UNIVERSITY OF LJUBLJANA
SCHOOL OF ECONOMICS AND BUSINESS

MASTER’S THESIS

ENERGY PRODUCTION MIX IN THE EU: A MACHINE LEARNING AND DATA MINING ANALYSIS

Ljubljana, October 2020 MARCO STAHLHACKE

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

aBIC – Adjusted Bayesian Information Criterion

ANN – Artificial Neural Network

BIC – Bayesian Information Criterion

BPTT – Backpropagation Through Time

BRP – Balance Responsible Parties

CO_2 – Carbon Dioxide

DM – Data Mining

DR – Demand Response

EM – Expectation Maximization

ENTSO-E – European Network of Transmission System Operators

EU – European Union

EU-28 – 28 Member States of the European Union

GDP – Gross Domestic Product

GHG – Greenhouse Gases

GMM – Gaussian Mixture Models

GT – Billion Tons

IDE – Integrated Development Environment

KDD – Knowledge Discovery in Databases

kWh – Kilowatt-Hours

LSTM – Long Short-Term Memory

ML – Machine Learning

MLE – Maximation Likelihood Estimation

MLP – Multilayered Perceptron

MSE – Mean Squared Error

Mtoe – Million Tons Of Oil Equivalent

mWh – Megawatt-Hours

TWh – Terawatt-Hours

RNN – Recurrent Neural Networks

SARIMA – Seasonal Auto-Regressive Integrated Moving Average

SVM – Support Vector Machine

TSO – Transmission-System Operators

UNFCCC – United Nations Framework Convention on Climate Change

V – Volt

W – Watt

1 INTRODUCTION

1.1 Background and Problem Identification

Climate change is a threat to the earth's ecosystem. This phenomenon is driven by natural as well as human forces. Anthropogenic contributions to climate change increased steadily since the pre-industrial era. This resulted in greenhouse gas (GHG) emissions reaching the highest point in the recent human history. As a consequence, the high concentration of GHG in the atmosphere contributes to rising ocean and surface temperatures, melting of ice covers, rising of average sea levels, the occurrence of extreme weather and climate events (IPCC, 2014).

The main drivers of anthropogenic GHG emissions are "population size, economic activity, lifestyle, energy use, land use patterns, technology and climate policy" (IPCC, 2014, p. 8). Without any action on mitigating the emissions of GHG more extreme and irreversible events will impact the ecosystem and humanity (IPCC, 2014).

Looking at recent statistics, it can be observed that 37,1 billion tons (Gt) of carbon dioxide (CO_2), one of the main GHG, were emitted globally in 2017 (Muntean et al., 2018). The member states of the European Union (EU-28) contributed 9,6% (3,5 Gt) to the total emissions (Muntean et al., 2018). The largest contributors within the EU-28 were Germany (22.4%), "the United Kingdom (10.7%), Italy (10.2%), France (9.5%) and Poland (9%)" (Muntean et al., 2018, p. 10). Moreover, the energy sector represents a key role in the CO_2 production. Globally, nearly 15 Gt of CO_2 emissions were emitted by the power generation sector (Muntean et al., 2018). This is why the energy sector of the European Union (EU) will undergo detailed examination in this thesis.

While most residents of developed regions, like North America, Europe or Japan, show a high level of awareness for the climate change in developing countries people are not aware or did not even hear about climate change (Lee, Markowitz, Howe, Ko, & Leiserowitz, 2015). Still, according to Lee, Markowitz, Howe, Ko, and Leiserowitz (2015) citizens from less developed parts of the world, who are aware of the climate change, perceive it as a bigger threat than citizens of developed countries.

Even if public awareness varies a lot across the world, politicians identified this as a threat to global wellness. The 2015 "Paris agreement" was defined as a part of the United Nations Framework Convention on Climate Change (UNFCCC). In article 2 1.a this agreement defines the following long-term goal: "Holding the increase in the global average temperature to well below 2°C above pre-industrial levels and pursuing efforts to limit the temperature increase to 1.5°C above pre-industrial levels, recognizing that this would significantly reduce the risks and impacts of climate change" (United Nations, 2015, p. 3). The convention was ratified by 185 out of 197 nations (United Nations, n.d.).

