at the basic, opportunistic, systematic, differentiating or transformational level. Evolving from one level to another requires taking steps in the areas of strategy, people, governance and technology. Figure 3 shows an overview of the five levels of data analytics maturity and explores the differences between them. In the basic level, the BI capabilities are generally limited to spreadsheet- based and data extraction is performed personally. Organizations at the opportunistic level, have individual initiatives rising as a stand-alone project. However, the organization misses a common structure gathering these projects.

At the systematic level, a cooperation between the analytics teams and tier projects starts to appear, and techniques such as Agile emerges in development process. At the differentiating level, the organization considers data analyzing customer as an essential factor for business development and innovation. “Organizations at transformational levels of maturity enjoy increased agility, better integration with partners and suppliers, and easier use of advanced predictive and prescriptive forms of analytics. This all translates to competitive advantage and differentiation” (“Gartner Survey Shows Organizations Are Slow to Advance in Data and Analytics,” 2018).

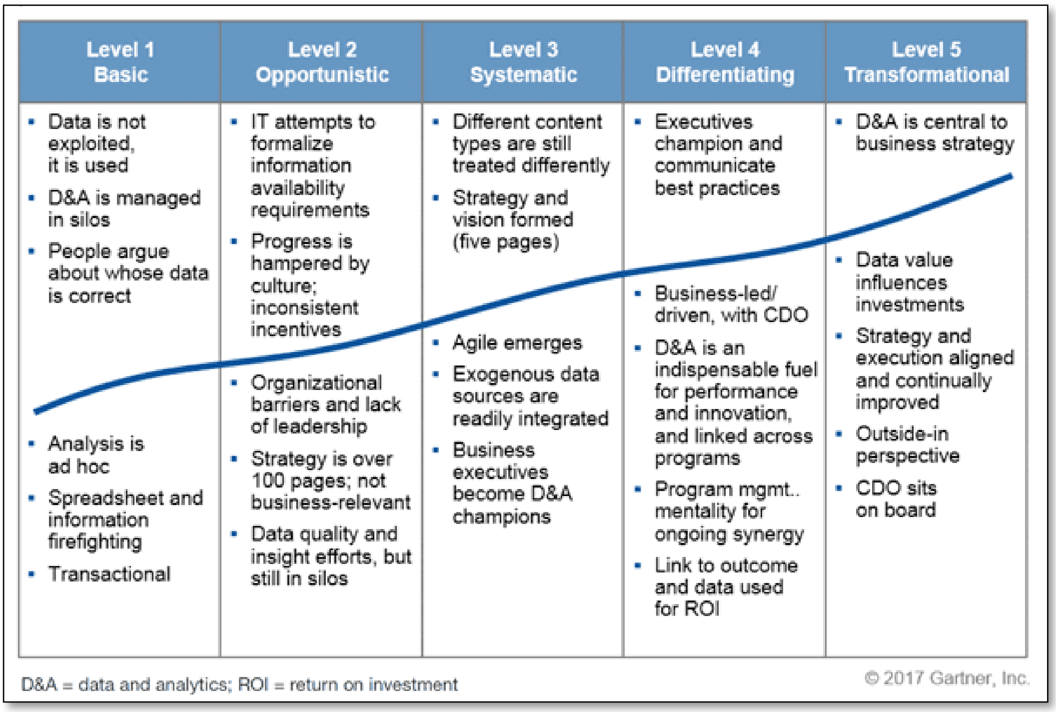


Figure 3 Overview of data analytics maturity model

Image available at: https://www.gartner.com/en/newsroom/press-releases/2018-02-05-gartner-survey-shows-organizations-are-slow-to-advance-in-data-and-analytics

Having all that said, it is obvious that embedding an advanced analytics system inside an enterprise structure is a very important necessity, decisions should not be made based on instincts, decisions should be made based on facts and insights. Even though it takes a while, leading organizations are already following this perspective, advanced analytics are going far beyond the traditional usages, analytics are being used whenever are wherever they are needed regardless the type of the enterprise and the objective that the organization is employing the analytics for, whether it was supporting the next big strategic decision or optimizing existing business transactions and interactions.

2.1.3. EDP Maturity Level

Since that EDPP has already started the process of integrating its customers in the decision-making process, where they form a part of the development activities and their experience is considered an important factor to the process of driving business projects to get the best outcome, and considering the previous two sections, it is possible to say that EDPP is on the Systematic level.

Regarding the team SKIPPER, the reporting process is based on a feedback loop where the customers participate in defining the KPIs, the design and the platform of the report, which helps to get the expected output that satisfies their needs and demands.

2.2. Technological Framework (Data Structure)

The data system of EDP Group, either in Spain, Brazil or Portugal, is based on implementing an OSI-PI system[2], as shown in Figure 4. The OSI-PI servers are distributed in the areas of the power plants, they are responsible for collecting real-time data, archive it and distribute it, and these servers are connected eventually to the EDP network (Rede EDP). Portal SKIPPER, where this work will be focusing on and held up within, is connected to the EDP network in a way that it has access to the data stored there. In a general view, the structure of SKIPPER starts from PI Data Archives where real-time data signals coming from the power plants are stored. These signals are divided into two types: (1) energy signals and (2) process signals that are stored in defined servers depending on this typology. Energy signals are values with aggregation of 15 minutes while the process signals are instant values.

[2] OSIsoft is a support software used to capture, analyze, and store real-time data, targeting markets such as: oil and gas; chemicals and petrochemicals; metals and power and utilities.

In the PI Data Archives structure, every thermal power plant, Lares, Ribatejo and Sines, has its own OSI-PI server that is located physically in the power plant, while this is different regarding hydraulic power plants where they are divided into three categories depending on their geographical location, Cavado de Lima, Douro and Tejo Mondego, this categorization is based on the main rivers in Portugal and the power plants near them, each one of these three categories has two types of servers in the PI Data Archives, one for energy signals and another for process signals.

More in details, the data stored in the PI Data Archives servers is acquired from different sources, mainly form two systems: SCADA and DCS located in the power plant groups, these systems consist of a set of sensors and actuators in the fields connected to networked supervisory computers that do the logical calculations of collected real-time data.

However, there is data that is manually fed up to the OSI-PI using VBA and PI-SDK which are tools used to access the OSI-PI Server. In the current data structure, the collected data is stored in PI Data Archives and prepared to be uploaded to Rede EDP (EDP network).

To express the collected data signals in an understandable way, SKIPPER has followed the KKS system to give these signals meaningful names called tags. KKS (Kraftwerk-Kennzeichensystem / Power Plant Designation System) is a German identification system for the uniform and systematic identification of power plants, and by looking to any tag, it would be possible to identify the power plant, the group and the equipment from which the signal was measured. These tags are important when an operation failure or an information delay is detected, and it is required to analyze the causes of the failure, and thus it is important to configure them correctly and in an equal naming manner on both sides, the data archives and the main OSI-PI server interface, to avoid mistakes and make it easier to access the data and define the failure source and reason.

At the same time of uploading the data signals to EDP network, it is also transformed into an Oracle database, 15 minutes of aggregation for the energy signals and instant, hourly, daily and monthly aggregated values for the process signals. However, the data existed in the Oracle database is not only the real-time data coming from the power plants, the Oracle database contains data related to the equipment existed in the fields, as well as their maintenance, and this data is retrieved from the SAP System to the Oracle database. The second data source for the Oracle is the TDMI which is the system available at UNGE that contains marketing data related to selling, buying and pricing processes.

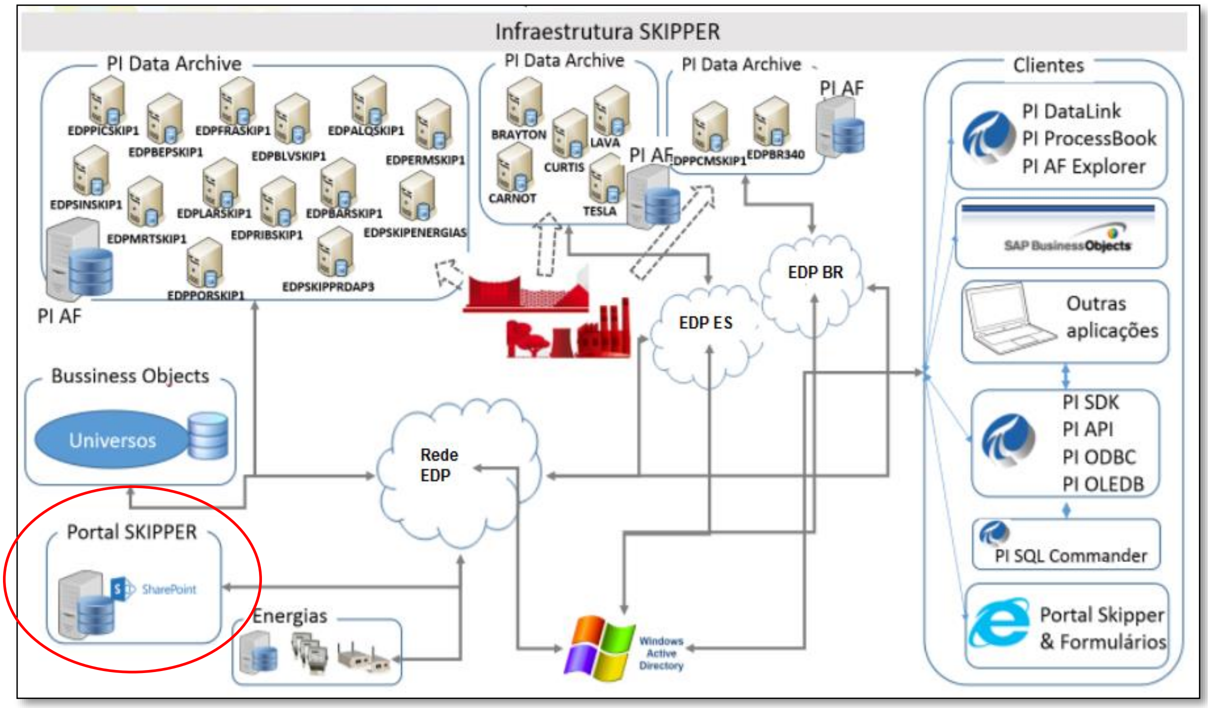


Figure 4 EDP – SKIPPER Infrastructure

2.2.1. New Data Architecture:

Skipper is working now on implementing a new data architecture, where the data stored on the EDP network is moved to an EDP Data Centre that holds a group of OSI-PI servers recording the signals with different time intervals depending on the signals types aiming to enable clients to perform queries on data as they fit. The main idea of this new architecture is to replace the old architecture based on several servers in the PI Data Archives. The new architecture is structured as follows:

1. One server responsible for two thermal power plants, Lares and Ribatejo, since both depend on natural gas in the production process, in terms of data processing.
2. One server responsible for Sines, the thermal power plant that depends on coal to produce energy, also in terms of data processing.
3. One server responsible for all the hydraulic power plants, regardless of the previous geographical aggregations.
4. Energies Server, that is responsible for storing energy data signals from both types of power plants hydraulic and thermal.
5. Interfaces Server, and is responsible for navigation between the incoming requests and the rest of the servers.

The first three servers are already implemented and Figure 5 shows a closer view on the structure (old and new) for the thermal power plant of Sines that uses coal for energy production and includes four production groups.

The structure starts with SCADA and SCD systems, located in each one of the four groups, collecting real-time data signals, the data is then subjected to a simple processing and aggregating operations, transferred to the OPC Server and then to the OPC Client, then the process signals are moved to the EDP Network and then to the defined server of Sines in the Data Centre. It is important to highlight the fact that not all the power plants have the same structure regarding the first part, some of them might have the OPC server and client in the same machine to reduce maintenance cost, some of the power plants have a firewall running before uploading the data to the network.

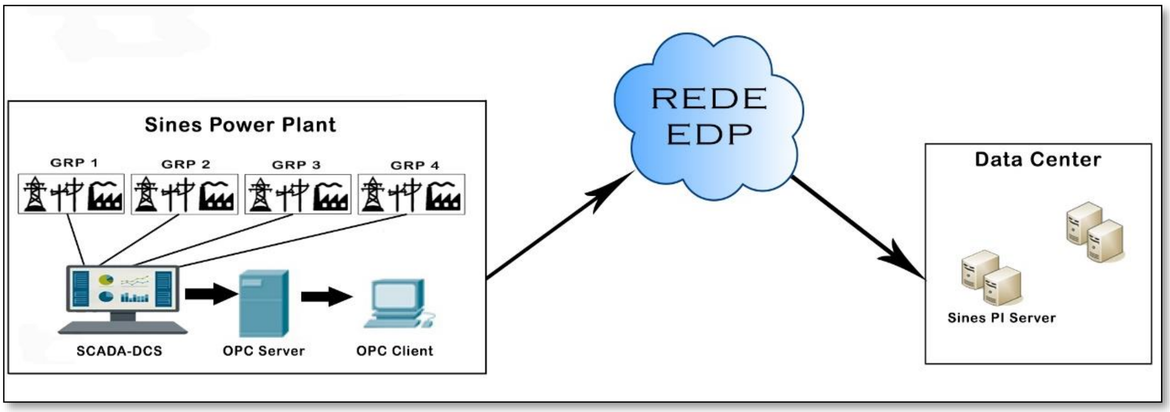


Figure 5 Sines Infrastructure, old and new

2.3.Reporting

The success of advanced analytics, within an organization, is not complete without presenting the resulting information so it is used in the best way, having an effective reporting system provides a greater depth of information to help the managers to make better informed decisions that lead to increase operating efficiency, understand the current position of the organization and consequently remain competitive. Without reporting management, employees may notice the existence of a problem, but they will not be able to identify its origin, reporting sets the first step toward finding the solution.

As a definition, a business report is a document that provides information, and sometimes analysis, to assist a business in making informed decisions, with the purpose of making the company's relevant data easily available to everyone in the company. Additionally, the business report needs to be as clear and concise as possible, and data needs to be understandable for the reader and this is done by clearly defined structure, formats, sections with labels and headings, and by representing the ideas with the right graphs, charts, or tables ("Structuring a business report | Oxford Dictionaries," n.d.)

The structure of one report depends on its type, and reports are usually categorized into two categories, Analytical and Informational reports, and both have a great importance inside an organization even though they are different in structure and usage. Informational reports contain data, facts, and results, without any analysis or recommendation, and are used to share this information with a group. They are, generally, brief and direct without explanations, suggestions or personal opinions, thus they are not helpful for instant decision making. Analytical reports, on the other hand, include information, analyze and interpret this information and the objective is to present some recommendations regarding an existed problem, thus they are useful for decision making.

2.3.1. Data Tools in EDP-SKIPPER

As soon as the data signals are stored in the EDP network, it is possible to use different types of tools to perform queries on the data and start the reporting and the analysis phases. The main tool used in SKIPPER is the Asset Framework (AF) as the tool to build up the basic structure of the data as a hierarchy that includes the power plants classified depending on the type of technology used to produce energy (thermal or hydraulic) and the geographical location (Portugal, Spain or Brazil). The hierarchy also includes the equipment existed in the power plants. For every element in AF, a set of attributes are defined and some attributes may be calculated values. The data hierarchy that is created in AF performs the base hierarchy that SKIPPER uses in other tools such as DataLink which retrieves the data from AF and is linked to Excel. DataLink functions are embedded in Excel, users from different teams in EDP can perform queries, retrieve data and build spreadsheets and reports with either compressed, sampled or calculated data during a period of their own choice.

SKIPPER also implements PI Vision for visualization purposes where a set of graphs is created to simulate the data hierarchy of the power plants and the actual production operation with actual numbers on the graphs. Another PI System tool used in SKIPPER is ProcessBook that works as an integrated component providing real-time values.

2.3.2. Reporting in EDP-SKIPPER

Reporting process in SKIPPER is basically presenting measurements and indicators of production and maintenance operations that have taken place during the current year and the previous years, each report delivers a set of data, facts, and results regarding a specific matter, which means that the reports belong to the informational category. Part of the reporting process is based on systems that have direct connections to the Oracle database store. These systems provide a set of tools to perform queries and select what kind of information to present in a report. SKIPPER usually provides reports and dashboards to the DOT (Direção Otimização Térmica) an administration team that is responsible for thermal power plants. For internal reporting, DOT requires using Microsoft Power BI – Desktop to present the indicators of production operation in the three thermal power plants in Portugal, Sines, Lares, and Ribatejo.

- **Power BI in SKIPPER**

The data source of the dashboards is a set of excel files uploaded manually, the indicators include electricity production, emission, reception, energy consumption, hours of activity, hours of inactivity, number of planed, non-planed, successful and unsuccessful stops for each group of the power plants. The dashboards cover a period that starts from 2016 until the current date, they contain different types of charts, line, bar and radar charts, with the options of choosing which power plant, group of the power plant, month, or year to display on the chart, and provide single statistical values according to the filtration applied on the charts.

Figure 6 shows an example of one page from the dashboard created in Power BI, the page has one line-chart that explores the Eficiência Bruta (or Total Efficiency) measurement that indicates efficiency of the power plant during the production process, mathematically, Efficiency is the quotient between the energy produced, by the energy value of the fuel consumed in parallel.

The page also includes three bar charts, the first one, with its cluster grouping style, explores the Utilização de Energia (or Energy Usage) measurement that indicates the percentage of the actual work of the group/power plant and the maximum work capacity that it can perform, mathematically it is the quotient between the value of the emitted energy and the value of the nominal energy.

The other two bar charts have the stack grouping style, one of them explores the Produção (or Production) measurement that indicates the amount of produced energy in a group(s) or the power plant as a whole, and it is measured by [GWh], and the second chart explores the Consumo Energia Primaria (or Primary Energy Consumption) measurement that indicates the primary energy needed in order to put the group/power plant working.

On the top of the page, three single statistical values are placed presenting average values of the indicators in the charts, and on the left side of the dashboard exists the filtration options where, as showed, Lares power plant is selected along with the two sub-groups and displayed for only the year 2018 until May.

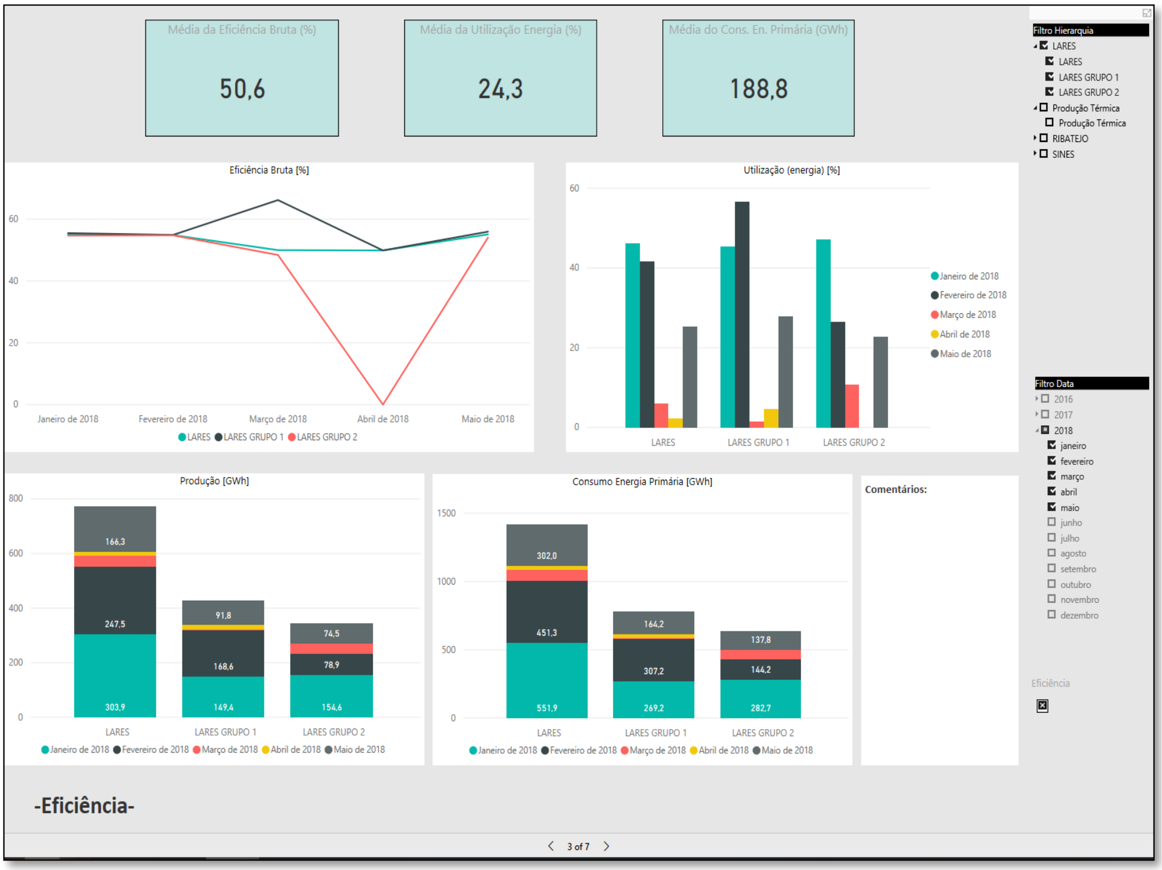


Figure 6 A page from the Power BI dashboard, presented by SKIPPER to the DOT

- **SAP in SKIPPER**

The second system used for reporting in SKIPPER is SAP Business Objects (SAP BO). SAP is an enterprise that was found in 1972 in Germany when five entrepreneurs had the vision of creating a standard application software for real-time business processing. SAP developed through several releases and several years, started from integrating real time data in business during the 80s, lunched the concept of client-server during the 90s, upgraded the performance to include the web and cloud computing during the first ten years of this century, and finally introduced the SAP HANA platform that is based on in-memory computing with the objective of enabling the users to access and analyze data anytime and anywhere in a period of seconds. During its development, SAP has acquired various companies and included their services into its own, and in 2008 SAP completed successfully its acquisition of Business Objects which was a French enterprise software company, specializing in business intelligence (BI) that was first founded in 1990. SAP BO is a business intelligence analytical suit produced by SAP combining the French business intelligence solutions with its software portfolio, it provides tools for reporting, building dashboards and applications and for data discovery[3].

The key feature of SAP BO is the Semantic Layer, or the universe, which performs as a view of the organization's data. And because universes are based on business terminology that is familiar and shared throughout the organization, they shield users from the technical complexities of the databases where the corporate information is stored (Aldeeb, 2017)

SKIPPER has been using SAP BO since the start of the team as the main system that has a direct connection to the Oracle database, and it is used to create reports that contain mostly tables regarding all types of operations in both types of power plants. The data in SAP BO includes technical information about the power plants structure, their affiliated groups, dates, statistical numbers, the tags used in the OSI-PI system and their indications, measuring units, and power resources.

Figure 7 shows an example of an SAP BO universe used in SKIPPER containing technical and operation exploration data.

[3] Historical information available at https://www.sap.com/corporate/en/company/history.html

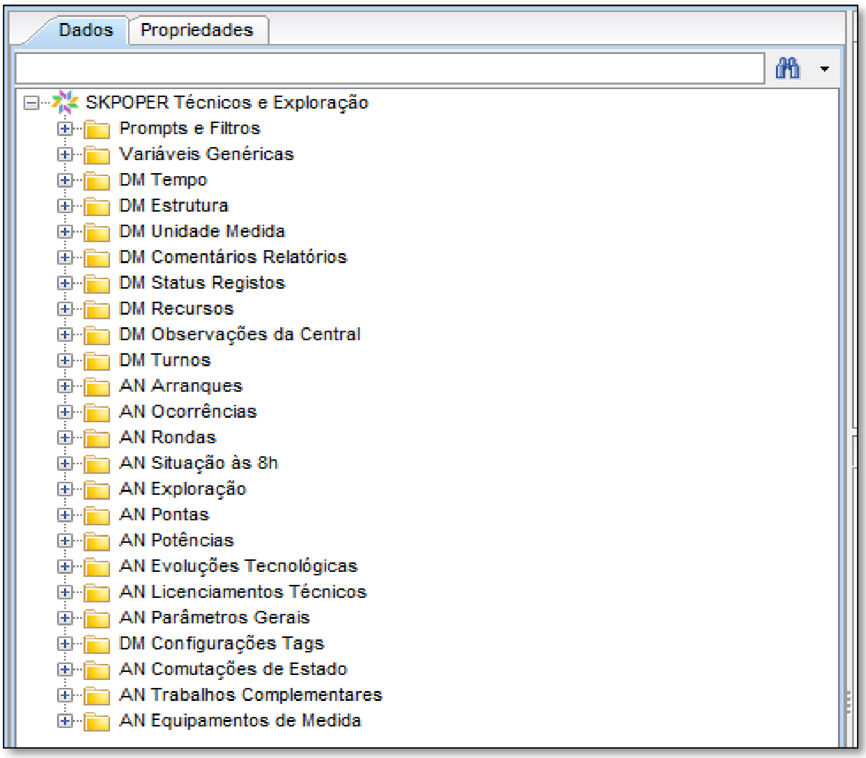
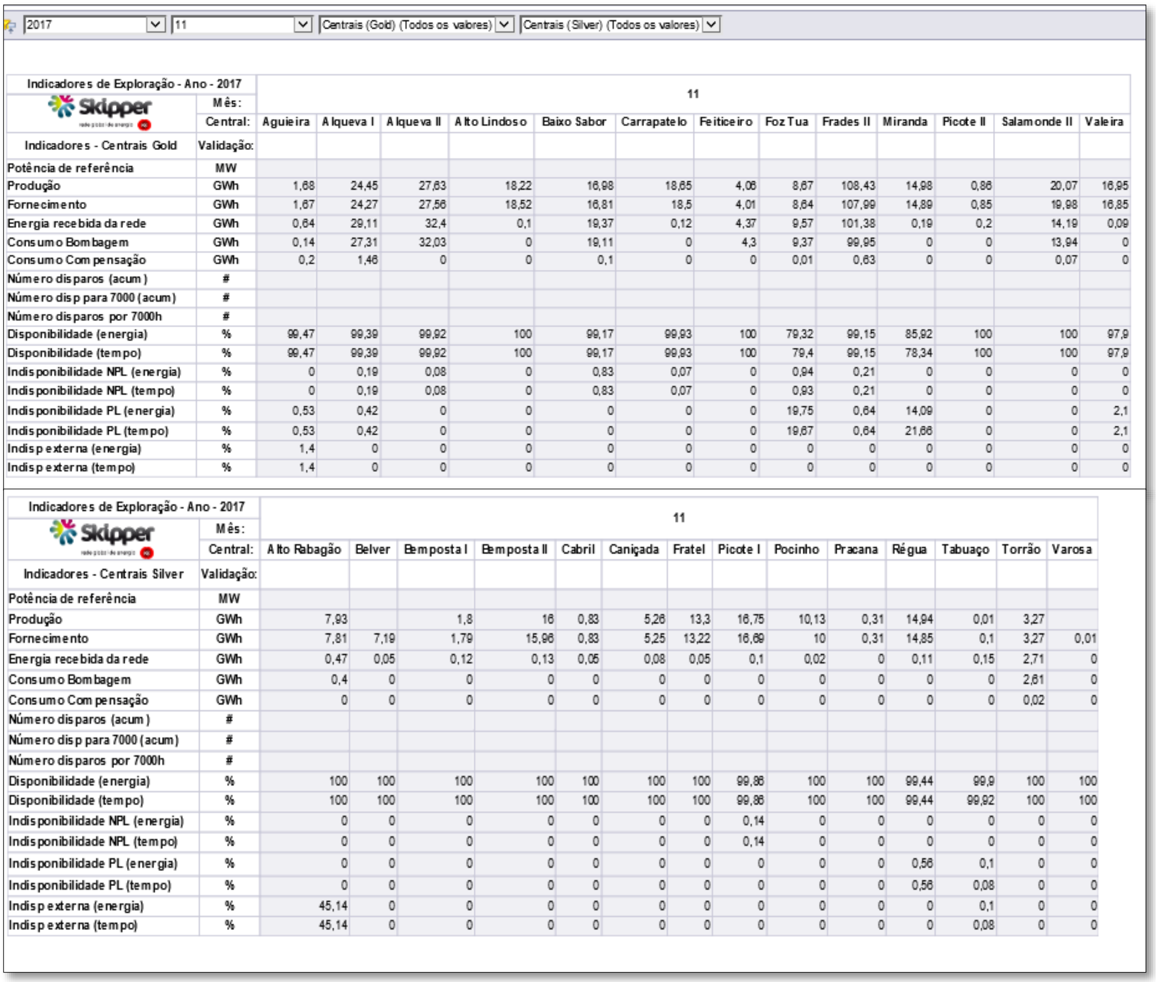


Figure 7 An example of a SAP BO Universe used in SKIPPER

SAP is being used in SKIPPER in two modes, the first one is a set of reports included in the Portal SKIPPER, these reports are accessible to SKIPPER clients and they contain tables of information regarding both types of power plants and their production and maintenance operations indicators. Figure 8 shows an example of a report made using SAP BO, and available at Portal SKIPPER for the clients. This report contains monthly values of exploration indicators, along with their measurements units, from all the hydraulic power plants (Centrais), during November in 2017, dividing them into two categories, Gold and Silver, depending on their income. The report offers, on the top, filtering options where the client can choose the desired year, month, and the type of power plants he wishes to explore. The second mode of SAP BO usage is data consulting, where the client/user can build a report with a table that contains data of his choice by performing a drag and drop action to the desired data, and apply the desired filters, the structure of the data implemented in SAP BO contains prefabricated filters ready to use, and at the same time the user can choose the types of filters that he wants to apply. The filtering options, included in the universe shown in Figure 7, are shown in Figure 9 and Figure 10. SAP is not limited to use only inside SKIPPER, different employees from different teams or departments have access to SAP but with different permissions depending on their usage profile.



2017 | 11 | Centrais (Gold) (Todos os valores) | Centrais (Silver) (Todos os valores)

Indicadores de Exploração - Ano - 2017

Mês: 11

Indicadores - Centrais Gold	Central:	Aguieira	Alqueva I	Alqueva II	Alto Lindoso	Baixo Sabor	Carrapatelo	Feiticeiro	Foz Tua	Frades II	Miranda	Picote II	Salamonde II	Valeira
	Validação:													
Potência de referência	MW													
Produção	GWh	1,68	24,45	27,63	18,22	16,98	18,65	4,06	8,67	108,43	14,98	0,86	20,07	16,95
Fornecimento	GWh	1,67	24,27	27,56	18,52	16,81	18,5	4,01	8,64	107,99	14,89	0,85	19,98	16,85
Energia recebida da rede	GWh	0,64	29,11	32,4	0,1	19,37	0,12	4,37	9,57	101,38	0,19	0,2	14,19	0,09
Consumo Bombagem	GWh	0,14	27,31	32,03	0	19,11	0	4,3	9,37	99,95	0	0	13,94	0
Consumo Compensação	GWh	0,2	1,46	0	0	0,1	0	0	0,01	0,63	0	0	0,07	0
Número disparos (acum)	#													
Número disp para 7000 (acum)	#													
Número disparos por 7000h	#													
Disponibilidade (energia)	%	99,47	99,39	99,92	100	99,17	99,93	100	79,32	99,15	85,92	100	100	97,9
Disponibilidade (tempo)	%	99,47	99,39	99,92	100	99,17	99,93	100	79,4	99,15	78,34	100	100	97,9
Indisponibilidade NPL (energia)	%	0	0,19	0,08	0	0,83	0,07	0	0,94	0,21	0	0	0	0
Indisponibilidade NPL (tempo)	%	0	0,19	0,08	0	0,83	0,07	0	0,93	0,21	0	0	0	0
Indisponibilidade PL (energia)	%	0,53	0,42	0	0	0	0	0	19,75	0,64	14,09	0	0	2,1
Indisponibilidade PL (tempo)	%	0,53	0,42	0	0	0	0	0	19,67	0,64	21,66	0	0	2,1
Indisp externa (energia)	%	1,4	0	0	0	0	0	0	0	0	0	0	0	0
Indisp externa (tempo)	%	1,4	0	0	0	0	0	0	0	0	0	0	0	0

Indicadores de Exploração - Ano - 2017

Mês: 11

Indicadores - Centrais Silver	Central:	Alto Rabagão	Belver	Bemposta I	Bemposta II	Cabril	Caniçada	Fratel	Picote I	Pocinho	Pracana	Régua	Tabuaço	Torrão	Varosa
	Validação:														
Potência de referência	MW														
Produção	GWh	7,93		1,8	16	0,83	5,26	13,3	16,75	10,13	0,31	14,94	0,01	3,27	
Fornecimento	GWh	7,81	7,19	1,79	15,96	0,83	5,25	13,22	16,69	10	0,31	14,85	0,1	3,27	0,01
Energia recebida da rede	GWh	0,47	0,05	0,12	0,13	0,05	0,08	0,05	0,1	0,02	0	0,11	0,15	2,71	0
Consumo Bombagem	GWh	0,4	0	0	0	0	0	0	0	0	0	0	0	2,61	0
Consumo Compensação	GWh	0	0	0	0	0	0	0	0	0	0	0	0	0,02	0
Número disparos (acum)	#														
Número disp para 7000 (acum)	#														
Número disparos por 7000h	#														
Disponibilidade (energia)	%	100	100	100	100	100	100	100	99,86	100	100	99,44	99,9	100	100
Disponibilidade (tempo)	%	100	100	100	100	100	100	100	99,86	100	100	99,44	99,92	100	100
Indisponibilidade NPL (energia)	%	0	0	0	0	0	0	0	0,14	0	0	0	0	0	0
Indisponibilidade NPL (tempo)	%	0	0	0	0	0	0	0	0,14	0	0	0	0	0	0
Indisponibilidade PL (energia)	%	0	0	0	0	0	0	0	0	0	0	0,56	0,1	0	0
Indisponibilidade PL (tempo)	%	0	0	0	0	0	0	0	0	0	0	0,56	0,08	0	0
Indisp externa (energia)	%	45,14	0	0	0	0	0	0	0	0	0	0	0,1	0	0
Indisp externa (tempo)	%	45,14	0	0	0	0	0	0	0	0	0	0	0,08	0	0

Figure 8 A SAP BO report for monthly indicators of hydraulic power plants

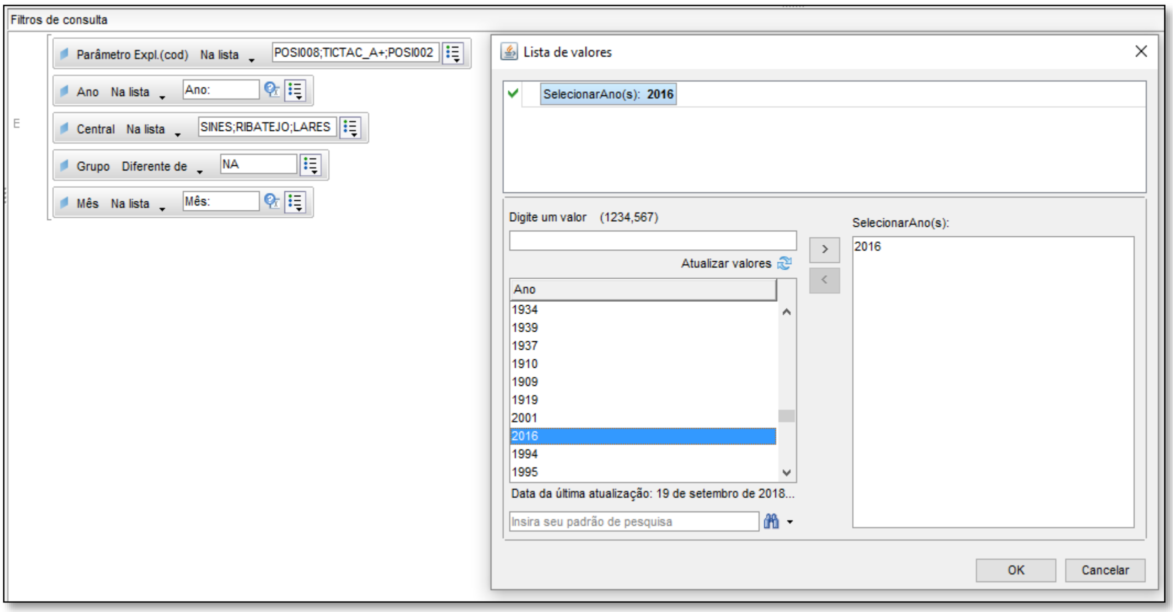


Figure 9 Applying filters from the choice of the client

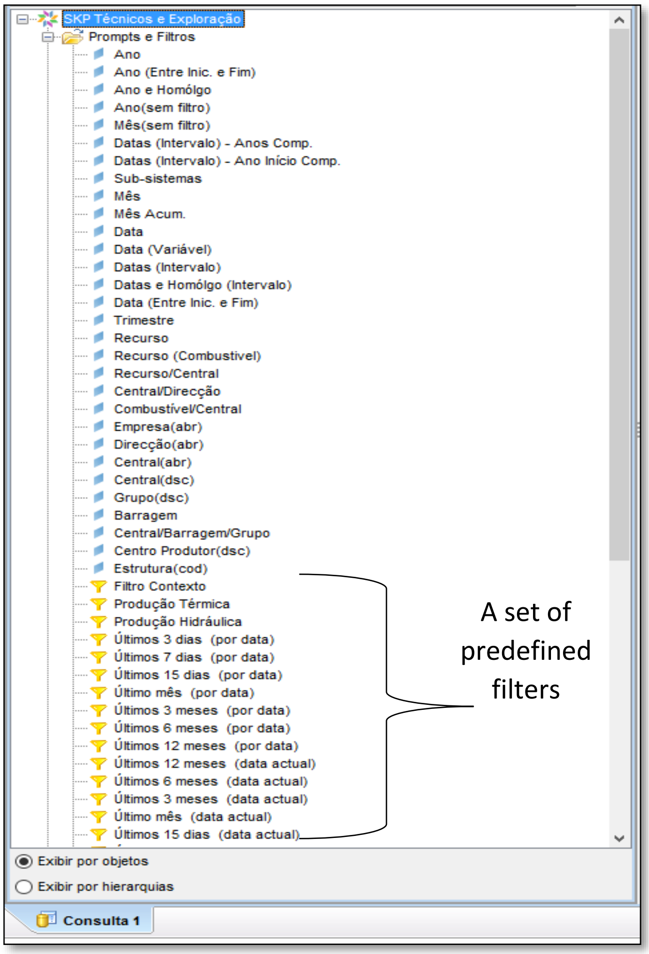


Figure 10 A set of prefabricated filters in universe shown in Figure.7

CHAPTER 3

Methodological Approach

3.1. Proposed Approach: from Better Reporting to Advanced Analytics

Over the last years, different energy forecasting models have been applied to solve different types of problems in a wide range of industries, each one with different needs data and available timescale. In this work, it is desired to develop a forecasting model that will enhance SKIPPER reporting process, focused on the electricity production in thermal power plants in Portugal, and integrate it with a new data advanced analytics solution.

The proposed methodology for this research is divided into two approaches, the first approach starts with implementing a new reporting framework in SKIPPER using SAS, which is a set of solutions that provides the possibility to perform data management, reporting, statistical and predictive analytics. The objective of this first approach is to compare the existing reporting framework (Power BI) with a novel report built on the new platform (SAS) and check whether the latter can satisfy the analytical requirements of SKIPPER, and eventually apply the necessary changes in the reporting process.

The second approach consists in integrating a forecasting model to complement the reporting process by applying two types of forecasting models, comparing them, defining their differences, and identifying the best uses for each one of them. Energy production is characterized by stability, and stability denotes an independence on weather conditions. Two years of historical data is available to predict future production levels. One of the available forecasting techniques proposed in this research is statistical time series modelling. The two types of forecasting models are tested using two platforms: a SAS forecasting tool in SAS Visual Data Explorer and R Studio. Furthermore, a regression model is tested using R Studio to define the important factors that affect thermal production.

Briefly, based on the Gartner Analytics Continuum shown in Figure 11and on EDP / SKIPPER current level of analytics maturity, the main objective of this research is to improve the descriptive and the diagnostic analytics by the first proposed approach and then moving to the predictive analytics using forecasting techniques in the second approach.

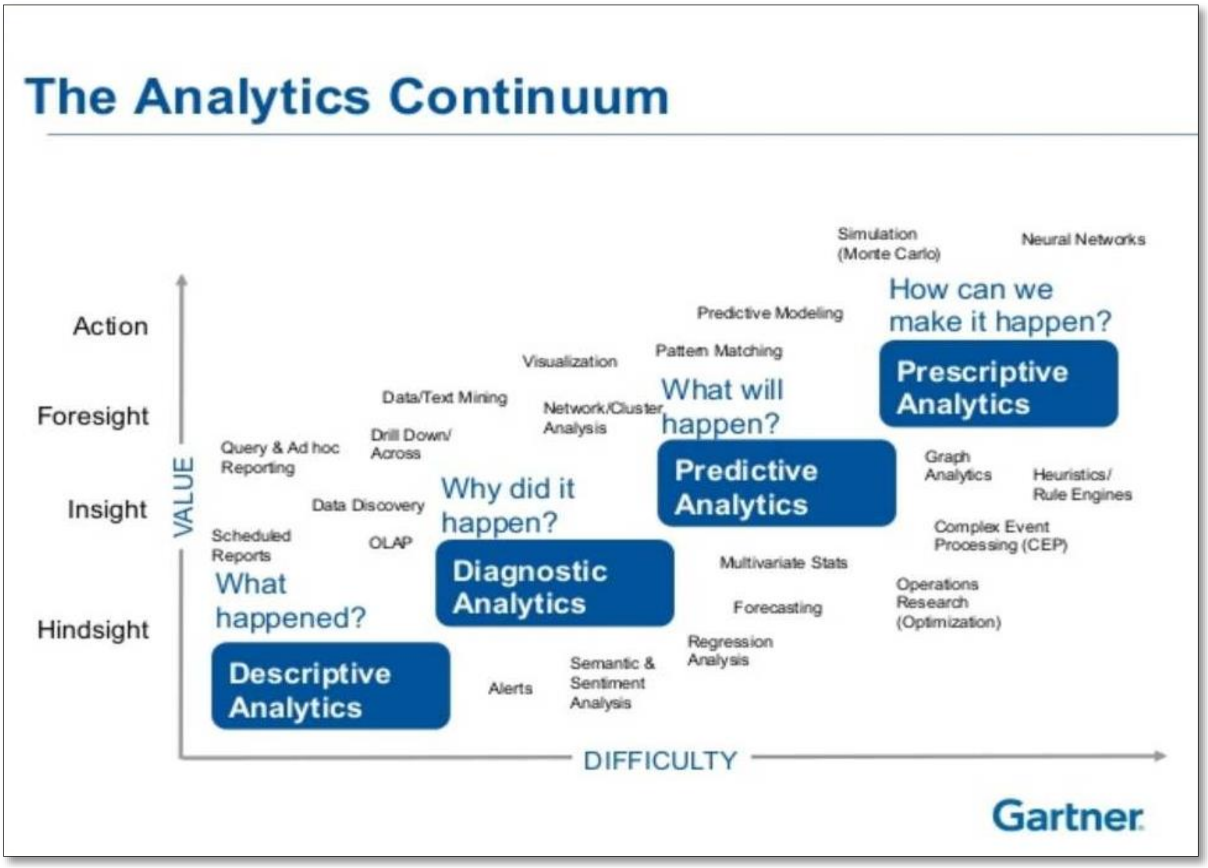


Figure 11 Gartner Anlaytics Continuum

3.2. Reporting Framework Improvement

To achieve the objective of this research, the first implemented step was participating in a process of improving the existing reporting process in SKIPPER. As mentioned before the main tools used for reporting are SAP Business Objects and Power BI. However, the DGU previously named “Direção de Sistemas de Informação (Information Systems Management) responsible for implementing new technologies in EDPP, has recently tested SAS as a new reporting tool, the tests had good results and DGU asked the EDP teams to start applying SAS in their internal work in the most suitable way based on their needs.

SAS is a software suite used for advanced data analysis, data management, and business intelligence. SAS has first appeared in 1966 when the USDA was collecting vast amounts of agricultural data but there was no computerized statistics program existed to analyse that data.

A consortium of eight universities came together under a grant from the National Institutes of Health (NIH) to solve that problem. The resulting program, the Statistical Analysis System, gave SAS both the basis for its name and its corporate beginnings. North Carolina State University, NCSU, became the leader in the consortium through two of the faculty members Jim Goodnight and Jim Barr who emerged as project leaders. Over the next few years, SAS software was licensed by pharmaceuticals, insurance companies and banks, as well as the academic community, and two more specialists joined the SAS project, Jane Helwig and John Sall. Together, the four members left the NCSU and started a SAS private company in March 1976, the company started to grow, products like SAS Base and SAS Graph started to be released, and SAS reached outside of the US [4] .

3.2.1. SAS in SKIPPER

SAS has been recently adopted by SKIPPER based on a decision made by DGU. The first efforts were focused on comparing SAS with the previous reporting tools and based on the obtained results, decide if this was the right choice. SKIPPER started integrating SAS into the reporting process through wo SAS products: SAS Enterprise Guide 7.1 and SAS Visual Analytics7.4. Figure 12 shows how these two products are integrated into EDP architecture.

- **SAS Enterprise Guide (EG)**

SAS EG has its first release in 1999 with the objective of creating an easy-to-use menu and wizard driven tool. Usually, SAS EG is connected to the data source that might be simple as an excel or CSV files or on a high level of complexity such as an Oracle database that exists on a server, which is the case for EDP. More in detail, SAS EG has a direct connection to SKIPPER's Oracle database. However, it is important to notice that in this case, SAS EG depends on having a data federation system, which means providing a virtual access to the sources and thus there is no business logic applied to the data at a source level. In SKIPPER, SAS EG is used to prepare data and this includes: querying required data, performing statistical calculations and aggregations, deriving new data items from existed ones, joining tables based on business needs, and finally uploading the processed data to the LASR[5] server or exporting data to external files.

[4] Historical information available at https://www.sas.com/en_us/company-information/profile.html
[5] The SAS LASR Analytic Server is an analytic platform that provides a secure, multi-user environment for concurrent access to data that is loaded into memory

In EDP, LASR server is composed of four repositories to present the base data store for SAS VA, and particularly, in SKIPPER, the processed data from SAS EG is uploaded to the EDPP LASR repository and then accessed from SAS VA. SAS EG could be also used for reporting as it provides drag and drop tools to build charts and dashboards, however, this is not implemented in SKIPPER.