On a smaller scale, also the European Union developed agreements dealing with the challenges pose by climate change. Since the energy sector can be identified as a major contributor, there are three EU strategies that need attention: the "2020 Energy Strategy", the "2030 Climate and Energy Framework" and the "Energy Roadmap 2050".

In 2014 the member states of the European Union (EU) mapped out the "2030 Climate and Energy Framework" (European Council, 2014). This strategy follows on the "Energy 2020" strategy from 2009 (European Commission, 2010). Together with the "2050 Energy Strategy" the European Union currently follows three strategies, which set short- and long-term goals on reduction of GHG emissions, energy consumption, share of renewable energy or on energy efficiency. These agreements are a necessary step to reduce the impact of the EU-28 energy sector on climate change (European Council, 2014). Along with other sectors, within the EU the sector with the highest contribution to GHG emissions is again the energy sector. This industry emitted nearly 80% of EU's greenhouse gases when the "Energy 2020" strategy was mapped out (European Commission, 2010).

But even with these agreements passing, criticism over the commitment that governments have shown for environmental questions rose. Moreover, some political decisions resulted in new discussions and conflicts within the EU-28. Also, natural or man-made disasters related to the energy sector and climate change, show the threats some energy carriers pose, are influencing opinions and lead to new discussions.

This can be observed i.e. by looking at the two major incidents that involved nuclear power plants. These incidents illustrate the potential devastation provoked by nuclear power production. In some countries, this has led to a rethink of suitable energy sources and initiated policies that move away from this energy carrier. A first major disaster in the Chernobyl power plant is dated back to 1986, while the more recent one happened 35 years later in Japan (Funabashi & Kitazawa, 2012; Yablokov, Nesterenko, & Nesterenko, 2009).

The disaster in Japan's Fukushima Daiichi power plant was triggered by a Tsunami that was the result of a 9.0 magnitude earthquake in the Pacific sea on March 11, 2011. In both cases meltdowns of the energy cores were a consequence and led to the release of nuclear material in the environment with far-reaching consequences (Funabashi & Kitazawa, 2012; Yablokov, Nesterenko, & Nesterenko, 2009). In Europe the incident of Fukushima led to a rethink of the energy mix. Germany for example showed an instant reaction to the disaster. On March 15 seven nuclear reactors were shut down temporarily and in June a law was passed that regulates a nuclear power phaseout by 2022 (Hake, Fischer, Venghaus, & Weckenbrock, 2015).

But also, natural disasters forced actions towards a change in the energy sector. In 2018, heat waves and forest fires in Sweden were the reason a movement for actions against climate change started. After the natural disasters, the Swedish Greta Thunberg was the initiator of the "School Strike for Climate" movement (The Economist, 2019). In August 2018, instead of attending school, the 15-year-old decided to sit in front of the Swedish parliament with a sign that said “Skolstrejk för klimatet”, which means “school strike for climate” (The Economist, 2019). The objective of this protest was to raise awareness around climate change and demand firm actions to prevent further human-induced climate alteration. Her protest was deemed to conclude only when the polices of the Swedish government would line with the climate goals of the Paris agreement (The Economist, 2019). However, within the next few months, children in many other cities around the globe became aware of Greta Thunberg’s actions and organized school strikes each Friday (The Economist, 2019). By doing so, a movement that attracts international attention has begun.

Not only actions taken by citizens, but also the ones initiated by governments show that economic and political interests have a big influence on the energy politics. E.g. the “Nord Stream 2” project with the purpose of building a new gas pipeline from Russia to Germany in the North Sea is does not in any way align with the renewable energy carriers project. Still, Nord Stream 2 might currently be one of the most criticized energy projects due to many different interests of several countries and the EU (The Economist, 2018).

As illustrated above, the energy sector and its politics are a complex subject. Many different factors had an influence on the development of the energy sector in the past and will have an influence in the future. Therefore, taking more recent events into consideration is reasonable since they might also be an indicator of recent and possible future changes in the energy policy of EU member states.