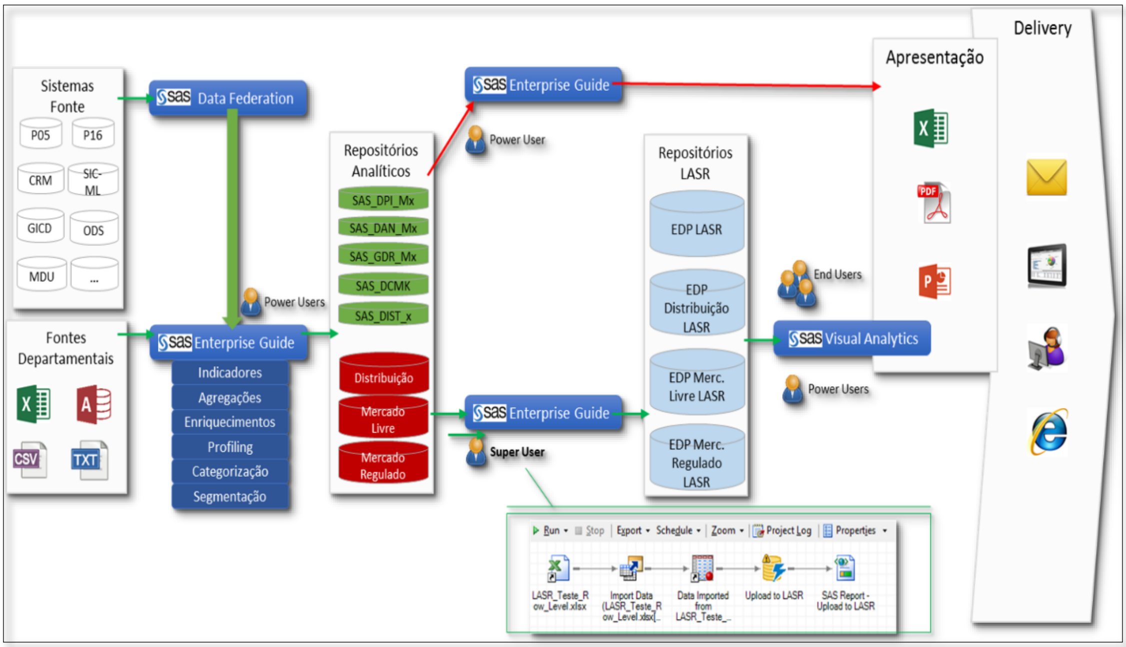


Figure 12 Structure EDP - SAS

- **SAS Visual Analytics (VA)**

SAS VA is a web-based environment that was first released in 2013 with the objective of providing graphical user interface that can be used to visually explore and report data, perform self-service analytics and present the results in the form of graphs and dashboards. SAS also provide the possibility of viewing the results on mobile devices. In SKIPPER, SAS VA is connected to the LASR server where the preprocessed data is coming from SAS EG, however, SAS VA is also used to perform part of the required calculations and aggregations to build the needed dashboards that meets the requirements of the client.

3.3. Forecasting Techniques

3.3.1. Forecasting Overview.

Forecasting is the process of determining the direction of the future based on past and present data and most commonly by analysis of trends. Forecasting means moving to the future, living in it and preparing for it in a way that delivers accurate forecasts and reduces the risks and uncertainty which are central to this matter, considering the dynamic nature of the external environment. In a general view, forecasting has been important in business for so many years and businesses have been using it to determine how to allocate their budgets or plan for anticipated expenses for an upcoming period, and this is typically based on the projected demand for goods and services.

However, the operating context of forecasting has changed over time. Forecasting started as an easy task when the production was on a small scale and the demand was stable; at that early stage, forecasting was built on the assumption that whatever happened last period would happen again this period in a high probability. Statistical demand forecasting was a big leap forward, as it started using time-series of previous data and building a few charts displaying aggregated values to create predictions about the future.

Forecasting evolved to the demand planning stage, where more hierarchical and casual factors were taken into consideration in the forecasting process. However, as the business complexity kept increasing, the planning era was dead and gone.

Today, forecasting is not the easy and simple task it used to be, it became more complex depending on more than just doing a statistical analysis of historical data. In the global and competitive business environments, the companies' need for accurate forecasting is becoming more important in terms of several success factors including on-time delivering and cost-effectiveness, these factors are linked directly to understanding the customer needs and demands, that is why bad forecasting can lead to serious consequences.

> "An example of forecasting going wrong are the alleged words of Ken Olsen, president of Digital Equipment Corporation (DEC) in 1977: "There is no reason for any individual to have a computer in their home". Subsequently, DEC entered the personal computer market too late and struggled until they were acquired in 1992." (EGNELL & HANSSON, 2013) .

Analysts and statisticians utilize various forecasting techniques depending on the situation that requires their purposes. Forecasting has been one of the major areas of interest for many academic researches and getting the forecasting into the current form was a long and tricky way.

Companies started with using corporate planning to organize their decision-making process and to plan long-term development. Traditionally a corporate plan utilized macroeconomic forecasts as basis and consisted of goals, objectives, forecasts of key economic trends, priorities for products and business areas and allocated capital expenditures (Hagerty, 2016). However, the changes in the business environment challenged the companies process of predicting demands and prices, and this challenge encouraged the managers to develop more specific, flexible and adaptable strategies that reduce ambiguity, add simplicity and simultaneously, are aware that the business environment may be too complex to allow strategies to develop all at once, therefore strategies must emerge in small steps as the company adapts and learns.

When thinking about forecasting, it is always important to be aware of few terms, while forecasting means exploring the future, backcasting is the process of exploring the past events given the information that is known to date, Forecasting Origin is the last point in the historical data, Forecasting Horizon is the distance between the forecast origin and the farthest point we are forecasting and finally a Step is the distance between the two adjacent observations.

3.3.2. Energy Forecasting

The importance of energy production management has been more vital in recent decades, considering that resources are getting less and, on a global level, there is a trend to use renewable and clean energy. Furthermore, forecasting energy production plays an important role in energy supply management for both government and private companies and therefore, using the appropriate models to accurately forecast the future energy production is an important issue for power production and distribution systems so they can be able to determine the possibility of meeting energy demands, and taking the right preventive procedures.

Recently, energy consumption had been increased due to population growth, besides the development in the industrial, economic and business environments. These changes have challenged companies to perform efficient production planning which requires accurate forecasts to cover the demand. Choosing the right technique and getting accurate forecasts helps the company to build a better understanding of the power generation supply, and thus make the right business decisions that will lead to increased profits, better utilization of the energy resources, and avoiding future problems.

3.3.3. Forecasting Techniques

Defining the forecasting type to be used depends on the organization's specific objective and the available data for the forecasting. An organization needs to develop a forecasting system that includes several approaches and expertise to handle forecasting problems and requirements, define the forecasting objective and the suitable forecasting type. Organizations can use short-term forecasting type when the objective is scheduling of production, personnel and transportation, while they can use medium-term forecasting when the objective is to determine future resource requirements and long-term forecasting when the objective is to set up strategic plans.

Furthermore, defining the appropriate forecasting method depends on knowing exactly what data is available, if there are no data available or if the available data is not relevant to the forecasts, then Qualitative Forecasting methods can be applied. Quantitative Forecasting, on the other hand, can be applied when the available data is numerical data about the past, and when it is likable that some of the past aspects will continue to appear in the future.

3.3.3.1. Qualitative Forecasting

Qualitative forecasting methods are the type of methods that combine intuitive judgment, opinions and subjective probability estimates, and these methods are very common in practice as organizations depend on them in situations when they have lake of historical data, when they are launching a new product in the market, when new market conditions are easy to predict, or even to forecast the effect of applying a new policy. Judgmental forecasting is also used as a solution when existing data isn't complete or only becomes obtainable with a delay such as forecasting the current level of economic activity in the central banks where some needed data is only available on a quarterly basis.

The accuracy of this type of forecasts increases when teams have a good knowledge about the forecasted area, the data available to the time is up-to-date as much as possible and the approaches implemented within an organization are well defined and systematic. During the last few years, organizations accepted the usage of judgmental forecasting as the solution when they are not able to use statistical methods, recognized its needs and its limitations, tried to confine these limitations by increasing the accuracy. When using a judgmental forecasting within an organization, it is important to be aware of the effect of personal or political agendas on the forecasts, for example, setting the sales forecasts to be low to show a good performance. Forecasts and judgments can be also affected by the optimism or wishful thinking of the forecaster, or even by prior familiar reference points, this is referred to as anchoring which might lead to brief and undervalue information and thereby create a systematic bias. To reduce these limitations and guarantee better judgmental forecasts, the forecasting team needs to follow some principles starting with setting up the forecasting task as clear and concise as possible, defining the challenges that might appear and implementing a systematic approach that includes a specification of important information that is relevant to the forecasting task. Organizations should also depend on the documentation of rules and assumptions implemented in the systematic approach, and justification of the forecasts made by the forecaster to increase accountability and reduce system bias.
Furthermore, organizations need to engage potential users in the forecasting process and clarify the basic assumptions that led to the forecasts, because this step will provide more assurance to the users and make them feel more connected to the process, and at the same time, some adjustments can be made on the forecasts according to a feedback from these users.

3.3.3.1.1. Delphi Method

The Delphi method, a structured forecasting technique, has been widely used for business forecasting. The Delphi method was invented by Olaf Helmer and Norman Dalkey of the Rand Corporation in the 1950s for the purpose of addressing a specific military problem (Hyndman & Athanasopoulos, 2018a).

The Delphi method is based on the idea that forecasts coming from a group are generally more accurate than those coming from individuals, usually, a group of experts are gathered, a facilitator is appointed, and a set of forecasting challenges/questionnaires are distributed to the experts. This is done for two rounds or more, wherein every round the experts return a set of forecasts and justifications and receive, afterwards, a feedback and they review their forecasts in accordance with the feedback, these rounds continue until a predefined sufficient level of consensus is reached and the final forecasts are built up by combining the experts' forecasts.

A group of challenges comes along with each one of the previous stages, and to guarantee the success of the method, several key characteristics need to exist during the process. The first challenge of the facilitator is selecting the experts, and the key feature here is that the experts are drawn from both inside and outside the organization, and it is recommended that the identities of these experts remain anonymous all the time the thing that reduces the influence of political, personal and social pressures, and gives every single expert an equal say. Furthermore, the experts don't need to meet at the same physical location which makes the process cost-effective and increases the possibility of gathering experts with diverse skills and expertise.

Another key feature is the feedback delivered to the experts in each round as it should involve summary statistics of the forecasts and outlines of qualitative justifications, the feedback is usually handled by the facilitator and it is his responsibility to direct the experts' attention to the areas that are most important or to the responses that fall outside their vision. Regarding the iteration of the process, many rounds should be avoided as experts are more likely to drop out if this number increases, the iterations are repeated until a satisfactory consensus level is reached. Satisfactory consensus does not mean complete convergence in the forecast value, it simply means that the variability of the responses has decreased to a satisfactory level (Hyndman & Athanasopoulos, 2018a). Finally, while building the final forecasts, the facilitator should pay attention and gives an equal weight to all experts' forecasts.

Overall, applying the Delphi method can be time-consuming, as it is almost impossible to construct final forecasts during a few hours, and taking a long time may cause the panel to lose interest and cohesion. However, organizations often use a variation of Delphi called "Estimate-Talk-Estimate" method, where even though forecasts submissions are anonymous, the experts are offered the possibility to interact between iterations. Even though this variation solved the problem of consuming time, it increases the possibility of the loudest person exerting too much influence.

3.3.3.2. Quantitative Forecasting

There is a wide range of used quantitative forecasting methods where often every method is developed within specified practices for specified objectives. There are no unified characteristics for all the methods, every method has its own properties, costs, and accuracies. Quantitative methods use mathematical analysis to identify patterns in historical data and use those patterns to forecast future ones.

Most quantitative prediction problems use either time series data (collected at regular intervals over time) or cross-sectional data (collected at a single point in time) (Hyndman & Athanasopoulos, 2018a).

3.3.3.2.1. Cross-Sectional Forecasting

The cross-sectional forecasting is the process of predicting a future value of a variable that has not been observed, based on information of other variables, called predictors, that have been observed, all at a point of time and regardless of the predictors' observations' ordering. It is necessary to understand the relationship between the variable to be forecasted and the predictors, so that it would be possible to estimate their effects on it. In this system scope, any change in the behaviour of the predictors will affect the output values in a predictable way, supposing that the relationship has no alteration. To gather data, cross-sectional designs generally use survey techniques, the survey presents a snapshot of the predictors' population at a certain point of time. These surveys may be repeated periodically and constructed in a panel where it would be possible to measure the change of the predictors over time.

3.3.3.2.2. Time Series Forecasting

While a time series is a sequence of data points indexed in time order and these points may or may not be observed at successive equally spaced points in time, the time series analysis covers a set of mathematical methods that analyse the time series data to extract meaningful statistics and other characteristics of the data. Time series forecasting is the process of predicting future values based on the previous observations and estimating how the sequence of these observations will continue into the future.

Considering the definition of the time series, an example of the time series could be a sequence of random variables, x_1, x_2, x_3, …, where the variable x1 denotes the value taken at the first-time point, the variable x_2 denotes the value taken at the second-time point, and so on. In general, a collection of random variables, $\{x_t\}$, indexed by t is referred to as a stochastic process (Shumway & Stoffer, 2017), which means that prediction of the future value will be with probability by finding the upper and lower limits as estimating the mathematical equation that determines the process is performed with an error value as it is always influenced by many factors and some of these factors are impossible to formulate mathematically.

For describing a time series data pattern, three terms may be used. The first is "trend" and it is used when a long-term increase or decrease in the data exists. The term "seasonal" is used when the time series data is affected by seasonal factors such as a specific day in the week or a specific month of the year, the seasonal pattern is always fixed and has a known frequency. The time series is "cyclic" when the data shows rises and falls and these changes are not of a fixed frequency, usually these variations occur as a resonance to economic conditions. It is very common for a time series to contain the three patterns together.

- **Autocorrelation Function**

An autocorrelation function (ACF) is the function that measures the linear relationship between lagged values of a time series at different times. In other words, an ACF illustrates how correlated two points with each other, based on the number of time steps they are separated by. Using ACF helps to determine how past and future data point are related in a time series.

For example, assuming r_1 is the measurement of the relationship between y_t and y_{t-1}, and r_2 is the is the measurement of the relationship between y_t and y_{t-2}, and so on, then the value of r_k can be written as the following:

$$rk = \frac{\sum_{t=k+1}^{T}(yt - \overline{y})(yt - k - y)}{\sum_{t=1}^{T}(y_t - \bar{y})2}$$

where T is the length of the time series [15].

When a time series shows values that are not related to each other with no autocorrelation, it is called a white noise. For a white noise, it is expected to have an autocorrelation close to zero, yet, in practice, the autocorrelations are never equal to zero due to the existence of random variations.

- **Stationary and Non-Stationary Process**

The stochastic nature of a process indicates that the observations that are close to each other in the current time, will be more closely related than observation further apart. Furthermore, a stochastic process can be either stationary or nonstationary, the former indicates that the static properties of the process such as the mean and variance don't change over time, and the later indicates the existence of changing in the properties. Thus, time series with trends or seasonality data patterns are not stationary, as the trend and the seasonality affect the data values of the time series at different times. Yet, a time series with cyclic data pattern, without any trends or seasonality, is considered stationary and this is because the various frequency of the cycles.

While stationary processes are much easier to analyse, in practice, the nonstationary time series processes exist much more. However, exists a way to transform a non-stationary time series into a stationary time series by computing the differences between sequential observations, this process is known as Differencing and it helps stabilize the variance of a time series.

The first step in any application of time series analysis method involves an accurate study of the gathered time series data over time and a definition of the data pattern as this will help to choose the analysis method as well as the statistics that will be the best to use to extract information from the data. The variables values of the time series are then plotted and connected at adjacent time periods as these values are usually produced as a discrete sample. Adjusting time periods or defining a finite number of points in time using a distribution function to specify a continuous time series does not mean that the selection of sampling rate is not extremely important as choosing an insufficient sampling rate could affect the appearance of data.

- **Time Series Models**

Models of time series data can be generated by different stochastic models to present different stochastic processes, and when it comes to practical importance, three classes of these models appear: autoregressive (AR) models, integrated (I) models, and moving average (MA) models. These three classes depend linearly on previous data points, and they can be combined to produce autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models.

3.3.3.2.3. Time Series Regression Models

It is important to be aware that the time series methods only use information from the past observations of a series to be forecast, without exploring any other factors that might affect its behaviour. However, exist time series models that are based on the concept of forecasting a time series *y* of interest assuming that it has a relationship with other time series ***x*.** These models are known as the Time Series Regression Models, where *y* is the forecast variable and *x* the predictor variable.

Understanding the regression models starts with understanding the very basic case where only one variable exists and it is needed to predict the next value of the variable depending on its current and past values. An example that illustrates this case would be forecasting the amount of the next income salary for an employee who has a job that is paid depending on the number of customers he serves and the amount of money he makes per week out of the deals he closes, while the only available data is the last six weeks income without any other details.

In this case, the week number is only a descriptor, and usually predicting the value of the next income salary would be done by calculating the mean value of the previous incomes, and setting it up as the value of the next income value. Visualizing the data points in a graph with the mean value will show the variability in the income values, however, the mean is considered as the best prediction for the next week income value and any other given week.

Measuring the distance of the data points from the mean line would give an idea of how good the best prediction line fits these data points; these distances are called Residuals or Errors as they show how far the observed points from the mean line. When performing statistics, it is usual to square the residuals to make all the values positive, as some of the distance values are negative ones, and emphasize larger deviations, adding the squared values presents a value called the Sum of Squared Errors (SSE).

- **Simple Linear Regression**

Moving forward to the linear regression where the simplest form in the regression models implies a linear relationship between the forecast variable y_t, for $t = 1, \ldots, n$, and the predictor series $x_{t1}, x_{t2}, \ldots, x_{tn}$, where it is possible to say that the forecast variable is a function of the predictor variable. As in the basic case discussed earlier, several data points will be observed and plotted in a graph, scattered around the linear prediction line, however, the goal of simple linear regression is to create a linear model that minimizes the value of SSE, which means minimizing the error values. Linear predictors estimated by minimizing the mean square prediction error are called best linear predictors (BLPs).

The relationship between the dependent/forecast variable y and the independent/predictor variable x can be expressed as a slope-intercept form of a line, as the following:

$$y_t=\beta_0+\beta_1 x_t+\varepsilon_t.$$

Where the coefficient β_0 denotes the intercept, β_1 denotes the slope, and ε denotes the error term or the unexplained variation in y that can be caused by any other factors than x.

Updating the previous example of an employee weekly income to fit in the concept of linear regression would be by getting new data, such as the number of customers the employee serves per week, plotting both variables to detect the relationship type based on the slope of the fitted line, study the changes of the income based on the number of customers and predict the next income.

When performing simple linear regression model, it might be important to compare it with the previous model where the independent variable didn't exist to be able to determine how important the independent variable is and how much its existence helped to explain the dependent variable.

- **Multiple Linear Regression**

There are cases in which there is more than one independent variable that is affecting the variable of interest, these models are called multiple linear regression models and they are expressed as the following:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t,$$

Where x1, x2…. xk are the predictor/independent variables and y is the variable to predict. The coefficients, β1, β2. βk, measure the marginal effects of the predictor variables (Hyndman & Athanasopoulos, 2018a), which means that every one of them presents the effect of one predictor taking into consideration the effect of other predictor variables.

Expanding the previous example of predicting weekly income of an employee involves using more than just number of customers, and including data such as the amount of money the employee makes out of closing deals and the number of working hours per week. Building a multiple linear regression model will produce more accurate forecasts and give a clearer idea about the strength of the relationship between the variables.

3.3.3.2.4. Time Series Models

3.3.3.2.4.1. Backshift Notation

Using B, the backshift operator, before either a value of the series x_t or an error term w_t means shifting the data back one period, for example:

$$Bx_t = x_{t-1}$$

The "power" of B indicates the number of the backshifts that need to be applied to the data, as an example,

$$B^2x_t=x_{t-2}$$

Where x_{t-2} represents x_t with two periods back in time.

It is important to know that the backshift operator B doesn't operate on coefficients as they are fixed values that don't change with time.

3.3.3.2.4.2. Autoregressive Models (AR)

In an autoregressive model, forecasting the current value of the variable of interest depends linearly on its own previous values. It can be explained as a function of p past values, x_{t-1}, x_{t-2}, ..., x_{t-p}, where *p* indicates the number of immediately preceding values in the series that are used to predict the value at the current time. The *p* number is referred to as the order of the autoregressive, and an autoregressive model of order *p* is abbreviated as AR (*p*), thus, writing AR (2), for example, means that the prediction is based on the last two lagged variables.

The expression "autoregressive" indicates that the regression process is a regression of the variable against itself.

An autoregressive model, AR(*p*), can be expressed in the same way as a multiple regression model but with having the lagged values of x as the predictors, like the following:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t$$

Where xt is stationary, wt is a white Gaussian noise, and $\varphi_1, \varphi_2, \ldots, \varphi_3$ are constants that don't equal 0.

The same equation can be redefined when the mean, μ, of xt is not zero, in the following form:

$$xt - \mu = \phi_1(x_{t-1} - \mu) + \phi_2(x_{t-2} - \mu) + \cdots + \phi_p (x_{t-p} - \mu) + w_t$$

or can be defined as

$$xt = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdot \cdot \cdot + \phi_p x_{t-p} + w_t$$

with defining $\alpha = \mu (1 - \phi_1 - \cdot \cdot \cdot - \phi_p)$.

Using the backshift operator produces a useful form to write an AR(p), as

$$(1 - \phi_1 B - \phi_2 B^2 - \cdot \cdot \cdot - \phi_p B^p)\, x_t = w_t,$$

and in a more briefly form:

$$\phi(B)x_t = w_t.$$

The difference between autoregression and regression models lies in the fact that the regressors, x_{t-1}, …, x_{t-p} are random components, where the values of z_t, in the regression model, are assumed to be fixed, and this might develop technical difficulties.

Figure 13 shows two time series from AR (1), left, and AR (2), right, models where:

$y_t=18 - 0.8y_{t-1} + \varepsilon_t$, for AR (1), and

$y_t=8 + 1.3y_{t-1} - 0.7y_{t-2}+ \varepsilon_t$ for AR (2).

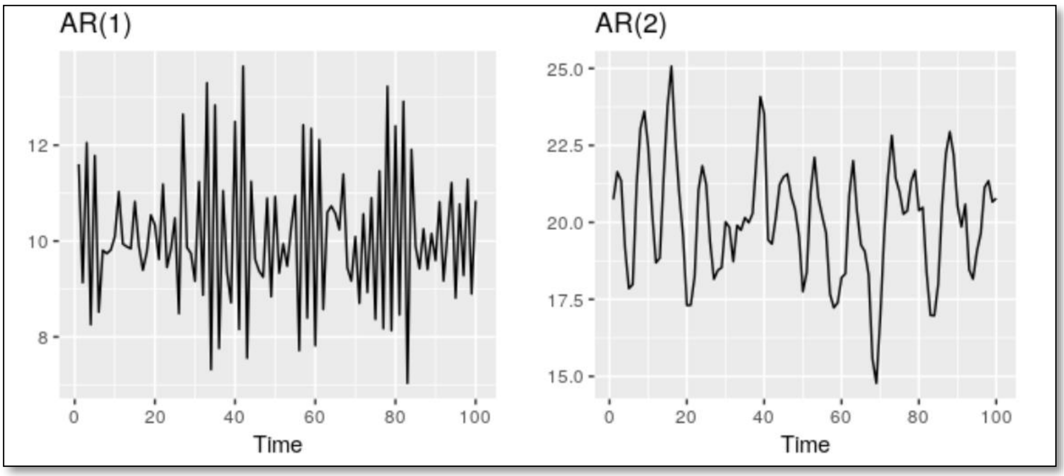


Figure 13 Time series from an AR (1) model and an AR (2) model

Image available at: https://otexts.org/fpp2/AR.html

In both cases, the error is a white noise that is normally distributed, and the variance of the error term εt will only change the scale of the series (Hyndman & Athanasopoulos, 2018a), while changing the parameters $\phi_1, \ldots, \phi_p$ will produce different patterns of the time series.

Usually, autoregression models are restricted to stationary data and this implies requiring some constraints on the parameters' values, for example:

$-1 < \phi_1 < 1$ in an AR (1) model, and

$-1 < \phi_2 < 1, \phi_1 + \phi_2 < 1, \phi_2 - \phi_1 < 1$ in an AR (2) model.

When p ≥ 3, the constraints are much more complicated, and usually the statistics software take care of them when estimating a model.

3.3.3.2.4.3. Moving Average Models (MA)

Instead of using the previously observed values of the variable to forecast in a regression, a moving average model depends on combining linearly the past forecast errors in a regression-like model to form the observed data.

A moving average model of order q, abbreviated as MA(q), is defined as the following:

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdot \cdot \cdot + \theta_q w_{t-q},$$

Where w_t is a white noise with mean equals to zero and a variance of σ2w, and $\theta_1, \theta_2, \ldots, \theta_q$ are parameters ($\theta_q \neq 0$). Using the backshift operator, the moving average can be written as the following:

$$x_t = \theta(B)w_t.$$

Unlike the autoregressive process, the moving average process is stationary for any values of the parameters $\theta_1, \ldots, \theta_q$.

Figure 14 shows two examples of moving average models with different parameters, where:

$y_t=20+\varepsilon_t+0.8\varepsilon_{t-1}$ for an MA (1), and

$y_t=\varepsilon_t-\varepsilon_{t-1}+0.8\varepsilon_{t-2}$ for an MA (1).

In both models, ε_t is a normally distributed white noise with mean zero and variance one.

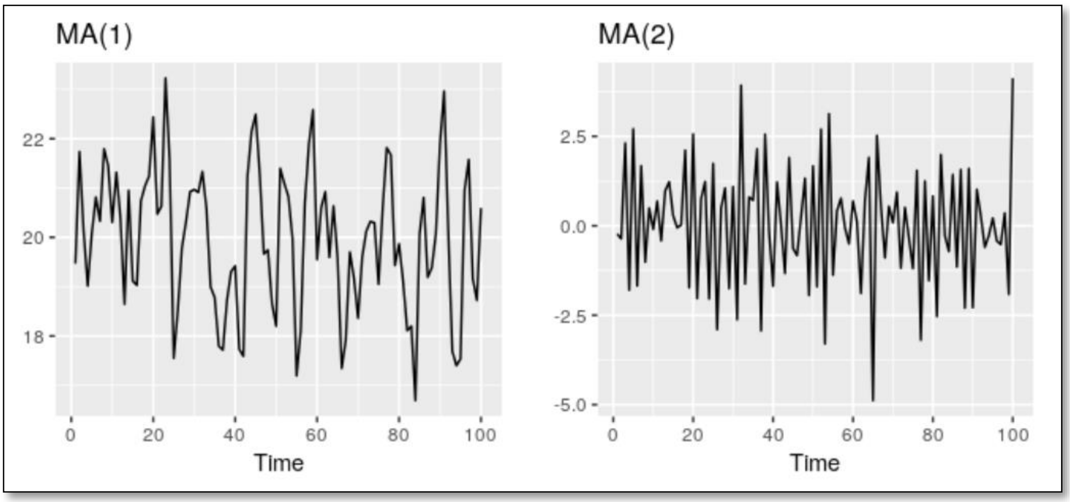


Figure 14 Time series from an MA (1) model and an MA (2) model.

Image available at: *https://otexts.org/fpp2/MA.html*

- **MA models Invertibility**

It is possible to express any stationary AR(p) model as an MA (∞) model (Hyndman & Athanasopoulos, 2018a). For example, for an AR (1) model, it is possible to write, using repeated substitution, the following equations:

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \varepsilon_t \\ &= \phi_1(\phi_1 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \phi_1^2 y_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\ &= \phi_1^3 y_{t-3} + \phi_1^2 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\ &\text{etc.} \end{aligned}$$

On the condition that -1 < ϕ1 < 1, the value of ϕk1 will get smaller as the value of k gets larger. So finally, the following form is acquired:

$$y_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \phi_1^3 \varepsilon_{t-3} + \cdots,$$

as an MA (∞) process.

The reverse process, writing an MA(q) model as an AR(p) model, can be held when applying some constraints to the MA model parameters, and in this case, the MA model is called an invertible model. To convert an MA invertible model to an AR model, a group of mathematical properties are required.

For example, considering an MA (1) model such as $y_t=\varepsilon_t+\theta_1\varepsilon_{t-1}$, in its AR ($\infty$) representation, the most recent error can be written as a linear function of current and past observations:

$$\varepsilon_t = \sum_{j=0}^{\infty}(-\theta)^j y_{t-j}.$$

The previous equation implies that when $|\theta| > 1$ the weight increases since the lags are increasing, and this means that most distant observations have a greater influence on the current error value. it also implies that when $|\theta| = 0$ the weights have the same size and all the observations, distant or recent, have the same influence. However, since the two previous cases don't make sense, it is required, in an invertible process, to have $|\theta| < 1$, the thing that makes the influence of the most recent observations higher than the influence of the distant ones.

The invertibility constraints for other models are similar to the stationarity constraints,

$-1 < \phi_1 < 1$ in an AR (1) model, and
$-1 < \phi_2 < 1$, $\phi_1 + \phi_2 < 1$, $\phi_2 - \phi_1 < 1$ in an AR (2) model.

3.3.3.2.4.4. ARMA

By combining the properties of AR and MA processes a newly mixed autoregressive-moving average model (ARMA), for stationary time series, is acquired. In a time-series, the ARMA model is used to understand the series and predict future values in it, by regressing its lagged variables and modelling the error terms. In practice, the ARMA model is more used than AR and MA models.

ARMA model is noted by two orders, *p* and *q*, as ARMA (*p*, *q*), where *p* is the autoregressive order, and *q* is the moving average order, and an ARMA model of a stationary time series can be expressed as the following:

$$x_t = \varphi_1 x_{t-1} + \cdot\cdot\cdot + \varphi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdot\cdot\cdot + \theta_q w_{t-q},$$

Clearly, when $q = 0$, the model turns into an autoregressive model of order p, and when $p = 0$, the model turns into a moving average of order q. Using the operators of AR and MA, the previous quotation is written as:

$$\varphi(B)x_t = \theta(B)w_t.$$

3.3.3.2.4.5. Integrated Models (I)

In many situations, time series can be thought of as being composed of two components, a nonstationary trend component and a zero-mean stationary component. This means that if x_t is a random walk where $x_t = x_{t-1} + w_t$, then by performing differencing on x_t, it can be said that $\nabla x_t = w_t$ is stationary. An example, from (Shumway & Stoffer, 2017), presents the model

$$x_t = \mu_t + y_t,$$

considering μ_t as stochastic and changing slowly based on a random walk, and can be written as

$$\mu_t = \mu_{t-1} + v_t$$

$$\nabla x_t = v_t + \nabla y_t$$

As a definition, an integration of order d, I(d), of a time series is a summary statistic, which determines the minimum number of differences needed to acquire a stationary series out of a non-stationary one.

3.3.3.2.4.6. ARIMA

The integrated ARMA, or ARIMA, model is a broadening of the class of ARMA models to include differencing (Shumway & Stoffer, 2017).In the same way of ARMA, ARIMA is a combination of Autoregressive and Moving Average models, and both of ARMA and ARIMA are fitted to time series data either to understand it or to forecast future values of the series.

However, there are cases when ARIMA is applied on a data that shows non-stationary behaviour, and this is where differencing, or integration, can be applied, more than ones, to remove the non-stationary.

These models are denoted as ARIMA (p, d, q), where p is the autoregressive order, d is the degree of differencing, and q is the moving average order. The full model can be written as

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

where y'_t is the differenced time series, and the predictors on the right side include the lagged values of y_t and lagged error values.

Many of the models that have already been discussed are special cases of the ARIMA model, as shown in

Table 1, noting that random walk and random walk with drift are two simple forecasting methods that assume the forecast values to be equal to the last obtained observation.

White noise	ARIMA (0,0,0)
Random walk	ARIMA (0,1,0) with no constant
Random walk with drift	ARIMA (0,1,0) with a constant
Autoregression	ARIMA (p,0,0)
Moving average	ARIMA (0,0, q)

Table 1 Special cases of ARIMA models

Table available at: https://otexts.org/fpp2/non-seasonal-arima.html

Identifying the appropriate orders *(p, d, q)* can be difficult, and usually the functions in the statistics software do it automatically. However, it is sometimes possible to use the ACF plot, and the closely related PACF plot, to determine appropriate values for p and q (Hyndman & Athanasopoulos, 2018a).

- **Forecasting with ARIMA**

Forecasting process with ARIMA can be finalized in the following steps:

a. Stationarity Check, where ARIMA models require the time series to have a constant mean and variance over time.
b. Differencing, which is the method of transforming a non-stationary time series into a stationary one, and the method is repeated more than once until the time series is stationarised. The best way to check if the series is differenced enough is to plot it and test to see if a constant mean and variance exist.
c. Selecting AR and MA terms, usually in ARIMA models, one of AR or MA is only used, and in order to define which one it is possible to use the ACF and the PACF plots, whereas AR terms are used and the data may follow an ARIMA (p, d ,0) model if:
   - The ACF is exponentially decaying or sinusoidal;
   - There is a significant spike at lag p in the PACF, but none beyond lag p (Hyndman & Athanasopoulos, 2018a).

   And MA terms are used and data may follow an ARIMA (0, d, q) model if:
   - The PACF is exponentially decaying or sinusoidal;
   - There is a significant spike at lag q in the ACF, but none beyond lag q (Hyndman & Athanasopoulos, 2018a).
d. Building the model and defining the number of periods to forecast depending on the needs.
e. Validating the model and comparing the predicted values to the actual ones in the validation sample.

3.3.3.2.5. Hierarchical Time Series

A hierarchical time series can be defined as “a collection of several time series that are linked together in a hierarchical structure”(Hyndman, 2013). Time series are, often, of a disaggregated nature where exists various attributes of interest, and each one of these attributes can be disaggregated into finer categories.

Since these finer categories are nested within a bigger group of categories, then the combining their time series will follow a hierarchical aggregation structure.

Figure 15. shows a two-level hierarchical structure, where the top level, with index k = 0, represents the most aggregated level of the hierarchy, the Total. In level k = 1, the Total is disaggregated into two time series, which are in turn are disaggregated into more two and three time series at the bottom level, k = 2. In this hierarchy, the total number of time series is n = 8.

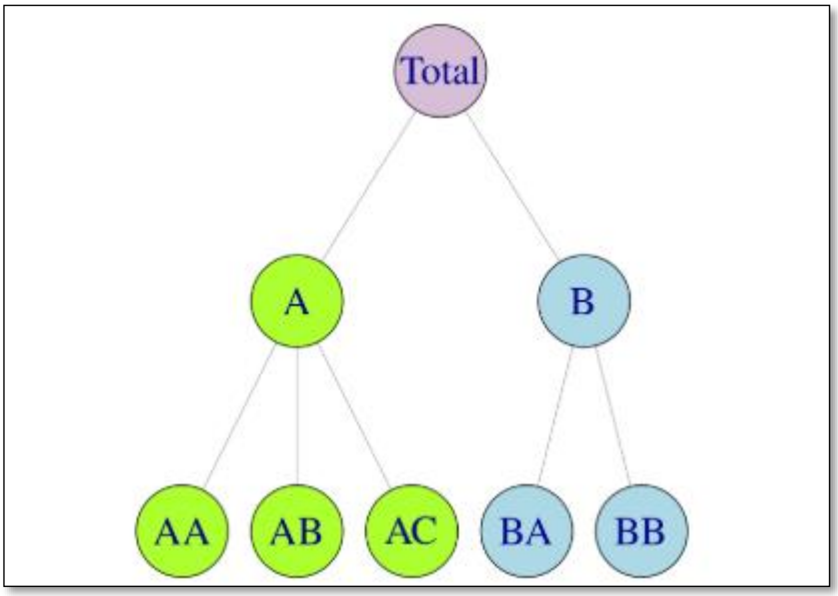


Figure 15 A two level hierarchical tree diagram

Photo available at https://otexts.org/fpp2/hts.html

At a given time, t, in order to capture the observation at the top level, the following can equations can be used:

$$y_t = y_{\mathrm{AA},t} + y_{\mathrm{AB},t} + y_{\mathrm{AC},t} + y_{\mathrm{BA},t} + y_{\mathrm{BB},t}$$

Where the observations at the bottom level will sum up to form the observations of the series in the level above, as the following:

$$y_{\mathrm{A},t} = y_{\mathrm{AA},t} + y_{\mathrm{AB},t} + y_{\mathrm{AC},t} \quad \text{and} \quad y_{\mathrm{B},t} = y_{\mathrm{BA},t} + y_{\mathrm{BB},t}.$$

And consequently:

$$y_t = y_{\mathrm{A},t} + y_{\mathrm{B},t}.$$

Advances in data collection and storage have resulted in large numbers of time series that are hierarchical in structure, and clusters of which may be correlated (Hyndman , Athanasopoulos, & Shang, hts: An R Package for Forecasting Hierarchical or Grouped Time Series).

- **Forecasting Hierarchical Time Series**

Hierarchical forecasting methods can follow several approaches which usually involve a top-down method, a bottom-up method, or even a method that combine both methods and which is referred to as middle-out method. "Hierarchical forecasting methods allow the forecasts at each level to be summed giving the forecasts at the level above" (Hyndman, Athanasopoulos, & Shang, n.d.). The interest is generating forecasts for each series in the hierarchy.

- **Bottom-up Method**

Process of generating forecasts, following this approach, starts from the bottom-level of each series in the hierarchy, summing them up to produce forecasts for all the series on the structure.

For example, in the hierarchy in figure 20, the forecasting starts by summing up the forecasts of the bottom-level series to get coherent forecasts for the rest of the series, as the following:

$$\begin{aligned} \tilde{y}_h &= \hat{y}_{\mathrm{AA},h} + \hat{y}_{\mathrm{AB},h} + \hat{y}_{\mathrm{AC},h} + \hat{y}_{\mathrm{BA},h} + \hat{y}_{\mathrm{BB},h}, \\ \tilde{y}_{\mathrm{A},h} &= \hat{y}_{\mathrm{AA},h} + \hat{y}_{\mathrm{AB},h} + \hat{y}_{\mathrm{AC},h}, \\ \text{and} \quad \tilde{y}_{\mathrm{B},h} &= \hat{y}_{\mathrm{BA},h} + \hat{y}_{\mathrm{BB},h}. \end{aligned}$$

> "An advantage of this approach is that we are forecasting at the bottom-level of a structure, and therefore no information is lost due to aggregation. On the other hand, bottom-level data can be quite noisy and more challenging to model and forecast." (Hyndman & Athanasopoulos, 2018b)

- **Top-down Method**

On the contrary of the bottom-up approach, the first generated forecast in the hierarchy is for the Total series on the top and then disaggregating them down the hierarchy.

According to (Hyndman & Athanasopoulos, 2018b), and assuming that $p1, \ldots p_m$ as a set of disaggregation proportions that define how the forecasts of the top level has to be distributed in order to get forecasts of time series at the bottom-level of the hierarchy. Following the previous example, in the hierarchy of figure.15, using the disaggregation proportions, the bottom-level forecasts can be:

$$\tilde{y}_{\mathrm{AA},t} = p_1\hat{y}_t, \quad \tilde{y}_{\mathrm{AB},t} = p_2\hat{y}_t, \quad \tilde{y}_{\mathrm{AC},t} = p_3\hat{y}_t, \quad \tilde{y}_{\mathrm{BA},t} = p_4\hat{y}_t \quad \text{and} \quad \tilde{y}_{\mathrm{BB},t} = p_5\hat{y}_t.$$

The generated bottom-level forecasts can be then used to generate the rest of the series of the hierarchy. Defining the disaggregation proportions in the top-down approaches is usually based on the historical proportions of the data, and the two most common top-down approaches, Average Historical Proportions and Proportions of the Historical Averages, relay on the historical proportions of the data (Hyndman & Athanasopoulos, 2018b).

“In general, these approaches seem to produce quite reliable forecasts for the aggregate levels and they are useful with low count data. On the other hand, one disadvantage is the loss of information due to aggregation” (Hyndman & Athanasopoulos, 2018b).

- **Middle-out Approach**

This approach combines the previous two approaches where the first forecasts to generate are the forecasts of the middle-level as the base to generate the forecasts for the upper and lower levels. Regarding the upper time series, the forecasts are generated following the bottom-up approach which means aggregating the middle-level upwards, while the forecasts of the lower levels are generated following the top-down approach which means disaggregating the middle-level forecasts downwards.

- **The Optimal Reconciliation Approach**

The optimal combination approach performs forecasting process in a different way from the previous approaches, where it starts by handling every series in the hierarchy independently and generating base forecasts without any aggregation at this stage. The independent base forecasts are then combined to generate a set of revised forecasts that are aggregated based on the hierarchal structure of the time series.

> "Unlike any other existing method, this approach uses all the information available within a hierarchy. It allows for correlations and interactions between series at each level of the hierarchy, it accounts for ad hoc adjustments of forecasts at any level, and, provided the base forecasts are unbiased, it produces unbiased revised forecasts." (Hyndman et al., n.d.)

CHAPTER 4

Data Description and Forecasting Experiments

4.1.The Case Study – EDPP – Team SKIPPER

EDP is a global manufacturer of energy production that supplies energy for both regulated and not regulated market. SKIPPER, the target case study of this research, is part of the DEC team in EDPP and, as explained before, SKIPPER is responsible for providing all types of information (including the thermal and hydraulic electricity production indicators) to other departments and teams at EDP Group. However, it is not SKIPPER's responsibility to decide the amount of electricity to be produced or delivered to the market.

The task of defining the target energy to produce, monthly or yearly, is done by UNGE, a department in EDPP, which is responsible for planning, managing and controlling sales in the energy market with the final goal of optimizing all decisions related to energy generation in the EDP Group. In the process of managing energy generation in the thermal production sector and defining selling and buying prices, UNGE defines a strategy based on monetarizing electricity market and some external factors such as the prices of needed gas and coal, besides the amount of money that EDP needs to pay in exchange of CO2 emitted during electricity generation operation.

This work focuses on two main goals: (1) improve the reporting process in SKIPPER using a new platform; and (2) perform production forecasting experiments where part of these experiments is based on the electricity production targets defined by UNGE, internal exploration indicators and previously mentioned external factors.

4.2.Data Collection

Part of the dataset used for forecasting in this research is extracted from the Oracle database in SKIPPER through SAS EG by performing a query that returns three attributes: timestamp, power plant or a group of the power plant and energy production. The dataset's time frame starts in July 2016 until October 2018, and it includes a monthly indicator of electricity production in the thermal sector for the three thermal power plants: Lares, Ribatejo, and Sines along with the groups of every power plant.

The rest of the dataset is extracted from the database of UNGE by performing a query that returns values of external factors related to thermal energy production in the same previous time frame. These factors were chosen based on recommendations from SKIPPER. The data contains missing values of production in some points, however, it is important to note that missing value, in this case, is not equal to zero, as zero means simply that the data was recorded, but the power plant or the group was not producing energy. Dealing with the missing values is handled differently based on the used framework to perform the forecasts.

4.3.Step 1: Reports Improvement

4.3.1. SAS vs. Power BI

The first implemented step in SKIPPER using SAS was creating the same thermal power plants dashboard that was built in Power BI, introducing it to the client (DOT) and comparing it with the Power BI dashboard. The process started by preparing the required data in SAS EG, querying the monthly and annual values from different tables in the Oracle database taking into consideration the time frame used in BI report, however due to data immigration process data was available from 2017 until the current time, performing the needed calculations, deleting duplicated data, setting up the right statistical formulas and finally collecting the processed data into one table. The second step was uploading the final table to the LASR server and getting it ready to be accessed from SAS VA. In SAS VA, the task required duplicating the structure in the BI report, creating the same number of pages/sections with the same number of charts in every section and placing the same statistical values in the upper part of the section. Figure 16 *Section from the SAS exploration dashboard in the classic display mode*shows a section from the dashboard made with SAS VA and presented in the classic mode, this section has an identical design to the page of the Power BI dashboard showed in figure 6. The section has the same three bar charts and the one line-chart, the same filtering options on the left of the section. It is important to highlight that SAS VA has another mode to visualize dashboards, which is the classic mode. The difference between the two modes lies mainly in the navigation process between the sections, wherein the modern mode a drop-down menu on the top right of the dashboard is used to move between the sections, as shown in Figure 18, while in classic mode the sections are displayed side by side, as shown in Figure 16. The second difference is the available possibility in the classic mode to export data of the charts into excel files and printing desired charts into PDF format files, as shown in Figure 17.

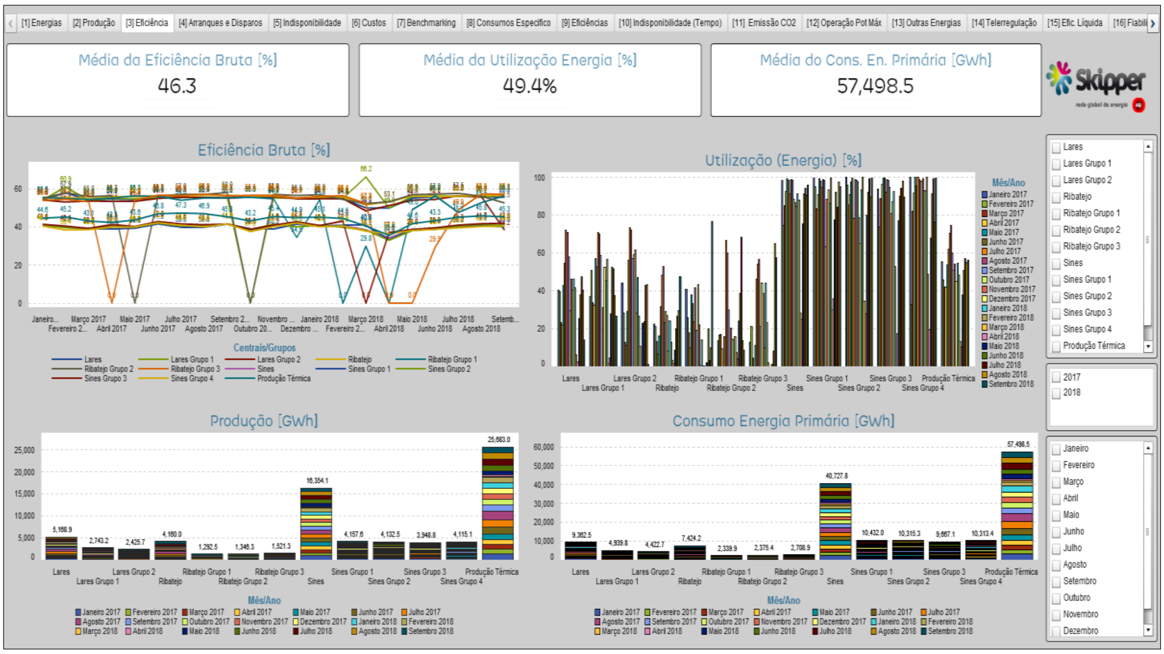


Figure 16 Section from the SAS exploration dashboard in the classic display mode

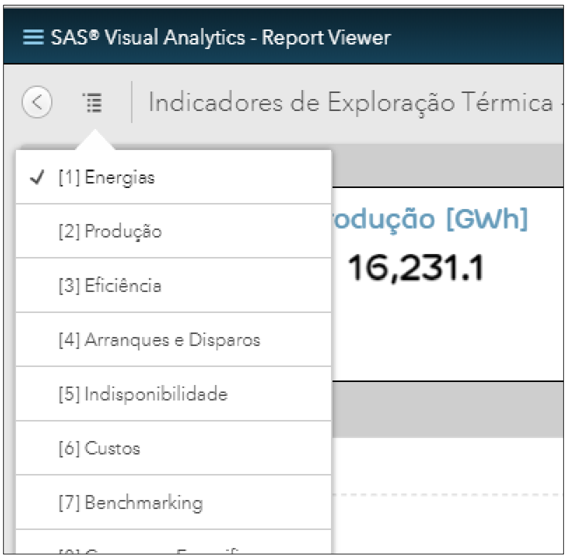


Figure 18 Navigation within SAS modern mode

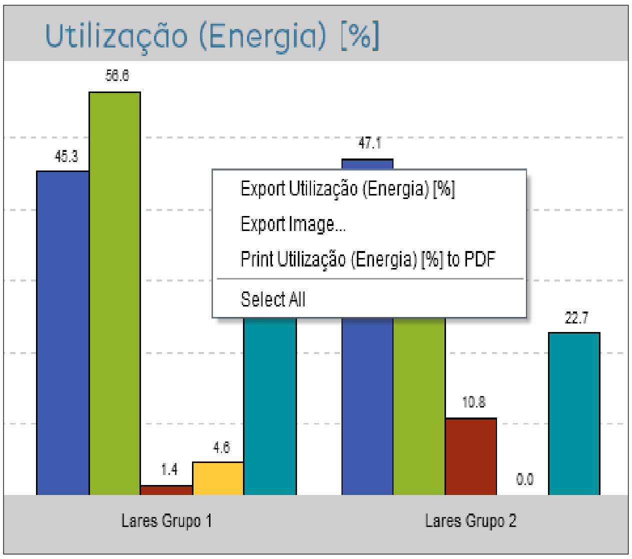


Figure 17 Data extracting options in SAS classic mode

4.3.2. SAS as a Reporting Tool

- **Maintenance Dashboard**

Another part of experimenting SAS in this research included proposing to enrich SKIPPER and developing another dashboard to the DOT. Its goal was presenting maintenance indicators for both types of power plants (thermal and hydraulic). This decision has been driven by two main factors: (1) the fact that the first dashboard has mainly exploration indicators, and (2) the importance of maintenance functions in EDPP through guaranteeing the functionality of the physical assets during their useful life using the most efficient techniques with the lowest cost, and thus contributing to the increase of the company's profitability.