1.2 Study Objectives and Relevance

Taking into consideration that recent energy related events may have led and still lead to changes in the energy policies of EU member states and that the EU developed encompassing strategies up to the year 2050, the European energy sector and the possibility of reaching the targets set should undergo special analysis.

The main purpose of this thesis is hereby to analyze historical and possible future developments in the EU’s energy sector with the help of data mining and machine learning techniques. The analysis will mainly focus on the energy production mix and will lay a special focus on the development of renewable energies.

Therefore, machine learning methods are applied on historical data to create forecasts and in the end acquire a wide-reaching projection until 2050. This projection would represent the foundation for the analysis of the European Union’s energy production mix. The major task is then to apply data mining methods to the data to carry out a final analysis on the development of the energy production.

This analysis aims to answer the following research questions:

With a special focus on renewables, are there any noticeable patterns in the behavior, in the stability and in the composition of member states' energy profiles? And moreover, is there a recognizable tendency towards a certain energy production mix or energy carrier?

Finally, when comparing the findings of the analysis to the energy strategies mapped out by the European Union, what conclusions can be drawn on the feasibility of these strategies?

1.3 Structure of the Paper

As an introduction the different theoretical topics are characterized and specified more into detail in an initial literature review in chapter 2. At first, a general understanding of the European energy system is imparted to the reader. A primary focus hereby lays on the fundamental mechanics and relationships between the parties involved in the EU's energy market and the energy strategies decided by the EU. Due to a strong relationship to the future development of the European energy production, the challenges of the energy transition towards renewable energy we are facing right now are examined more in depth afterwards. As certain energy incidents, not only in Europe, affected and will still affect the actions in the energy market, a short overview on some of the major events is given consequently.

The second main subject of the literature review gives further insights on the data mining and machine learning tools and technologies and their previous applications in the energy sector. After defining the relationship of data mining and machine learning, both fields are described more precisely with a focus on forecasting methods and clustering algorithms. Hereby, artificial neural networks (ANN) and the expectation maximization (EM) algorithm are of particular importance. To conclude, previous applications of such methods in the energy sector are summarized.

These energy sector insights combined with the comprehension of data mining and machine learning techniques will enable the reader to comprehend the approach of the following analysis of the European energy production in chapter 3. The analysis is based on a methodological approach that is defined precisely by a framework tailored to the needs of the present problem. Following this framework, in the subsequent chapter the dataset is examined and processed to fulfill the needs of the analysis. This allows for the subsequent application of a forecasting and a clustering algorithm. The emerged results are then part of an in-depth discussion and analysis in order to answer the research questions of this thesis. In the final lines of the paper conclusions are drawn and the insights gained through this investigation will be highlighted.

2 LITERATURE REVIEW

2.1 The European Union's Energy Market

To comprehend the following challenges of the energy transition in chapter 2.2 it is essential to get an overview of the European energy sector first. Therefore, some fundamental mechanics and relationships of national grids are described in the following sections. Afterwards an overview is given on the strategies and polices that are developed by the EU and on the goals set for the member states and the EU as a whole.