In the same way, the dashboard is interactive with filtering options, giving the client the option to choose the thermal or hydraulic power plant he desires, dividing the hydraulic ones into two categories, silver and gold, in a timeframe that covers the years of 2016 and 2017 where the data is grouped as annual values and the client can choose which quarter(s) of the year to visualize. The proposed dashboard contains various types of graphs, bar charts, bubble plots, tree maps and list tables. The presented indicators are defined in accordance with the European Standard EN 15341, and the terminology in accordance with the European Standard NP 13306.

The indicators include economic values regarding maintenance costs whether it is contracted, corrective or conditioned maintenance and the impact of these costs in the whole production operation cost as an average, the indicators also explore the time of unavailability of the groups in the power plants due to maintenance operations.

Figure 19 shows one section from the maintenance dashboard presented in the modern mode, the section shows a comparison between two hydraulic power plants, Alqueva1 and Alqueva2, which are from the Silver category, regarding two maintenance indicators, E0 and E3, where E0 is the total operation cost, and E3 is the maintenance cost per emitted energy and / or available energy, during the last quarter in 2016 and the full year of 2017.

Figure 19 Section from the SAS maintenance dashboard in the modern view

- **Mobile Dashboard**

SAS VA offers the possibility of viewing the designed dashboards on mobile devices through an application called "SAS Visual Analytics". This functionality added an additional requirement: design a dashboard that is suitable to display on a mobile device, taking into consideration its trending the small size of the screen compared with the screen of a computer device. The dashboard is designed to present the exploration indicators for the three thermal power plants, the same ones discussed in the first report but with a different design and different indicators' grouping concept where the dashboard is composed of five main sections and the sections are designed to present five main indicators categories, which are production, service quality, efficiency, consumption, and environment. Every one of these sections has a group of several indicators that are in the same area, for example, production indicators are divided based on the type of the row source, gas in both Lares and Ribatejo, and coal in Sines, efficiency is presented by two indicators: (1) percentage of energy usage measurement and reliability of the produced energy, and (2) environmental indicators include measurements that indicate the efficiency of the systems that minimize the emission of gases such as CO_2 and SO_2.

It is important to note that the design of the dashboard depends on using Info Pages option in SAS VA, where it is possible to identify a section of the dashboard to be displayed as an information page that is not presented directly, the client needs to click on an icon/link in one section to display the information page that is connected to it.

The dashboard covers a time frame of 2017 and 2018 and gives the client two options included in the dropdown menu on the left-top of the page, "Month" option that shows the values in one specific month of the year, or "YTD" that shows the values from the beginning of the year until the chosen month. Figure 20 shows the normal section of the dashboard that explores the Service Quality (Qualidade de Service) presenting three indicators: a count of the stops and the starts of the power plants, availability and unavailability of the power plant for emission in terms of time and energy. Clicking on each one of the sub-titles causes its related info page to be the next page to display, Figure 21 shows the information page that pops up when clicking on the availability (Disponibilidades), the info page presents a butterfly chart that compares the availability of time and energy during the year 2018 until April, for the three thermal power plants.

Figure 20 Service Quality section in the mobile dashboard

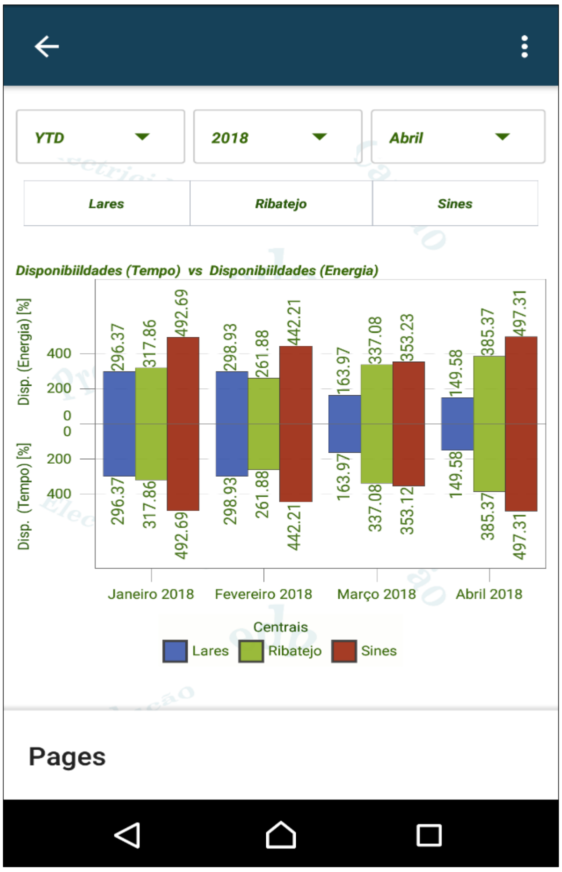


Figure 21 Availability information section

4.4. Step 2: Advanced Analytics

The advanced forecasting techniques implemented in this thesis are performed in three different platforms, based on the same dataset.

4.4.1. SAS Data Explorer – Forecasting Tool

SAS VA Explorer is a component of SAS VA that allows exploring the data source interactively using charts, histograms, and tables. SAS Explorer also enables performing data analysis such as correlation, fit lines, and forecasting which will be the focus of this work.

It is useful to point out to some few important notes on SAS VA Explorer, the work in the explorer is saved as a metadata object called an exploration, an exploration (sometimes called a visual exploration) contains all the visualizations, data settings, and filters from the explorer session ("What Is SAS Visual Analytics Explorer?," n.d.). The explorer has a connection to the LASR server and it is possible to export the explorations' results as reports, and thus it is possible to refine them in SAS VA or view them directly. However, it is not possible to edit the exported reports analysis options in SAS VA. Furthermore, SAS explorer also allows saving explorations in a pdf format, exporting the charts as image files, and exporting the data of the visualizations into excel or CSV files.

Applying forecasting in SAS data explorer requires the chart to be a line chart where it must contain a datetime, time or date item as a category, forecasting is done by implementing a scenario and a goal seeking analysis. In both cases, exists a measure to forecast and a group of underlying factors that affect the variable.

- **Assign Forecasting Variables**

As shown in Figure 22, a usual line chart is plotted with Month/Year (Mês/Ano [Date]) item as a category and the measure to forecast is the electricity production indicator (Produção) in the thermal power plants as a total.

Checking the forecasting option will perform a basic forecast of the production values for the next six months, a duration defined to forecast, based on the historical values of the production time series as shown in Figure 23.

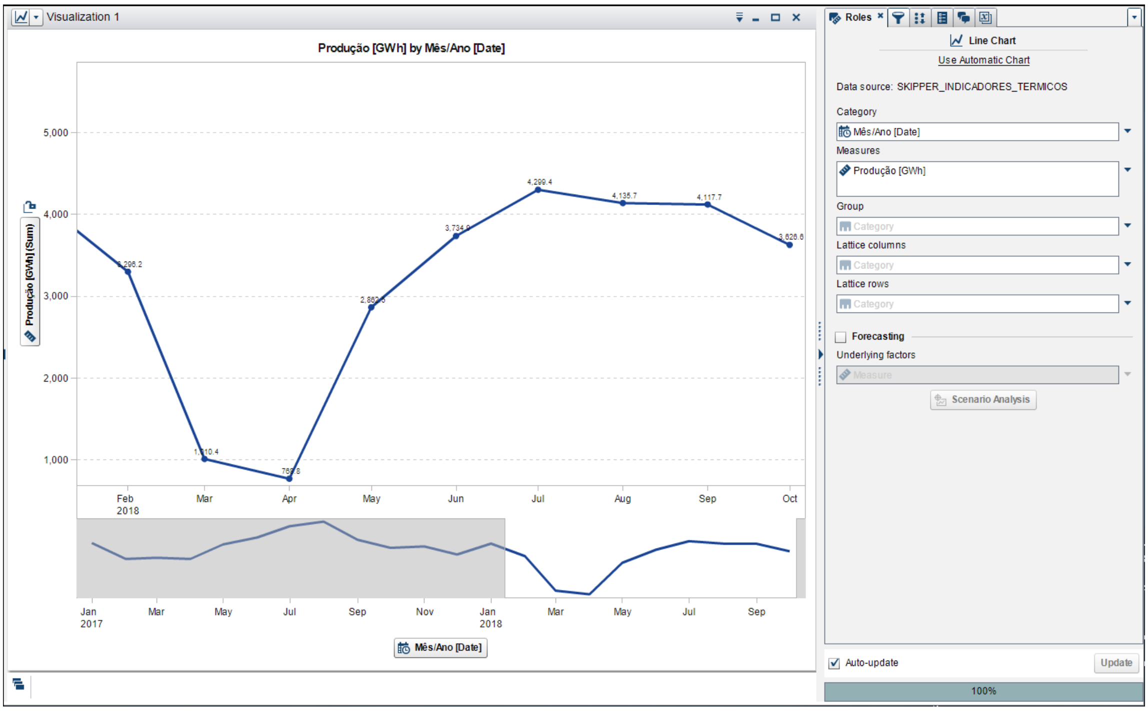


Figure 22 Line chart of production indicator

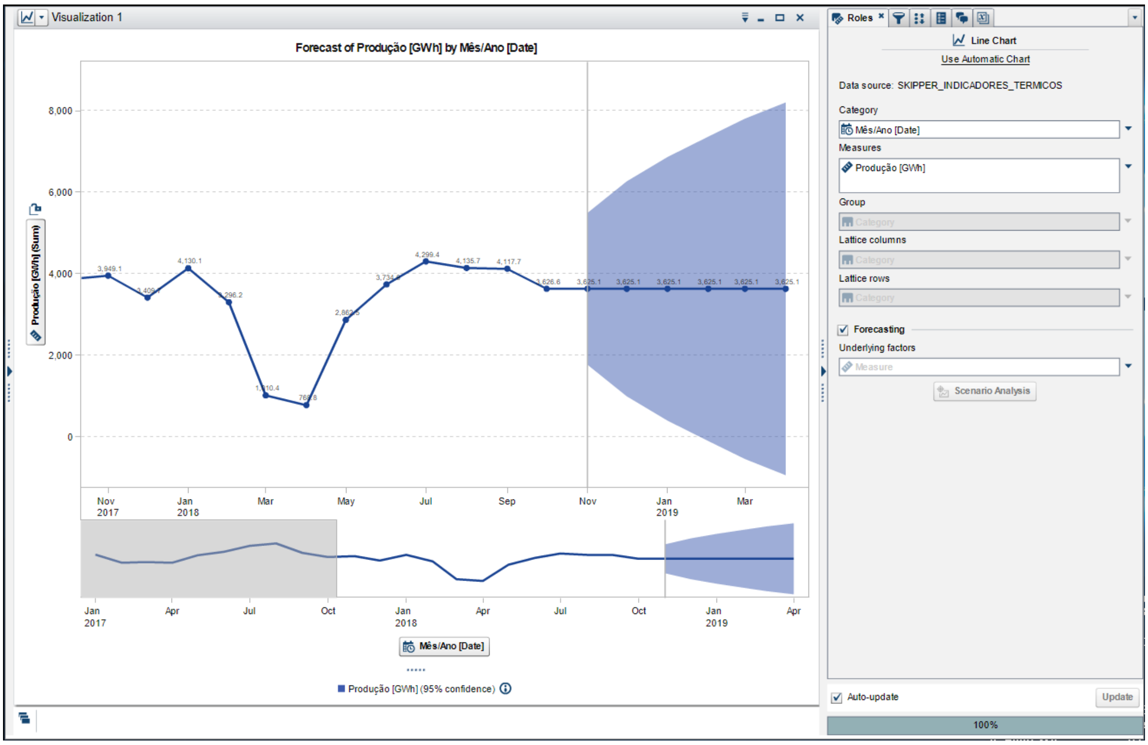


Figure 23 Basic forecasting of electricity production

However, to apply a forecast analysis, a group of seven indicators was selected as the most relevant and correlated ones to the production measure based on information from SKIPPER, as the following:

- Operating Hours [#], «Horas de Funcionamento [#]»;
- Number of start-ups [#], «Número de arranques [#]»;
- Usage (Time) [%], «Utilização (Tempo) [#] »;
- Usage (Energy) [%], «Utilização (Energia) [#] »;
- Number of stops [#], «Horas de Funcionamento [#]»;
- Availability (Time) [%], « Disponibilidade (Tempo) [%] »;
- Availability (Energy) [%], « Disponibilidade (Energia) [%] ».

SAS VA Explorer went through the underlying factors, choosing automatically the ones considered more related and significant to the forecast measure, which are in this case the first three indicators from the previous group, and then performing a forecast operation based on the historical values of the production time series along with the historical values of the three chosen underlying factors time series. Figure 24 presents the forecasting result with four separate visualizations, where the top visualization represents the overall production and the predicted six-month production going forward, and the other three visualizations present three channels that were identified as significant by the explorer.

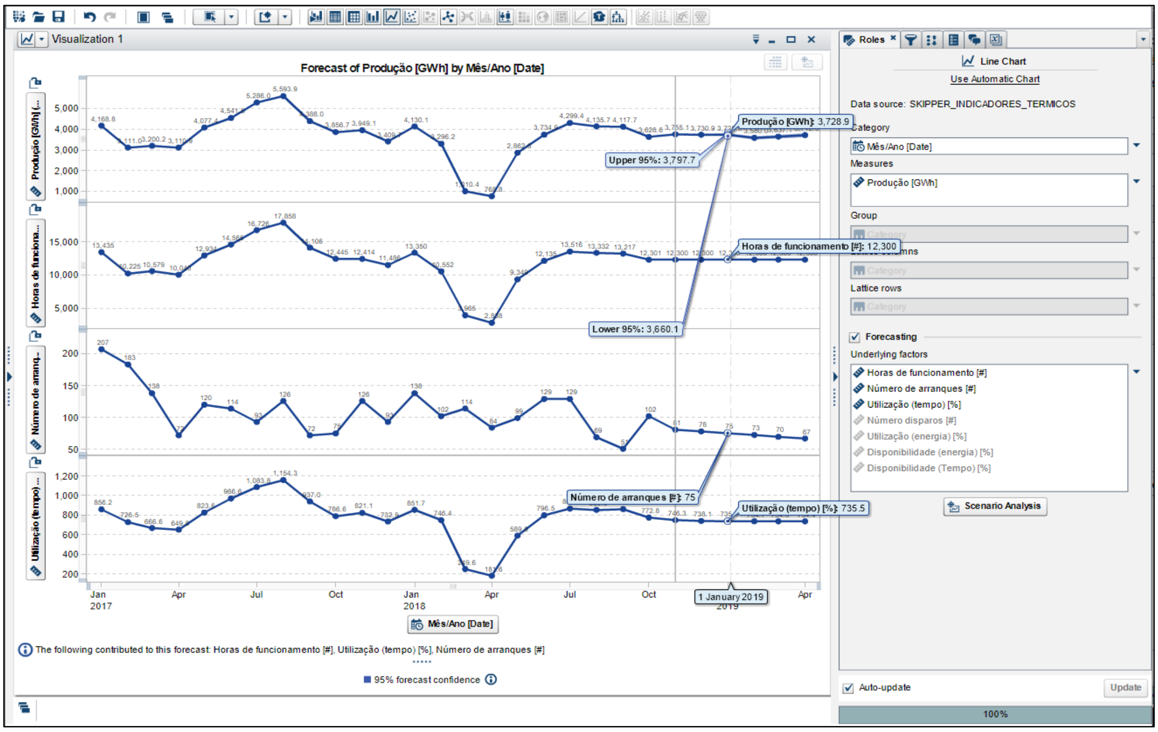


Figure 24 Production Forecast based on underlying factor

For a specific point in the future, January 2019, the chart presents the predicted production value, the upper and lower limits of the prediction, and the corresponding values of the underlying factors. It is important to note that at the bottom of the visualization it is possible to confirm the factors contributing in the forecast, besides the percentage of the forecast confidence which is in this case 95%.

- **Apply a Scenario Analysis**

Applying a scenario analysis enables exploring hypothetical scenarios and answers a group of questions such as: what if the value of the Operating Hours in a month increases by 50%, how this will affect the overall production forecast?

Performing this scenario starts by opening the Scenario Analysis window, Figure 25, choosing the required future point in the required factor's time series, Operating Hours, selecting the "Set Series Values" option which causes changes in the time series as a whole not only in the chosen point, Figure 26, and adjusting the future value of the time series by 50%, Figure 27.

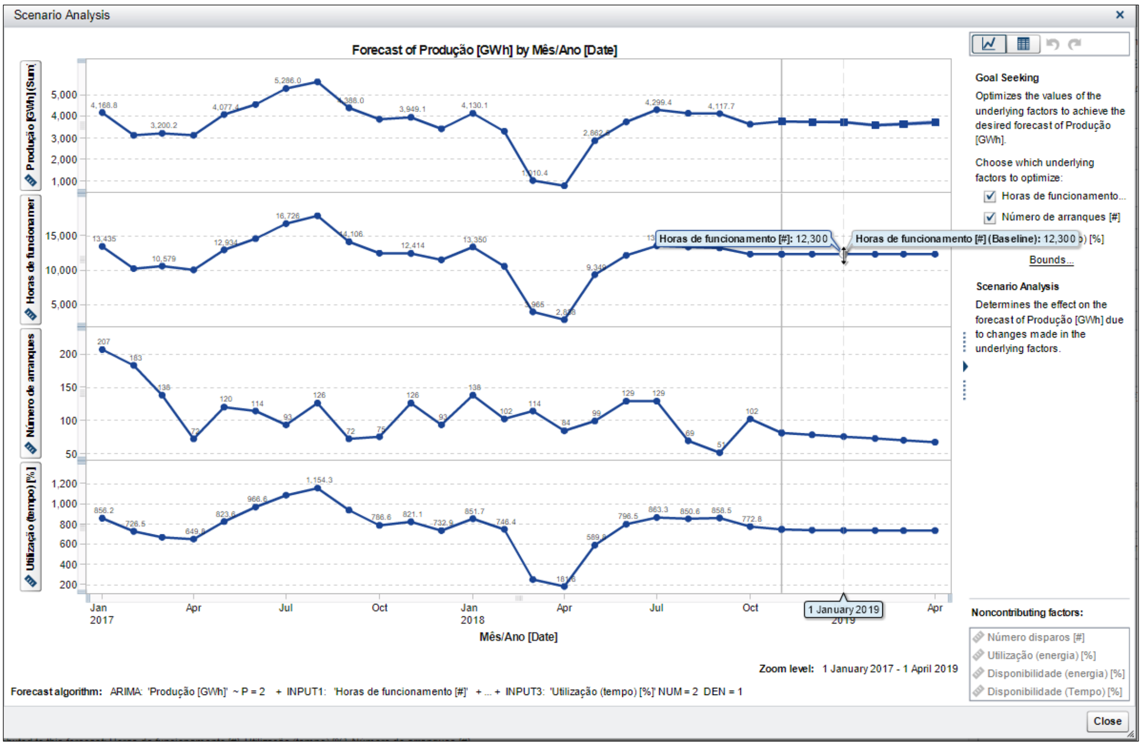
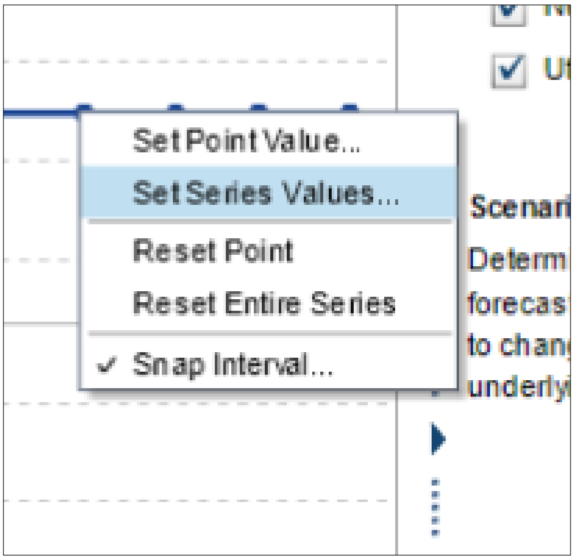


Figure 25 Scenario Analysis window



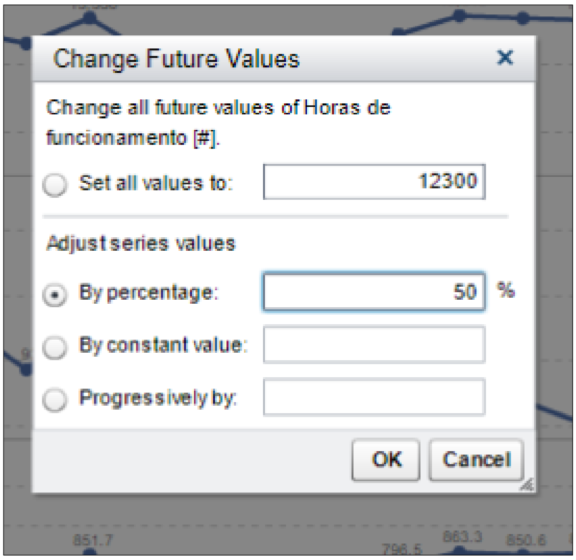


Figure 26 Applying Scenario analysis

Figure 27 setting up Scenario options

Applying the scenario analysis will affect the future values of the production time series, Figure 28, where it is possible to see the original baseline of the production time series as a gray line, the new future predicted values with the upper and lower limits of the prediction, the original baseline and the new values of the chosen underlying factor's time series, besides the original baselines and new future values of the other two underlying factors' time series, even though the difference might be very small regarding these two. At the bottom of the scenario, analysis window appears the used forecast algorithm (in this case ARIMA) with the used input parameters.

Implementing the scenario analysis enables SKIPPER's client to do the following:

- Performing different scenarios based on different factors and thus answering different questions based on his own assumptions and visions;
- In the scenario analysis window, it is possible to view the production prediction values in a table view;
- In the explorer window, the client can view the full analysis results in a table where both historical and prediction values exist;
- The client can export the data form the result table, in the explorer window, into a CSV file, as shown in Figure 29.