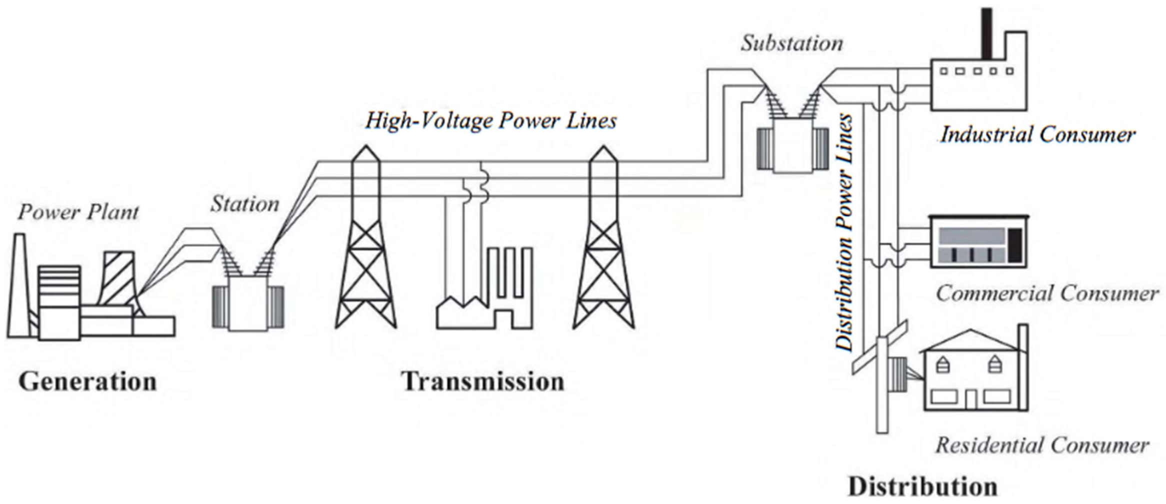
2.1.1 Mechanics and Relationships

To understand the mechanics and relationships of the European energy market, it is necessary to have a basic understanding on how a national power grid and the energy market operate. Therefore, the reader is introduced to the thematic by providing an overview of a common energy power system, the fundamental measurements in the energy sector, the structure of market participants and pricing mechanisms.

Figure 1 represents the basic concept of the classic electrical power systems. The system is a composition of three main building blocks, which are the electricity generator, the transmission network and the distribution network.

To get further insights in the building blocks and the energy sector it is first necessary to understand the basic measurements used in the energy field. The fundamental measurements and their units to express energy related values that will be needed in this paper are voltage, power and energy.

Figure 1: Energy Power System Building Blocks



Source: Blume (2017).

Voltage describes the potential energy that is flowing through the system and always occurs between two points (Blume, 2017). It is a force, that is comparable to the pressure in a water pipe (Blume, 2017). The unit to measure Voltage is defined as a Volt (V). For different applications power systems use different voltages reaching from 120 Volt low voltage to 765.000 Volt ultra-high voltages (Blume, 2017). Power, which is measured in the unit Watt (W), describes the real work that can be produced with the energy. Electrical power can be used e.g. "to create heat, spin motors, light lamps, etc." (Blume, 2017, p. 6).

The last fundamental unit describes the electrical energy, that "is the product of electrical power and time" (Blume, 2017, p. 6). Energy is defined as the product of the time a load is flowing and the power that is used during that time (Blume, 2017). The unit therefore is defined as watt-hours. More common measurements are kilowatt-hours (kWh) for private households and megawatt-hours (mWh) for large industrial and the power companies themselves. Specifically, one kWh is equal to one thousand and one mWh to one million watt-hours (Blume, 2017).

After knowing the fundamental measurements, a closer look can be taken at the building blocks of an energy power system. The first building block represents the electricity generators that are responsible to produce electrical energy. The generators exist in different types and sizes where each of them has specific attributes (Erbach, 2016). Generator sizes might vary e.g. from single solar panels, to wind farms and large-scale coal or nuclear energy plants.

Table 1 illustrates the main energy-generation technologies and their characteristics. The energy types can be differentiated between renewable (solar, wind, biomass, geothermal) and conventional (hydro, coal, oil, natural gas and nuclear) energy sources. The characteristics are the variability, the type of fuel, the degree of flexibility and the contribution to GHG emissions. These characteristics will be of utter importance to comprehend the following chapters and the final analysis of this paper.

Table 1: Characteristics of the Main Energy Types

Type	Firm / Variable	Type of Fuel	Flexibility	GHG Emissions
Coal	firm	fossil	medium	Yes
Oil	firm	fossil	high	Yes
Natural Gas	firm	fossil	high	Yes
Biomass	firm	renewable	medium	Emission compensation through biomass regrowth
Nuclear	firm	nuclear	low	Zero Emissions
Hydro	firm	renewable	very high	
Solar	variable	renewable	very low	
Wind	variable	renewable	very low	
Geothermal	firm	renewable	high	

Source: Adapted from Erbach and Stram.

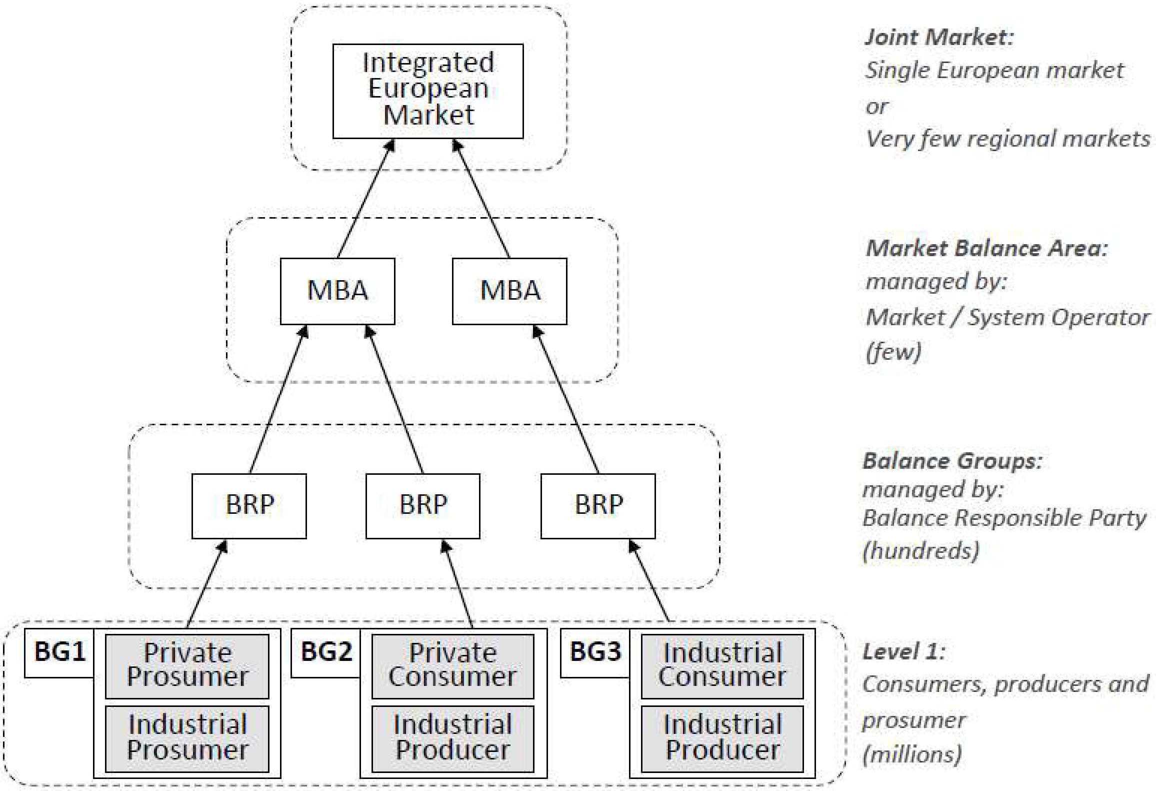
The second building block are the transmission lines, which provide an infrastructure to efficiently transport the produced electric energy. In Europe more than 300.000 km of power lines form the transmission grid (Erbach, 2016). To transport the electricity, high voltages are used for the simple reason, that an increasing voltage leads to a significant reduction of transmission losses (Blume, 2017; Erbach, 2016).

The final building block is the distribution network. The main task of a distribution network is to distribute the electric energy to industrial, commercial and residential consumers (Blume, 2017). For this purpose substations transform the energy into useable, lower voltages (Blume, 2017). Moreover, the network usually involves energy that is generated by smaller renewable suppliers with solar or wind systems. The distribution networks are operated by distribution-system operators (DSO) (Erbach, 2016). They are responsible to "connect consumers, install electricity meters and communicate the consumption to the energy suppliers" (Erbach, 2016, p. 4).