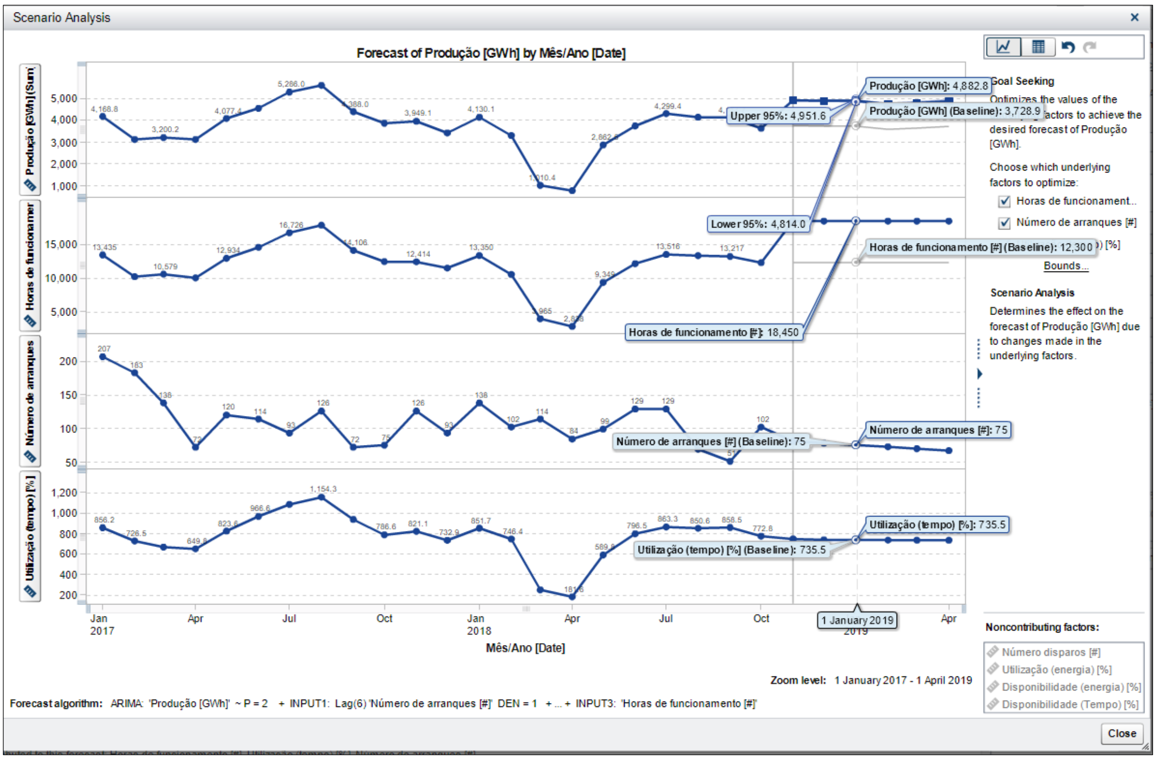


Figure 28 Result of applying Scenario Analysis by changing Operating Hours factor by 50%

Results | Analysis

Mês/Ano [Date] ▲	Produção [GWh]	Produção [GWh] (Lower 95%)	Produção [GWh] (Baseline)	Produção [GWh] (Upper 95%)	Horas de funcionamento [#]	Horas de funcionamento [#] (Baseline)	Número de arranques [#]	Número de arranques [#] (Baseline)
1 February 2018	3,296.2				10,552		102	
1 March 2018	1,010.4				3,965		114	
1 April 2018	768.8				2,838		84	
1 May 2018	2,862.5				9,349		99	
1 June 2018	3,734.9				12,135		129	
1 July 2018	4,299.4				13,516		129	
1 August 2018	4,135.7				[illegible]		69	
1 September 2018	4,117.7				[illegible]		51	
1 October 2018	3,626.6				12,301		102	
1 November 2018	4,908.9	4,858.3	3,755.1	4,959.6	18,450	12,300	81	81
1 December 2018	4,884.7	4,829.5	3,730.9	4,940.0	18,450	12,300	78	78
1 January 2019	4,882.8	4,814.0	3,728.9	4,951.6	18,450	12,300	75	75
1 February 2019	4,733.9	4,654.3	3,580.0	4,813.5	18,450	12,300	73	73
1 March 2019	4,791.0	4,708.0	3,637.1	4,874.0	18,450	12,300	70	70
1 April 2019	4,865.8	4,768.8	3,712.0	4,962.9	18,450	12,300	67	67

Export Data...
Hide Details
Hide Unselected Table Rows

Figure 29 Exporting data from result table of scenario analysis

- **Apply a Goal Seeking Analysis**

A goal seeking analysis is implemented in this work to enable the client to optimize the decision-making process regarding the production measure whereas it aims to predict how the factors affecting the production need to be changed in order to get to a target point in the production.

This analysis is performed based on information from UNGE indicating that the production in December 2018 needs to be 4654.865 MWh. To perform this analysis, the same scenario analysis window is used but with changing the time series to modify whereas, as shown in Figure 30, the option "Set Point Value" is selected to change the value of the production time series at the specified future date point from 3730.851500792141 to 4654.865 MWh, as shown in Figure 31.

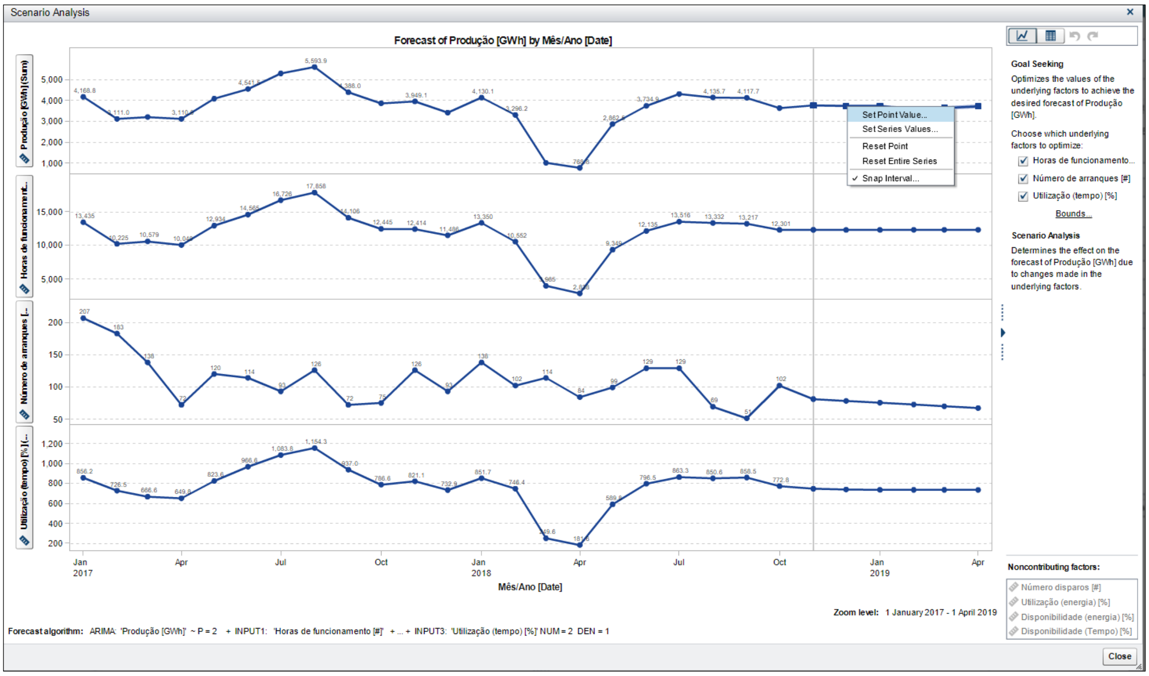


Figure 30 Applying Goal Seeking analysis on the production time series

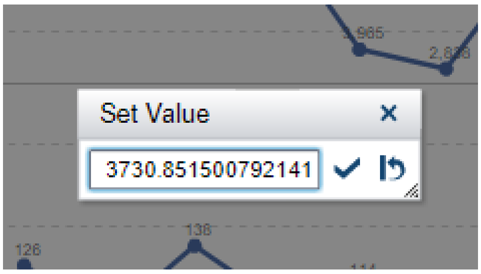


Figure 31 Changing a future value in the production time series

Applying the new value of the production time series, the target value, will cause changes in the time series of the underlying factors so that the new future values of these series are the required ones to obtain the target. Figure 32 presents the baseline and the new line representing the changes in the prediction measure along with the upper and lower values of the prediction, and the corresponding changes in the future predictions of the underlying factors whereas the biggest change took place in the time series of Usage (Time) factor, the Number of Stops time series is the same, and a small raise occurred in the time series value of Operating Hours factor.

Implementing the goal analysis enables SKIPPER' clients to do the following:

- The client can develop a better idea about the future since this analysis can answer question like: what needs to be done to increase the production by 50%?
- Make better decisions based on the changing values to meet the desired production target.
- Like the scenario analysis, the client is able to view the resulting analysis in a table view.

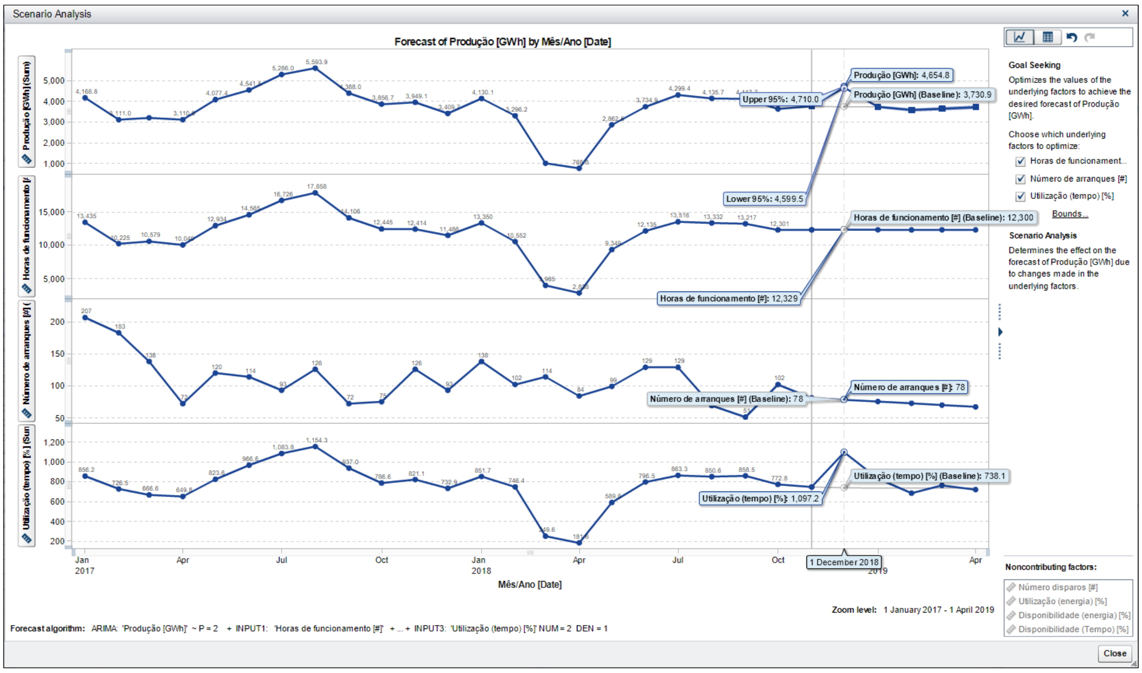


Figure 32 Results of applying a goal seeking analysis

4.4.1. R studio

R Studio is a free, open-source software and programming language developed in 1995 at the University of Auckland as an environment for statistical computing and graphics (Roecker, Yoast, Wills, & D'avello, n.d.). R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form ("What is R?," n.d.).R can be described as an environment where statistical techniques are implemented. R Studio is an integrated development environment (IDE) that allows the interaction with R to be more readily, and it is similar to the standard RGui but is considerably more user friendly (Roecker et al., n.d.). R Studio was found in 2008, and the first delivery of the free open source product was in 2009 (R Studio, 2012). R Studio runs either on a desktop or in a browser connected to R Studio Server.

This work is done on the R Studio desktop software, with the goal of experimenting different statistical forecasting methods on the data considering it as a time series data. The used dataset in this part is the same dataset used to perform forecasting with SAS Data Explorer but with more external factors. The work was provided these factors by UNGE with the required data that included:

- Prices of raw materials necessary for the thermal energy production (gas and coal);
- The recoupment that EDP needs to pay for the CO2 emission;
- Monthly averages of temperatures;
- The levels of water behind the dams in the hydraulic power plants.

4.4.2.1. Data Structure

Since the data is describing the energy production and the affecting factors in the thermal power plants as totals and the detailed values of the sub-groups of the plants, then the data has a hierarchical structure of 3 levels that is shown in Figure 33. Based on this fact and to perform forecasting operations on the production hierarchical time series, an extraction process is performed on the one table extracted from SAS EG that contains the full data to get the detailed production data reconstruct it in a hierarchical form that make it usable in R Studio. The time series is imported as a data set in the global environment in R, and treated as hierarchical time series (hts) object that has a time frame starting in July 2016 and ending in October 2018.

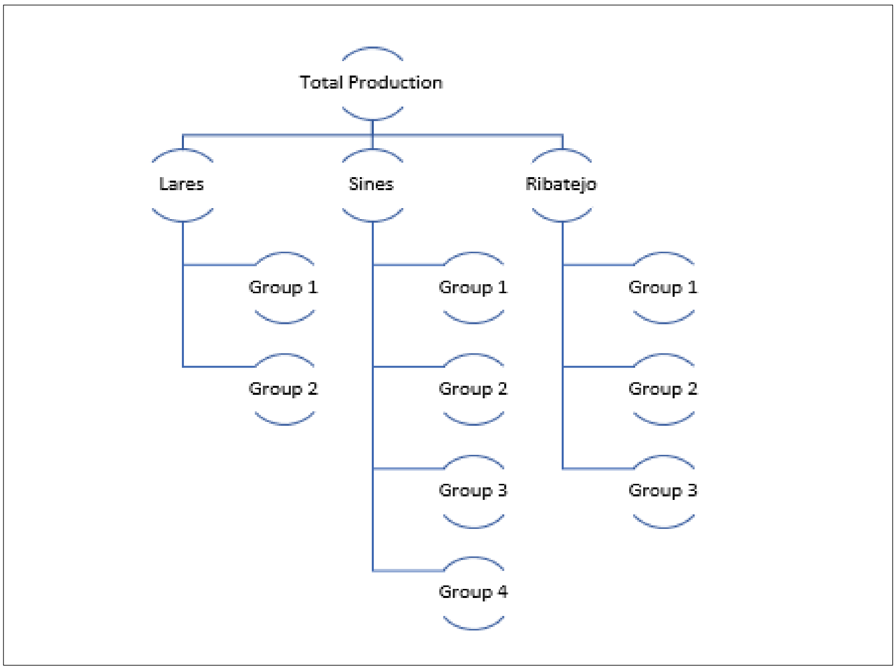


Figure 33 Hierarchical structure of each time series in the data

4.4.2.2. Production Hierarchical Time Series Forecasting / Arima Method

The first experiment was performed on the production hierarchical time series, and two libraries are needed at this stage, “forecast” and “hts”. The data set is imported into R, and stored as a time series object, with indication to the start date of the series and identifying the samples as monthly ones. The production time series object is then transformed into hts object using the following piece of code, and the results of plotting the time series of the hierarchy is shown in Figure 34. To perform the forecasting methods and measure the accuracy of each one, the data of the hierarchical production time series is divided into training and testing sets, training set is composed of data over 24 months starting July 2016 and the testing set is composed of data over 6 months starting May 2018.

```
Library (hts) #Loading hts library
# Transforming data set into time series object using “ts” function
production_ts <- ts (Production_data, frequency = 12, start = c (2016,7))

# Transforming ts object into hierarchical time series object using “hts” function
production_hts <- hts (production_ts, characters = c (3,6))

#extracting time series from the hierarchy using “aggts” function and plotting them
production_hts %>% aggts (levels = 0:2) %>%
  + autoplot (facet = TRUE)
```

```
production_training_set <- window (production_data_hts, start = c (2016,7), end = c (2018,4))

production_testing_set <- window (production_data_hts, start = c (2018,5))
```

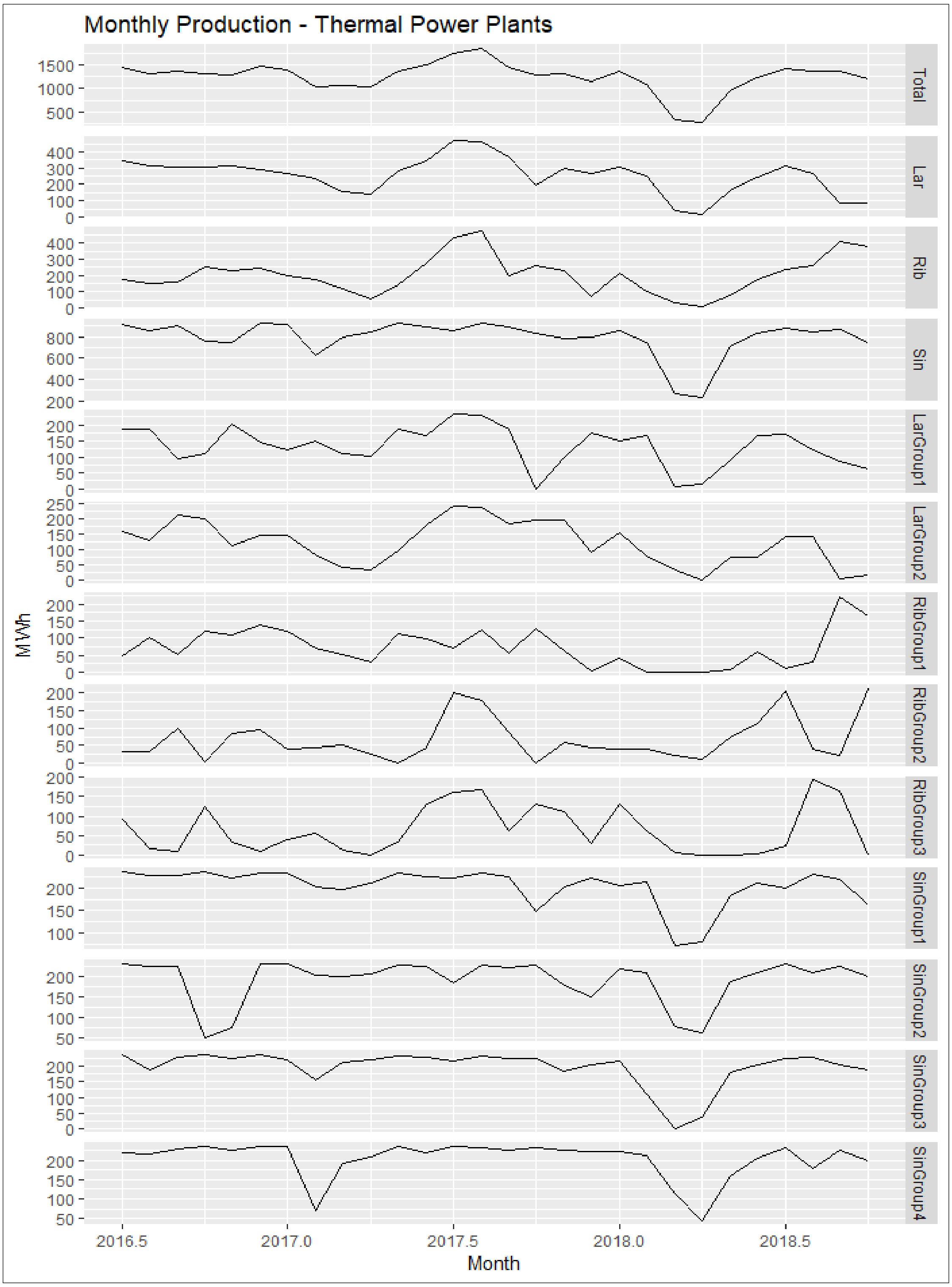


Figure 34 Plot of time series in the production hts object

- **Bottom – up Forecast**

The first approach applied on the production hierarchical time series is the bottom up method, which is done through the forecast function applied only on the training dataset as the following:

```
Bottom_up_production_hts <- forecast (production_data_hts, h = 6, method = "bu", fmethod = “arima")
```

The code line shows using “arima” as a forecasting method, which will apply the function “auto.arima” to every time series in the bottom level of the hierarchy, where the orders *(p, d, q)* are calculated automatically. The forecasting horizon is six months, and the method “bu” defines the use of bottom-up approach.