If we take a look not only at a single energy power system, but also at the national and European energy market, it becomes clear that the market is organized hierarchically (see Figure 2). On the lowest level the producers and consumers are representing balance groups, where several balance groups represent a market balance area. Since the energy supply and demand always need to be balanced, balance responsible parties (BRP) represent the superior level in the hierarchy. The BRPs are responsible to balance supply and demand within their group. If the energy is exceeding within this group they are able to trade exceeding energy to other groups or in the opposite case buy necessary energy to compensate a shortage (Dannecker, 2015). To manage and control the areas and transmission grids, transmission-system operators (TSO) often function as a market or system operator as well (Dannecker, 2015). On a European level, the responsible operators are organized in the European Network of Transmission System Operators (ENTSO-E) (Dannecker, 2015; Erbach, 2016). It develops plans and rules for the European grid network, that included 355 cross-border lines in 2015 (Erbach, 2016). Hence, the interconnection of different national grids allows TSOs also to trade energy to other countries if it is necessary to balance supply and demand internationally (Dorsman, Westerman, Karan, & Arslan, 2011).

Within the EU-28 different markets exist to trade electricity between the market participants. The markets can generally be divided in the retail and the wholesale market. On the retail market electricity contracts between energy suppliers and consumers are concluded (Erbach, 2016). The wholesale market instead is responsible for managing the trading of electricity between electricity generators, suppliers and larger industry consumers (Erbach, 2016). In the wholesale market the parties trade electricity in advance, since the electricity must be used at the point of production (Dorsman, Westerman, Karan, & Arslan, 2011). Within the wholesale market a distinction between four sub-markets can be made by considering their timescale.

Figure 2: Hierarchical Organization of the European Electricity Market



Source: Dannecker (2015).

In the forward and future market, the time period between trade and delivery can amount from weeks to years (Erbach, 2016). Trading with a next-day delivery is carried out in the day-ahead market. Market participants can purchase hourly or 24-hour block contracts and the TSO develops an operating schedule based on the transactions (Dorsman, Westerman, Karan, & Arslan, 2011). Still, it is unavoidable that the actual physical delivery and the demand settled in the contracts differ (Dorsman, Westerman, Karan, & Arslan, 2011). This results in the need of an intra-day market which would allow to conduct short-term trades (Erbach, 2016). In the last sub-market, the balancing market, the TSO is responsible to regulate the real-time supply or demand imbalances (Dorsman, Westerman, Karan, & Arslan, 2011). By allowing market participants to bid on prices, they can increase or decrease the electricity generation and consumption (Dorsman, Westerman, Karan, & Arslan, 2011).

The wholesale and retail market also involve different pricing policies. As mentioned afore, in the wholesale market energy is traded from the generators to large industrial customers and energy retailers, who subsequently distribute the electricity to private households in the retail market (Dutta & Mitra, 2017). This is the reason why the electricity prices vary between private household and industrial consumers. The electricity prices are defined by the balancing of supply and demand through generators who "offer bids for a certain amount of power at a certain price" (Kirschen, Strbac, Cumperayot, & Paiva Mendes, 2000, p. 613). These bids are ordered by their price. Following this order, retailers and large industrial customers place their bids to satisfy their demand (Kirschen, Strbac, Cumperayot, & Paiva Mendes, 2000). If this is the case, the market price for the specific time frame can be determined. It is set to the last bid that was accepted from the pool of bid prices (Kirschen, Strbac, Cumperayot, & Paiva Mendes, 2000). According to Kirschen, Strbac, Cumperayot, and Paiva Mendes (2000) this process of price determination can also be influenced by different design decisions in the market.

As Dutta and Mitra (2017) state the retail markets usually utilize static pricing policies, like a flat or block model. While electricity is offered for a fixed price to the customer in the flat model, in a block model prices customers are classified in tiers based on their consumption (Dutta & Mitra, 2017). Herby, the models are static since the pricing mechanisms ignore changes in demand and therefore the prices of the market (Dutta & Mitra, 2017).

2.1.2 Strategies and Policies

About a decade ago, the EU recognized the energy transition as one of the biggest challenges Europe must face. From 2010 on the EU developed several new strategies for the energy sector. These strategies were developed for different periods. As a result, strategies for the years 2020, 2030 and 2050 were defined. To get a better understanding of what the ambitious goals are, each strategy is described more into detail in this chapter. The focus is hereby laid on the main objectives of the strategies which are of particular relevance for this paper and the subsequent analysis.