Running a summary function on the previous forecasts shows the following details of the hierarchy forecasting process, the forecasts of the top level, the total production values for the three power plants per month, in the hierarchy starting November 2018 and ending in April 2019, Figure 35 shows the line chart resulting from the bottom-up forecasting.

```
> summary (Bottom_up_production_hts)
Hierarchical Time Series
3 Levels
Number of nodes at each level: 1 3 9
Total number of series: 13
Number of observations in each historical series: 28
Number of forecasts per series: 6
Top level series of forecasts:
       Jan       Feb      Mar      Apr May Jun Jul Aug Sep Oct Nov       Dec
2018                                                      1155.8210 1126.3975
2019 1154.1894 1093.2621 884.2143 857.7573

Method: Bottom-up forecasts
Forecast method: Arima
```

The full forecast values of the lower levels of the production hierarchy time series are viewed as the following:

```
> Bottom_up_production_hts %>% aggts(levels = 0:1)
             Total        Lar      Rib      Sin
Nov 2018 1155.8210  132.55192 277.8127 745.4563
Dec 2018 1126.3975  103.45192 226.4338 796.5118
Jan 2019 1154.1894  141.55192 216.3706 796.2669
Feb 2019 1093.2621   85.05192 211.9433 796.2669
Mar 2019  884.2143 -122.04808 209.9954 796.2669
Apr 2019  857.7573 -147.64808 209.1385 796.2669
> Bottom_up_production_hts %>% aggts(levels = 2:2)
         LarGroup1   LarGroup2 RibGroup1 RibGroup2 RibGroup3 SinGroup1 SinGroup2 SinGroup3 SinGroup4
Nov 2018      98.9   33.651923 115.20997 143.57182  19.03095  181.5741  173.4367  193.7009  196.7446
Dec 2018     173.5  -70.048077  92.33675  70.73266  63.36435  203.2386  192.7803  198.1038  202.3891
Jan 2019     149.4   -7.848077  82.27358  70.73266  63.36435  203.2386  192.7803  197.8590  202.3891
Feb 2019     168.6  -83.548077  77.84625  70.73266  63.36435  203.2386  192.7803  197.8590  202.3891
Mar 2019       4.7 -126.748077  75.89843  70.73266  63.36435  203.2386  192.7803  197.8590  202.3891
Apr 2019      14.8 -162.448077  75.04148  70.73266  63.36435  203.2386  192.7803  197.8590  202.3891
```

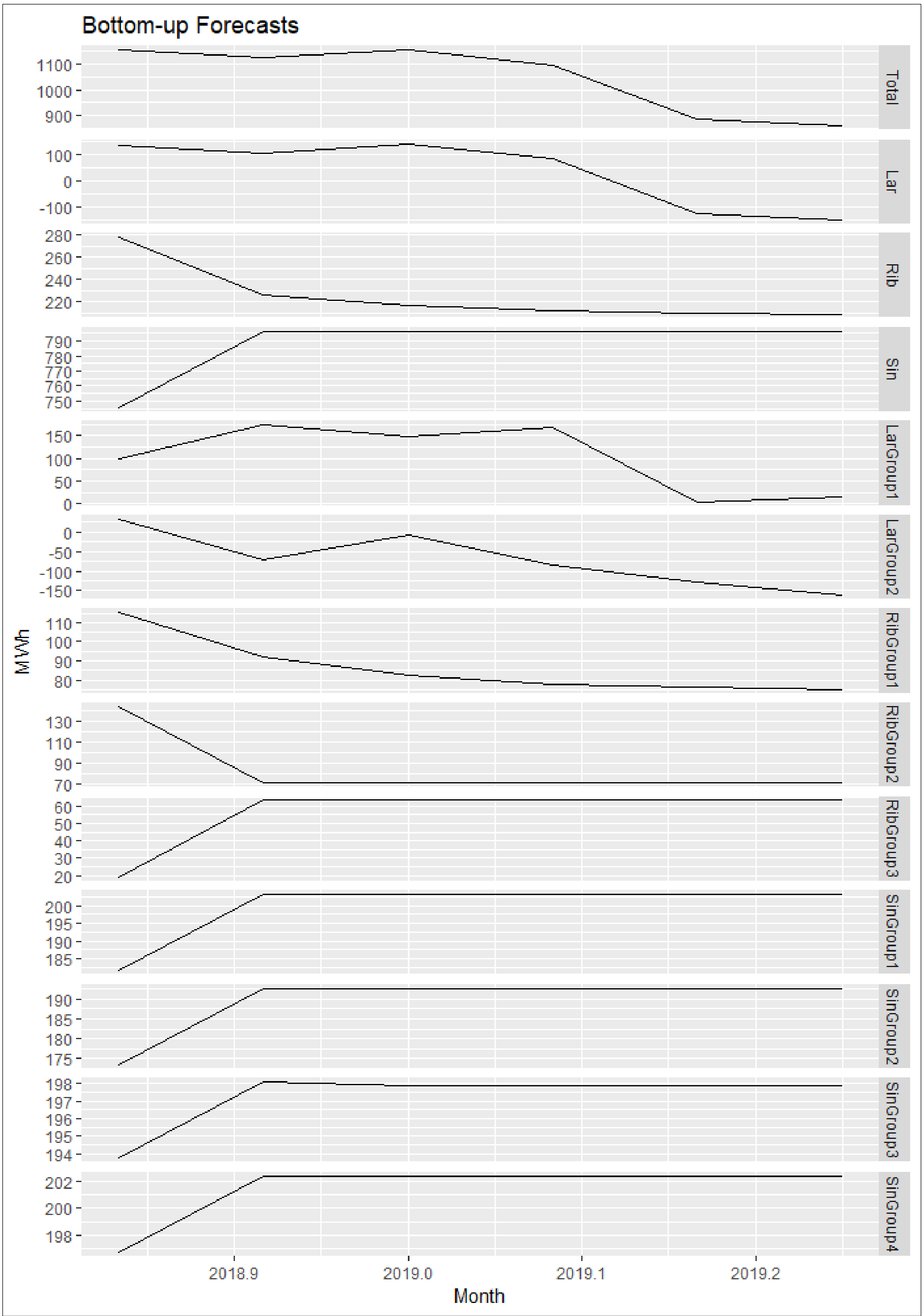


Figure 35 Line chart of bottom-up forecasting approach

- **Top – down Forecast**

The second applied approach on the production hierarchical time series is the top-down forecasting method, and this is performed using the following line code:

```
> Top_down_production_hts <- forecast (production_data_hts, h=6, method = "tdfp", fmethod="arima")
```

In the same way of the previous approach, the forecasting method is arima, the forecasting horizon is six months, and the same hierarchical time series is used, the only difference is using the value “tdfp”, which stands for “top-down forecast proportions”, for the method argument. Using “tdfp” argument indicates that the forecast proportions used for disaggregation are based on forecasts rather than historical data the thing that will produce more accurate forecasts at lower levels of the hierarchy since that proportions that are based on historical data don’t consider the changes that occurs over time (Hyndman & Athanasopoulos, 2018a). The following details result from summarising the previous forecasts:

```
> summary (Top_down_production_hts)
Hierarchical Time Series
3 Levels
Number of nodes at each level: 1 3 9
Total number of series: 13
Number of observations in each historical series: 28
Number of forecasts per series: 6
Top level series of forecasts:
        Jan    Feb    Mar    Apr May Jun Jul Aug Sep Oct    Nov    Dec
2018                                                     1239.7 1060.0
2019 1300.3 1022.1  260.2  179.6
Method: Top-down forecasts using forecasts proportions
Forecast method: Arima
```

The full forecast values of the lower levels of the production hierarchy time series are viewed as the following, and Figure 36 shows the line chart of the forecasting operation:

```
> Top_down_production_hts %>% aggts(levels = 0:1)
          Total       Lar        Rib      Sin
Nov 2018 1239.7 199.65145 254.211181 785.8374
Dec 2018 1060.0 159.13614  77.339129 823.5247
Jan 2019 1300.3 207.55267 232.251007 860.4963
Feb 2019 1022.1 133.53770 103.390405 785.1719
Mar 2019  260.2 -24.94070   9.976279 275.1644
Apr 2019  179.6 -24.73893   2.990977 201.3479
> Top_down_production_hts %>% aggts(levels = 2:2)
           LarGroup1  LarGroup2  RibGroup1  RibGroup2  RibGroup3 SinGroup1 SinGroup2 SinGroup3 SinGroup4
Nov 2018 148.9644797   50.68697 105.422319 131.374688 17.4141737 191.40989 182.83172 204.19357 207.40219
Dec 2018 266.8884268 -107.75228  31.537893  24.158953 21.6422830 210.13120 199.31829 204.82234 209.25289
Jan 2019 219.0600352  -11.50736  88.312012  75.924052 68.0149434 219.63242 208.33060 213.81891 218.71439
Feb 2019 264.7142578 -131.17656  37.975047  34.504888 30.9104688 200.40667 190.09417 195.10205 199.56900
Mar 2019   0.9604516  -25.90115   3.605716   3.360305  3.0102577  70.23275  66.61873  68.37375  69.93919
Apr 2019   2.4797891  -27.21871   1.073199   1.011577  0.9061999  51.39190  48.74738  50.03159  51.17709
```

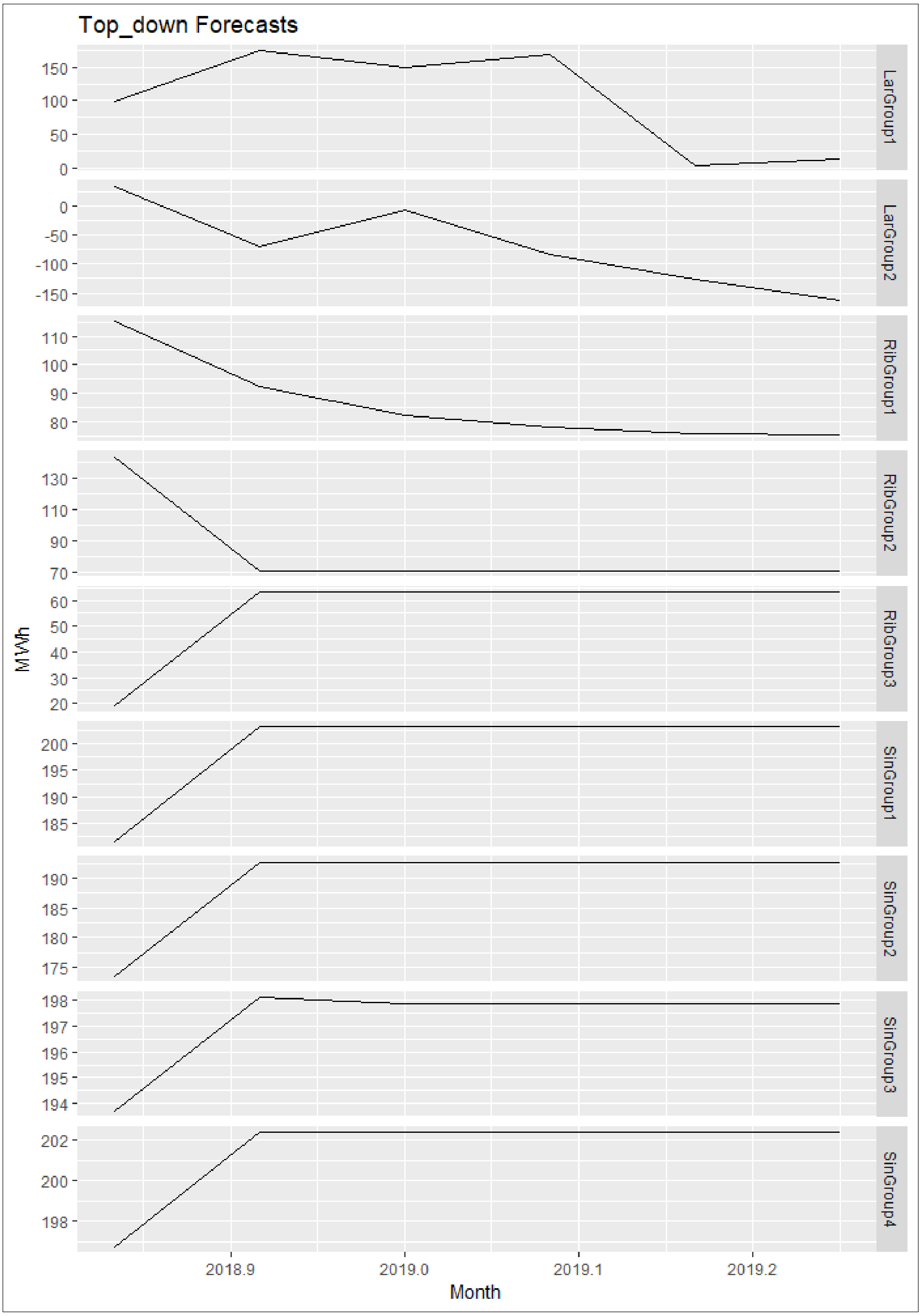


Figure 36 Line chart of top-down forecasting approach

- **Middle-out Approach**

The third approach applied to the production hierarchical time series is the middle-out method through the following code line where the method argument is set to “mo”, the forecasting horizon is six months and the arima method is used as well. The extra argument in this approach is the “level” argument which defines the level where the method should start forecasting from.

```
> Middle_out_production_hts <- forecast (production_data_hts, h=6, method="mo", fmethod = "arima"
, level = 1)
```

A summary of the previous forecast and the six months forecasts of the top level of the production hierarchical series can be shown as the following:

```
Summary (Middle_out_production_hts)
Hierarchical Time Series
3 Levels
Number of nodes at each level: 1 3 9
Total number of series: 13
Number of observations in each historical series: 28
Number of forecasts per series: 6
Top level series of forecasts:
          Jan       Feb       Mar      Apr May Jun Jul Aug Sep Oct   Nov       Dec
2018                                                                 1133.8221 1022.4579
2019 1200.3579 1034.0579 751.1579 708.5579
Method: Middle-out forecasts
Forecast method: Arima
```

Figure 37 shows the plot of the middle-out forecast, and the forecasts values of the three levels of the hierarchy are displayed in R as the following:

```
> Middle_out_production_hts %>% aggts(levels = 0:1)
              Total   Lar   Rib      Sin
Nov 2018 1133.8221 182.6 232.5 718.7221
Dec 2018 1022.4579 153.5  74.6 794.3579
Jan 2019 1200.3579 191.6 214.4 794.3579
Feb 2019 1034.0579 135.1 104.6 794.3579
Mar 2019  751.1579 -72.0  28.8 794.3579
Apr 2019  708.5579 -97.6  11.8 794.3579
> Middle_out_production_hts %>% aggts(levels = 2:2)
          LarGroup1  LarGroup2 RibGroup1  RibGroup2 RibGroup3 SinGroup1 SinGroup2 SinGroup3 SinGroup4
Nov 2018 136.242007   46.35799 96.418612 120.154491 15.926897  175.0623  167.2168  186.7542  189.6888
Dec 2018 257.436008 -103.93601 30.420912  23.303312 20.875776  202.6890  192.2590  197.5681  201.8418
Jan 2019 202.222898  -10.62290 81.524277  70.088466 62.787258  202.7513  192.3181  197.3846  201.9038
Feb 2019 267.811229 -132.71123 38.419329  34.908571 31.272100  202.7513  192.3181  197.3846  201.9038
Mar 2019   2.772678  -74.77268 10.409153   9.700690  8.690156  202.7513  192.3181  197.3846  201.9038
Apr 2019   9.783263 -107.38326  4.233986   3.990874  3.575140  202.7513  192.3181  197.3846  201.9038
```

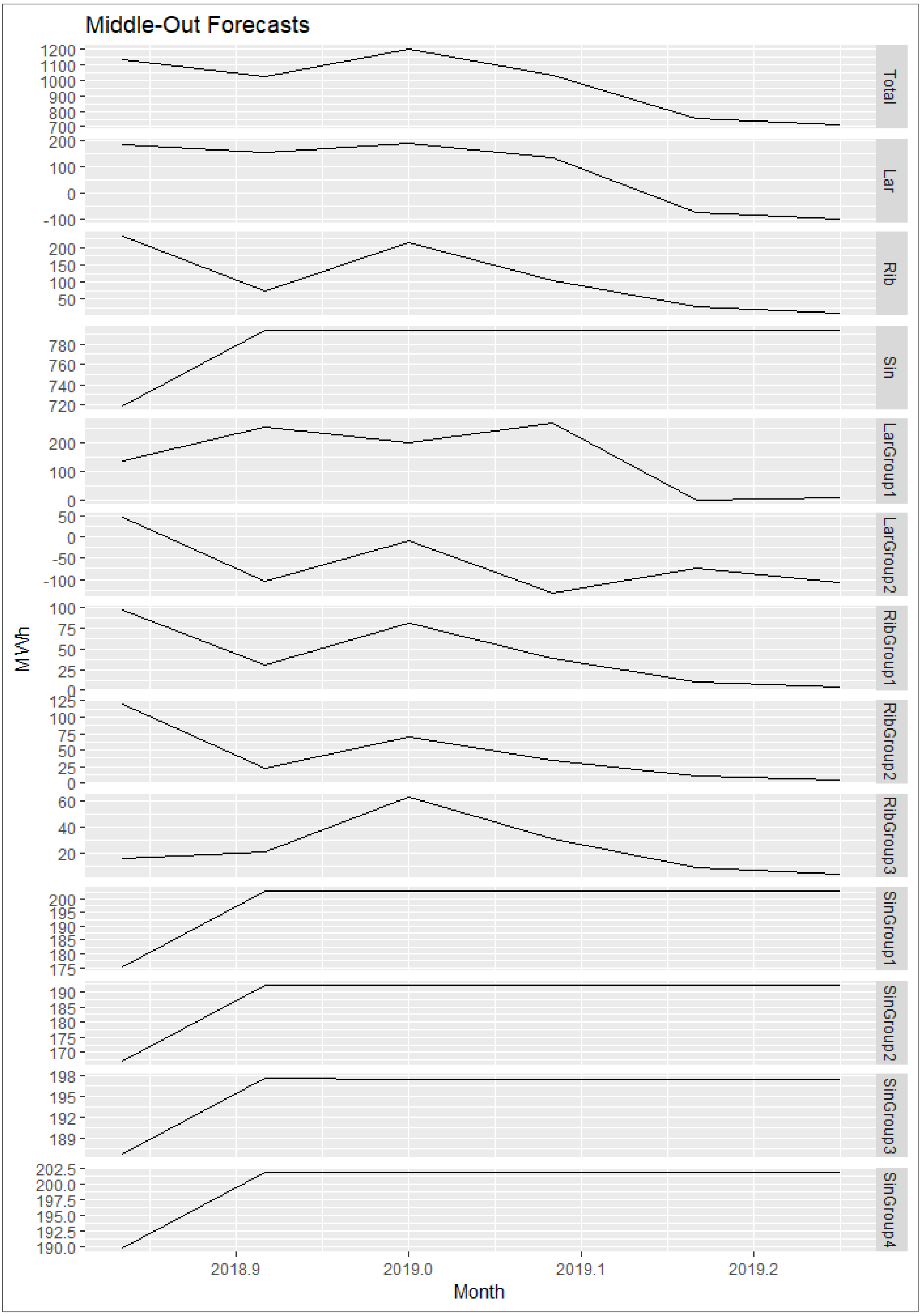


Figure 37 Line chart of middle-out forecasting approach

- **The Optimal Reconciliation Approach**

Implementing the optimal reconciliation on the production hierarchical time series is done by applying the forecast function and leaving the default value of the method argument, which will have the “comb” value, as the following:

```
> Opt_production_hts <- forecast (production_data_hts, h=6, fmethod = "arima")
```

In the same way, a summary of the forecasting process, the used method, the number of the hierarchy levels along with the number of nodes in each, and the top-level forecasts values can be displayed as the following:

```
> summary (Opt_production_hts)
Hierarchical Time Series
3 Levels
Number of nodes at each level: 1 3 9
Total number of series: 13
Number of observations in each historical series: 28
Number of forecasts per series: 6
Top level series of forecasts:
           Jan       Feb      Mar      Apr     May Jun Jul Aug Sep Oct  Nov       Dec
2018                                                                    1186.5785 1064.0900
2019 1234.0897 1043.6628 569.3258 513.9423

Method: Optimal combination forecasts
Forecast method: Arima
```

Figure 38 shows the plot of the optimal forecast combination forecast, and the forecasts values of the three levels of the hierarchy are displayed in R as the following:

```
> opt_production_hts %>% aggts(levels = 0:1)
              Total       Lar        Rib      Sin
Nov 2018 1186.5785  172.4417 266.793171 747.3436
Dec 2018 1064.0900  133.7501 136.288789 794.0511
Jan 2019 1234.0897  183.6361 234.567940 815.8856
Feb 2019 1043.6628  112.4207 142.755484 788.4866
Mar 2019  569.3258 -142.8899  13.494241 698.7214
Apr 2019  513.9423 -172.7175  -4.190098 690.8499
> opt_production_hts %>% aggts(levels = 2:2)
           LarGroup1   LarGroup2  RibGroup1  RibGroup2 RibGroup3 SinGroup1 SinGroup2 SinGroup3 SinGroup4
Nov 2018 121.8977937   50.543930 111.814271 139.724774 15.254126  181.9592  173.9706  194.1416  197.2721
Dec 2018 190.9678927  -57.217813  64.558398  39.262114 32.468277  202.7364  192.0842  197.5292  201.7013
Jan 2019 173.6629190    9.973171  87.881131  77.085546 69.601263  207.2420  198.3302  202.4408  207.8726
Feb 2019 184.3790218  -71.958298  56.525901  46.578488 39.651095  201.6509  190.5794  196.0419  200.2144
Mar 2019  -7.3159994 -135.573895  15.346196   2.132046 -3.984001  183.3331  165.1861  175.0777  175.1246
Apr 2019   0.3466442 -173.064147   9.303852  -3.742570 -9.751379  181.7268  162.9594  173.2393  172.9244
```

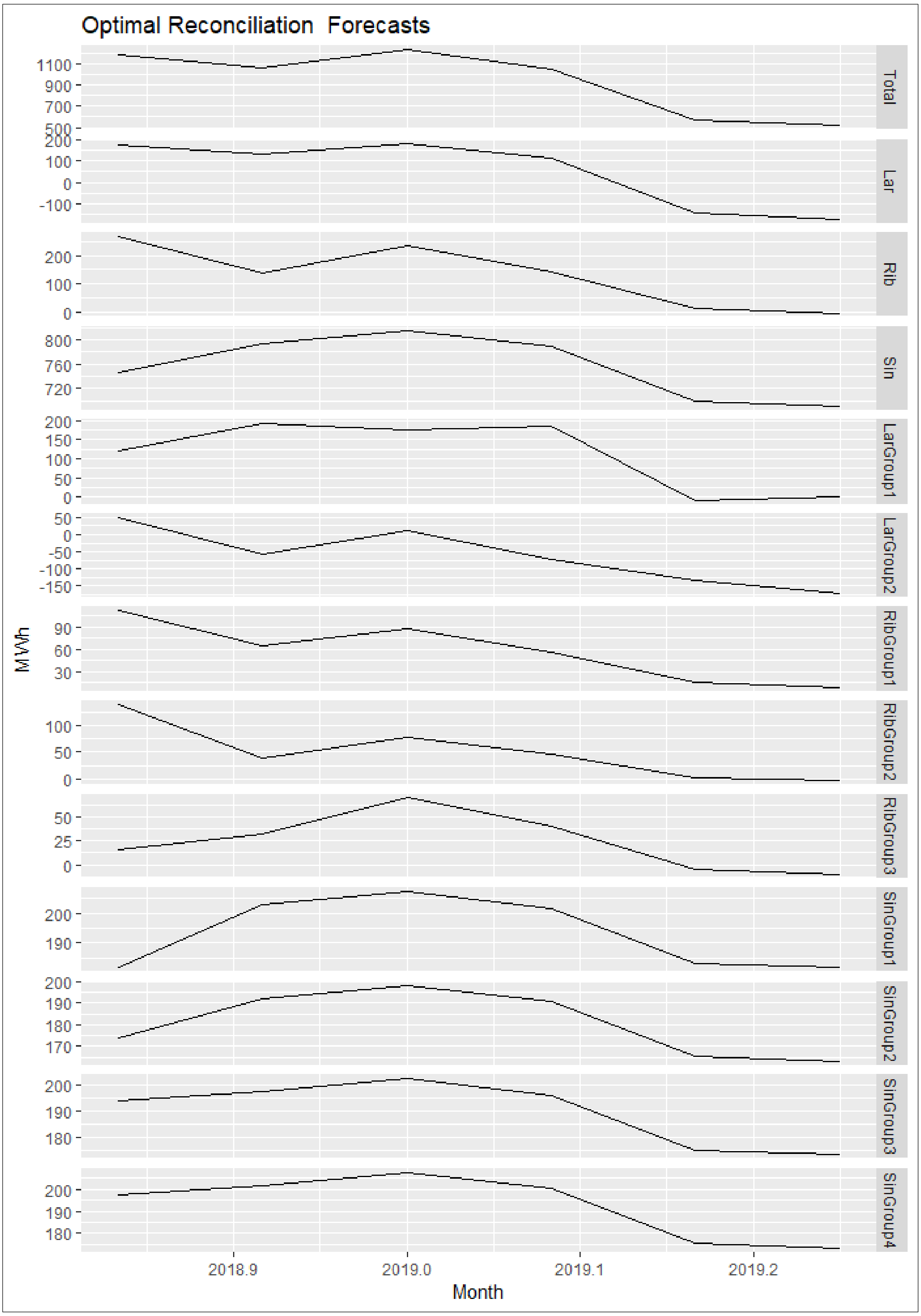


Figure 38 Line chart of optimal Reconciliation forecasting approach

4.4.2.3. Time Series Dynamic Regression Model

The applied time series regression in this work studies the effects of external factors on the total monthly production in the three thermal power plants, the data from the SAS EG table and the data from UNGE are collected in one table as total monthly values for every power plant and its sub-groups. The new table is imported into the global environment of R as a data set composed of 14 variables and 28 observations for each one of them.

Performing a dynamic regression model means extending Arima models to allow information other than past observations of a series to be involved in the models, and to allow errors from the regression models to include autocorrelation.

During the forecasting operation using a regression model, it is important to evaluate the model through evaluating the predictors that are involved in the regression model to choose the best ones. In R Studio, it is possible to measure the predictive accuracy through five measures where the model is better if the first four measures have lower values and the last measure has a higher value:

- CV, Cross-validation
- AIC, Akaike's Information Criterion
- AICc, Corrected Akaike's Information Criterion
- BIC, Schwarz's Bayesian Information Criterion
- AdjR2, Adjusted R2.

Based on (Hyndman & Athanasopoulos, 2018a), it is recommended to use one of CV, AIC or AICc statistics, since each one of them has forecasting as its objective.

- **Building Regression Model**

Building the regression model in this work follows the backwards stepwise regression approach introduced by (Hyndman & Athanasopoulos, 2018a) where all the potential predictors are used in the first model, and then removing one predictor with every iteration until the model shows no further improvements. However more models were tested in this research. Considering the existence of 13 predictor variables, there are $2^{13} = 8192$ possible regression model, the 13 predictors are listed in Table 2 with a description of each one.

Based on information from SKIPPER, there are a relation between Utilização (Tempo) and Utilização (Energia), and thus using only one of them in the regression model is enough. In the same way Disponibilidade (Tempo) and Disponibilidade (Energia) are related and one of them is only used in the model. Based on that, the model only has 11 predictor variables, and thus there are $2^{11} = 2048$ possible models.

Predictor Number	Predictor	Description	Time Series Name used in R
H. F	Horas de Funcionamento [#]	Working hours of the power plants;	hours_ts
U.T	Utilização (Tempo) [%]	Percentage of the actual working time and the possible working time;	usage_time_ts
U.E	Utilização (Energia) [%]	Percentage of the actual produced energy and the maximum energy that can be produced;	usage_energy_ts
N.A	Número de Arranques [#]	Monthly number of start-ups o the three power plants;	starts_ts
N.D	Número de Disparos [#]	Monthly number of stops o the three power plants;	stops_ts
D.T	Disponibilidade (Tempo) [%]	Percentage of the actual time when the power plants are available to work and the maximum possible available time;	availability_time_ts
D.E	Disponibilidade (Energia) [%]	Percentage of the available energy to use and the maximum possible available energy;	availability_energy_ts
C.P	Coal Price (€/Ton)	Price of coal measured by euro per ton;	Coal_ts
T.G	TFF Gas Price (€/MWh)	Price of Gas from the type TFF measured by euro per MWh	Tff_gas_ts
H.G	HH Gas Price (€/MWh)	Price of Gas from the type HH measured by euro per MWh	HH_gas_ts
H.C	Hydro lakes Capacity (MWh)	The capacity of energy production from the water stored behind the dams in the hydraulic plants	hydro_capacity_ts
T.R	Temperature (ºC)	Monthly average of temperature in Portugal	temperature_ts
CO2	CO2_price (€/Ton)	The price that EDP has to pay in exchange of CO2 emission.	Co2_price_ts

Table 2 Chosen predictors for the regression model

Every predictor variable is a time series, and in order to fit the right regression model the time series are turned into stationary time series, and the following code shows an example of the applied pre-processing and testing on the Stops time series. The used test is the Augmented Dickey-Fuller Test, and by looking at the p-value of the test output, it is possible to identify the time series as a stationary one if the value is less than 0.05. Since that the time series in question is not stationary, function of decomposing, seasonality adjustments and differences removing are applied.

```
# converting the Coal Price column into a time series
> stops_ts <- ts (Data.Table$ Numero_de_Disparosfrequency = 12, start = c (2016,7))

# Checking if the time series needs differencing
> adf.test (stops_ts, alternative = "stationary")
        Augmented Dickey-Fuller Test

data:  stops_ts
Dickey-Fuller = -3.0511, Lag order = 3, p-value = 0.1704
alternative hypothesis: stationary

# Decomposing, seasonality adjustments and differences removal
> stops_ts_diff <- diff (seasadj (stl (stops_ts, "periodic")))

# Checking if any more adjustments are needed
> adf.test (stops_ts_diff, alternative = "stationary")

        Augmented Dickey-Fuller Test

data:  stops_ts_diff
Dickey-Fuller = -4.6113, Lag order = 2, p-value = 0.01
alternative hypothesis: stationary
```

The experiments started by implementing a model that contains the all 11 predictor variables, as shown below, and then repeating the model with choosing one predictor or more to remove at the time, checking the values of AICc, and choosing which are the most important predictors for the production variable. The model was iterated 53 times as shown in Table 3, where every row presents one applied model, 0 means the predictor is not included, and 1 means that the predictor is included. The following R code shows the base regression model that includes the all 11 predictor time series after confirming that each one of them is a stationary series.

```
# Fitting a regression model for 11 predictors. Model number 1.

> fit_all_model <- auto.arima (production_ts_diff, xreg= hours_ts_diff
+ usage_time_ts_diff + stops_ts_diff + starts_ts_diff +
availablity_time_ts_diff + Coal_ts + TFF_Gas_ts_diff + HH_Gas_ts_diff
+ hydro_capacity_ts + temperature_ts + Co2_price_ts)
```

Model	Predicter Number												
	H.F	U.T	N.A	N.D	D.T	C.P	T.G	H.G	H.C	T.R	CO2	Arima model	AICc
1	1	1	1	1	1	1	1	1	1	1	1	(0,0,0)	133.68
2	1	0	0	0	0	0	0	0	0	0	0	(0,0,0)	234.74
3	0	1	0	0	0	0	0	0	0	0	0	(2,0,0)	241.78
4	0	0	1	0	0	0	0	0	0	0	0	(2,0,1)	346.96
5	0	0	0	1	0	0	0	0	0	0	0	(0,0,0)	354.46
6	0	0	0	0	1	0	0	0	0	0	0	(0,0,0)	241.78
7	0	0	0	0	0	1	0	0	0	0	0	(0,0,0)	354.43
8	0	0	0	0	0	0	1	0	0	0	0	(0,0,0)	302.81
9	0	0	0	0	0	0	0	1	0	0	0	(0,0,0)	350.52
10	0	0	0	0	0	0	0	0	1	0	0	(0,0,0)	352.85
11	0	0	0	0	0	0	0	0	0	1	0	(0,0,0)	354.51
12	0	0	0	0	0	0	0	0	0	0	1	(0,0,0)	182.76
13	0	0	1	1	0	0	0	0	0	0	0	(0,0,0)	354.49
14	1	1	1	1	0	0	0	0	0	0	0	(0,0,0)	235.78
15	0	0	0	0	0	1	1	1	0	0	1	(0,0,0)	131.19
16	0	0	0	0	0	0	0	0	1	1	0	(0,0,0)	352.85
17	0	0	0	0	0	0	0	0	1	1	1	(0,0,0)	177.88
18	0	1	0	0	1	0	0	0	0	0	0	(0,0,0)	241.78
19	0	1	1	1	1	0	0	0	0	0	0	(0,0,0)	321.72
20	1	1	1	1	1	0	0	0	0	0	0	(0,0,0)	235.82
21	1	0	1	1	0	1	1	1	0	0	0	(0,0,0)	202.43
22	1	0	1	1	0	1	1	1	0	0	1	(0,0,0)	89.22
23	1	0	0	1	0	1	1	1	0	0	1	(0,0,0)	89.3
24	1	0	1	0	0	1	1	1	0	0	1	(0,0,0)	89.4
25	0	0	1	0	0	1	1	1	0	0	1	(0,0,0)	131.81
26	0	0	1	1	0	1	1	1	0	0	1	(0,0,0)	133.65
27	1	0	1	1	0	1	1	0	0	0	1	(0,0,0)	97.89
28	1	0	1	1	0	1	0	1	0	0	1	(0,0,0)	124.12
29	1	0	1	1	0	0	1	1	0	0	1	(0,0,0)	90.6
30	1	0	1	1	0	1	0	0	0	0	1	(0,0,0)	124.13
31	1	0	0	0	0	1	1	1	0	0	1	(0,0,0)	89.51
32	1	0	0	0	0	0	0	0	0	0	1	(0,0,0)	122.9
33	1	1	1	1	1	1	1	1	1	1	0	(0,0,0)	304.81
34	1	1	1	1	1	1	1	1	1	0	0	(0,0,0)	304.81
35	1	1	1	1	1	1	1	1	0	0	0	(0,0,0)	206.06
36	1	1	1	1	1	1	1	0	0	0	0	(0,0,0)	202.07
37	1	1	1	1	1	1	0	0	0	0	0	(0,0,0)	236.72
38	1	1	1	1	1	0	0	0	0	0	0	(0,0,0)	235.82
39	1	1	1	1	0	0	0	0	0	0	0	(0,0,0)	235.78
40	1	1	1	0	0	0	0	0	0	0	0	(0,0,0)	234.81

41	1	1	0	0	0	0	0	0	0	0	0	(0,0,0)	234.82
42	1	1	1	1	1	1	0	1	1	1	1	(0,0,0)	182.91
43	1	1	1	1	1	1	1	1	1	1	0	(0,0,0)	304.81
44	1	1	1	1	1	1	1	1	1	0	1	(0,0,0)	133.68
45	1	1	1	1	1	1	1	1	0	1	1	(0,0,0)	93.72
46	1	1	1	1	1	1	1	0	1	1	1	(0,0,0)	144.68
48	1	1	1	1	1	0	1	1	1	1	1	(0,0,0)	141.56
49	1	1	1	1	0	1	1	1	1	1	1	(0,0,0)	133.68
50	1	1	1	0	1	1	1	1	1	1	1	(0,0,0)	100.58
51	1	1	0	1	1	1	1	1	1	1	1	(0,0,0)	105.45
52	1	0	1	1	1	1	1	1	1	1	1	(0,0,0)	133.68
53	0	1	1	1	1	1	1	1	1	1	1	(0,0,0)	140.25

Table 3 Applied regression models

4.5. Results and Discussion

Since the work in this research is divided into two main parts, and the first part is already implemented in the case company, every part will be discussed separately.

4.5.1. Reports Improvement

Comparing SAS with Power BI, it was very clear that SAS has more statistical power considering that SAS EG has a direct connection to the base Oracle data source and it is possible to process the data, transform it, and perform mathematical calculations which forms a stronger base in the reporting process. Furthermore, in the dashboard that was recreated in SAS, some of the miscalculations that existed in the Power BI dashboard were corrected, and based on the client feedback, DOT, SAS presents more interactivity with the available printing and extraction options. Besides that, SAS presents more reliability since logging into the dashboard from the SKIPPER Page requires the user to have a SAS account, however, SAS also presents the option of browsing the dashboard as guest user. Having these facts, Power BI was replaced in SKIPPER by SAS as the main reporting tool along with SAP BO, and the Power BI dashboard existed in SKIPPER Portal was replaced with the new dashboard designed by SAS VA permanently to be viewed by other EDP units.

4.5.2. Advanced Analytics

- **Forecasting Dashboard, SAS Forecasting Tool**

The dashboard built in SAS Data Explorer presents a line chart that gives a comprehensive idea about the thermal energy production, easy to read and offers the possibility to find out what could be done to meet a defined target. However, during the design of the dashboard it is not possible for the designer to change the used forecasting method or even the factors that are considered important to the production. Furthermore, when presented in the report viewer, the dashboard is limited regarding interactivity with the user, and the filtration options available for him.

- **Production Hierarchical Time Series Forecasting**

The production hierarchical time series is forecasted in four different methods in R Studio platform, and to compare the four methods and find out which one of them is more accurate and suitable for the data, an accuracy test is applied to evaluate the models.

Table 4 summarizes the accuracy of the proposed approaches, forecasting May 2018 to October 2018 as a testing period, these measurements are obtained by applying the function *accuracy ()* on every model of the four ones, the following code shows an example of applying the function on the bottom-up approach:

```
# Applying Bottom-up model on the training set, forecasting 6 months
> Bottom_up_production_training <- forecast (production_training_set, h = 6, method = "bu",
fmethod = "arima")
# fitting the previous model on the testing data set
> accuracy.gts (Bottom_up_production_training, production_testing_set)
```

Two measurements are chosen to compare the four models, MAPE and MASE, these measurements are more common since both measure the accuracy independently of the data scale. MAPE (Mean Absolute Percentage Error) uses a percentage of the error as the following:

$$MAPE = 100n^{-1}\sum_{t=1}^{n} {|y_t - f_t|}\Big/{|y_t|}$$

and MASE (Mean Absolute Scale Error) has a scale mathematical calculation implied in its formula as the following:

$$MASE = n^{-1} \sum_{t=1}^{n} |y_t - f_t| / q$$

Where y_t is the observed time series, f_t the forecasted time series of y, and q is a stable measure of the scale of the time series yt (Jeromy Anglim, n.d.).

Based on the results in the table below it is possible to see that the forecast with highest accuracy is the one based on the optimal reconciliation approach.

	Bottom-up		Top-Down		Middle-out		Optimal	
	MAPE	MASE	MAPE	MASE	MAPE	MASE	MAPE	MASE
Total	35.75	1.56	13.77	0.53	37.32	1.31	32.18	0.42
Lares	73.67	1.05	105.02	1.40	89.68	1.12	84.73	1.03
Ribatejo	47.32	1.42	33.80	0.94	29.87	0.94	25.34	1.57
Sines	41.38	1.96	15.00	0.68	8.58	0.41	8.68	0.52
Lares Group1	40.88	0.56	66.47	0.91	52.50	0.69	41.09	0.67
Lares Group2	556.50	1.38	740.05	1.38	683.41	1.25	568.07	1.41
Ribatejo Group1	98.49	1.74	98.62	1.74	98.40	1.74	73.62	1.59
Ribatejo Group2	67.35	1.58	103.37	1.63	92.40	1.51	61.46	1.34
Ribatejo Group3	69.35	1.51	101.37	1.46	82.40	1.60	63.56	1.08
Sines Group1	58.45	2.75	37.56	1.78	34.09	1.63	8.11	0.50
Sines Group2	13.59	0.41	32.85	1.00	37.24	1.12	14.31	0.57
Sines Group3	80.45	2.82	70.71	2.47	68.98	2.42	9.01	0.41
Sines Group4	14.86	0.66	37.08	1.61	37.80	1.64	14.01	0.69

TABLE 4 ACCURACY OF THERMAL PRODUCTION FORECASTS FOR DIFFERENT GROUPS OF SERIES

- **Time Series Dynamic Regression Model**

Moving forward to the applied regression model, and observing the models output obtained in Table 3, it appears that the best model output is obtained at row 22 when the following seven factors are included in the model:

- TFF Gas Price
- HH Gas Price
- Coal Price
- Co2 Fine
- Number of Stops
- Number of Starts
- Working Hours

However, taking a closer look at models between the numbers 21 and 32 can reveal that some of these 7 factors is more important than other, for example at line 21 when removing the factor of CO2 Fine time series, the value of AICc increased from 89.22 To 202.43. Furthermore, examining the models 23, 24 and 31 shows when removing Starts-up factor or Stops factor or even both factors the models' AICc values almost stay identical. Models 27, 28 and 29 also show a higher value of AICc when removing on factor of raw materials' price (gas and coal). Removing the factors of Working Hours, at the model 25, also increases the AICc value.

Based on the previous discussion, the most important factors that affect the thermal production are the following ones:

- TFF Gas Price
- HH Gas Price
- Coal Price
- Co2 Fine
- Working Hours

These final factors are in line with client's expectations and the removed factors improved the model accuracy.

Conclusions and Future Developments

Conclusions

This research is motivated by the increased importance of advanced analytics in business planning and decision-making processes, and thus the research proposed two main improvements to be implemented based on the current analytics maturity level of SKIPPER.

The first proposed approach was presenting SAS as new reporting platform and testing two SAS products (SAS Enterprise Guide and SAS Visual Analytics) regarding data pre-processing and dashboards deign. Comparing SAS with one of the already existing reporting platforms (Power BI) SAS showed a better statistical performance, higher level of analytics technologies and more user interactivity. Based on comparison results it was recommended that SKIPPER replaces Power BI with SAS as the main reporting tool along with SAP BO.

The second proposed approach is integrating the concept of advanced data analytics to complement reporting process by implementing two time series forecasting models and one regression model based on two years of historical data. The collected data and includes energy production of the three main thermal power plants in Portugal (Sines, Lares and Ribatejo), along with different factors' data to be tested regarding their affection on production process. The first forecasting model is presented using SAS Forecasting Tool of SAS Visual Analytics, the built dashboard presents long-term forecasting (six months) for thermal energy production, and defines automatically important factors that affect production process. The dashboard also helps in decision-making process by showing what need to be changed in order to meet a specific goal. The second production forecasting model is presented using RStudio, the model presents long-term forecasting as well, and considers production data as hierarchical time series composed of three levels (Total production, power plants, and sub-groups). Four different hierarchical time series forecasting approaches are tested (Bottom-up, Top-down, Middle-out, and Optimal Reconciliation), and the results showed that that the Optimal Reconciliation approach performed better than the others and delivered more accurate forecasts.

Furthermore, a regression model is implemented on R Studio to define the most important factors that affect thermal production, and thus help to create future strategies and identify the necessary actions. The model shows that the prices of the raw materials (gas and coal), working hours of the power plants, and CO2 fine are the most important predictors for the production. However, it is also recommended that SKIPPER considers the number of stops and start-ups of the power plants when defining a future strategy.

Comparing the differences of the two proposed forecasting models and the two utilized platforms in this research shows that the SAS model performs well, offers several options, and has an easy drag-and-drop interface, however it is limited form the parametrization point of view since it is not possible to control the structure of the data or the time series forecasting method that is being implemented. R Studio, on the other hand, offers more statistical and parametrization power, and unlimited options of testing different time series forecasting and regression models with several methods to check the accuracy and the validity of implemented models. Based on that, SAS is recommended to be used SAS by users that don't have much statistical experience considering the easy user interface, while R Studio is recommended for users with more statistical and analytical experience that are able to take full advantage of its analytical techniques and deliver the outcomes in a form that is more understandable for the clients.

Future Developments

The field of advanced analytics and forecasting is quite wide, and this research presented different models implemented within two different platforms. Since R Studio already offers the possibility of reading SAS dataset using a specific library, this research suggests integrating R Studio into the architecture of SKIPPER, training a team of statistical expert to use it and develop a new direct connection that enables R to access the LASR server of SAS, which would help to raise the level of analytics maturity of SKIPPER.

Bibliography

Aldeeb, A. (2017, March 5). Introduction to SAP BusinessObjects BI Suite - Mindtory. Retrieved January 13, 2019, from https://www.mindtory.com/introduction-sap-businessobjects-bi-suite/

As’ad, M. (2012). *Finding the Best ARIMA Model to Forecast Daily Peak Electricity Demand*. University of Wollongong, Australia.

Lans, R. van der. (2016, October 20). Developing a Bimodal Logical Data Warehouse Architecture Using Data Virtualization. Retrieved January 13, 2019, from https://www.denodo.com/en/document/whitepaper/developing-bimodal-logical-data-warehouse-architecture-using-data-virtualization

EDP. (n.d.-a). EDP | about us. Retrieved from https://www.edp.com/en/aboutedp

EDP. (n.d.-b). Energy Sector - Distribution. Retrieved from https://www.edp.com/en/edp/energy-sector/distribution

EDP. (n.d.-c). Generation and Supply | EDP Portugal. Retrieved from https://portugal.edp.com/en/energy-sector-in-portugal/generation-and-supply

EDP. (n.d.-d). Volunteering | EDP Portugal. Retrieved December 16, 2018, from https://portugal.edp.com/en/a-stronger-more-supportive-and-more-innovative-society/volunteering

EGNELL, J., & HANSSON, L. (2013). *Applicability and accuracy of quantitative forecasting models applied in actual firms A case study at The Company*. CHALMERS UNIVERSITY OF TECHNOLOGY, Göteborg - Swedn.

Gartner Survey Shows Organizations Are Slow to Advance in Data and Analytics. (2018, February 5). Retrieved January 13, 2019, from https://www.gartner.com/en/newsroom/press-releases/2018-02-05-gartner-survey-shows-organizations-are-slow-to-advance-in-data-and-analytics

Ghalehkhondabi, I., Ardjmand, E., Weckman, G. R., & Young, W. A. (2016). An overview of energy demand forecasting methods published in 2005–2015. *Energy Systems*, *8*(2), 411–447. https://doi.org/10.1007/s12667-016-0203-y

Gvaladze, S. (2015). *Evaluating methods for time-series forecasting; Applied to energy consumption predictions for Hvaler (kommune)*. Østfold University College, Halden , Norway.

Hagerty, H. (2016). *2017 Planning Guide for Data and Analytics* (No. G00311517) (p. 27). Gartner. Retrieved from https://www.gartner.com/binaries/content/assets/events/keywords/catalyst/catus8/2017_planning_guide_for_data_analytics.pdf

Hyndman, R. J. (2013, July 3). R tools for hierarchical time series | Rob J Hyndman. Retrieved January 13, 2019, from https://robjhyndman.com/seminars/hts-2/

Hyndman, R. J., & Athanasopoulos, G. (2018a). *Forecasting: principles and practice ; [a comprehensive indtroduction to the latest forecasting methods using R ; learn to improve your forecast accuracy using dozenss of real data examples.*

Hyndman, R. J., & Athanasopoulos, G. (2018b). *Forecasting: principles and practice ; [a comprehensive indtroduction to the latest forecasting methods using R ; learn to improve your forecast accuracy using dozenss of real data examples.*

Hyndman, R. J., Athanasopoulos, G., & Shang, H. L. (n.d.). *hts: An R Package for Forecasting Hierarchical or Grouped Time Series* (p. 5). Retrieved from https://cran.r-project.org/web/packages/hts/vignettes/hts.pdf

Jeromy Anglim. (n.d.). *Forecasting time series using R by Prof Rob J Hyndman at Melbourne R Users*. Retrieved from https://www.youtube.com/watch?v=1Lh1HlBUf8k

Kumar, M. (2017, March 1). Why Businesses Need Advanced Analytics Today. Retrieved from https://www.salesforce.com/blog/2017/02/why-businesses-need-advanced-analytics

Manual, N., Wiseman, B., & Zecha, C. M. (2016). How advanced analytics can drive productivity | McKinsey. Retrieved from https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/how-advanced-analytics-can-drive-productivity

Martín, L., Zarzalejo, L. F., Polo, J., Navarro, A., Marchante, R., & Cony, M. (2010). Prediction of global solar irradiance based on time series analysis: Application to solar thermal power plants energy production planning. *Solar Energy*, *84*(10), 1772–1781. https://doi.org/10.1016/j.solener.2010.07.002

Moore, S. (2018, December 6). Take Your Analytics Maturity to the Next Level. Retrieved January 13, 2019, from https://www.gartner.com/smarterwithgartner/take-your-analytics-maturity-to-the-next-level/

R Studio. (2012, July 31). Why R Studio? Retrieved January 13, 2019, from https://www.rstudio.com/about/

RapidMiner. (n.d.). Advanced Analytics vs Business Intelligence. Retrieved December 16, 2018, from https://rapidminer.com/glossary/advanced-analytics-vs-bi/

Rodrigues, Á., Lopes, J. A., Miranda, P., Laginha Palma, J., Monteiro, C., N Sousa, J., … Matos, J. (2007). *EPREV - A wind power forecasting tool for Portugal* (Vol. 2).

Roecker, S., Yoast, K., Wills, S., & D’avello, T. (n.d.). Chapter 1 Introduction to R and RStudio. Retrieved January 13, 2019, from http://ncss-tech.github.io/stats_for_soil_survey/chapters/1_introduction/1_introduction.html

SAS. (n.d.). Big data analytics: What it is and why it matters. Retrieved December 16, 2018, from https://www.sas.com/pt_pt/insights/analytics/big-data-analytics.html

Shumway, R. H., & Stoffer, D. S. (2017). *Time series analysis and its applications: with R examples* (Fourth edition). Cham: Springer.

Structuring a business report | Oxford Dictionaries. (n.d.). Retrieved January 13, 2019, from https://en.oxforddictionaries.com/writing-help/structuring-a-business-report

What is R? (n.d.). Retrieved January 13, 2019, from https://www.r-project.org/about.html

What Is SAS Visual Analytics Explorer? (n.d.). Retrieved January 13, 2019, from http://support.sas.com/documentation/cdl/en/vaug/66223/HTML/default/viewer.htm#n0xushcurzq19vn1m5b86bp2zule.htm