2.1.2.1 2020 Energy Strategy

Back in 2007 the first energy and climate targets for 2020 were adopted by the European Council. These targets were later included into the more detailed 2020 energy strategy, which was approved by the European Commission in 2010. The 2020 energy strategy defined a roadmap to be followed by all member states for the next decade. Compared to the targets of 2007, the focus hereby was extended to define a strategy that ensures an energy sector that is competitive, secure and sustainable (European Commission, 2010).

The European Commission then introduced, on the foundation of the previous objectives, specific target levels for relevant energy and climate specific fields in the 2020 energy strategy. In numbers, the EU aims on an overall minimum share of renewables of 20% in 2020. Precisely, target values on the renewable shares of each member state are specified in the DIRECTIVE 2009/28/EC and are illustrated in Appendix 2 (Official Journal of the European Union, 2009). In addition, the EU set as further targets the reduction of GHG emissions by at least 20% together with an increase of the energy efficiency by 20%.

Along with these specific targets, the European Commission (2010) defined and elaborated the following general priorities:

1. Achieving an energy efficient Europe;

2. Building a truly pan-European integrated energy market;

3. Empowering consumers and achieving the highest level of safety and security;

4. Extending Europe's leadership in energy technology and innovation;

5. Strengthening the external dimension of the EU energy market (pp. 5-6).

2.1.2.2 2030 Energy Strategy

The 2030 climate and energy framework defines the subsequent roadmap for the EU-28 for the decade that follows 2020. An agreement on this strategy was approved on 23. October 2014 by the European Council. Since the strategy lasts until 2030, the targets are more ambitious and expand the objectives of the 2020 energy strategy (European Council, 2014). More specifically, the strategy pursues the reduction of the GHG emissions by 40% compared to the levels of 1990 (European Council, 2014). Also, the share of renewables should be increased to at least 27% of consumption by 2030 (European Council, 2014). The strategy addresses the energy efficiency as well. In this regard, it sets a goal to improve the energy efficiency in the EU by 27% in comparison to conducted projections of the future energy efficiency in 2030 (European Council, 2014). The strategy explicitly states that the member states are still free to choose their preferred energy mix and are encouraged to set national targets even higher than the targets prescribed by the strategy (European Council, 2014).

In 2018, the EU revised the targets of the renewable share and the improvement of the energy efficiency defined in the initial 2030 energy strategy and increased them with the renewable energy directive 2018/2001/EU and the Directive on Energy Efficiency (2018/2002). From then on the EU is targeting a 32% share of renewables and a 32,5% improvement in energy efficiency until 2030 (Official Journal of the European Union, 2018a, 2018b). Moreover, a clause was included that defines the possibility of another upwards adjustment for the targets until 2023 (Official Journal of the European Union, 2018a, 2018b).

The incorporation of the national energy markets into one fully functioning European market remains a target that is promptly pursued by the EU. For this purpose, the EU-28 has set the target to an establishment of 15% of existing electricity interconnections until 2030 (European Council, 2014).

The strategy addresses the need for a high energy security and a lower dependency on gas and electricity as well (European Council, 2014). In this context, the European Council sees the need of a higher energy efficiency, access to indigenous resources and implementation of low carbon technologies as drivers to achieve the above target (European Council, 2014).

To ensure flexibility and reduce administrative burdens for the member states in developing their energy mix the EU also has announced the development a “reliable and transparent governance system” (European Council, 2014, p. 9) that ensures the accomplishment of the energy related goals.

2.1.2.3 2050 Energy Strategy

On 15. December 2011 the European Commission has endorsed the 2050 Energy Roadmap that represents the EU’s long-term energy strategy. Unlike the other two strategies, the 2050 roadmap presents a different approach due to its long-term nature. Therefore, the only specific target value defined is the reduction of GHG emissions by 80-95% in comparison with the levels of 1990. The strategy additionally clarifies that a successful energy transition requires urgent major investments in the energy sector since this will need years to produce results (European Commission, 2011).

In addition to the target definition, the paper examines different scenarios that ensure the energy security and competitiveness of the European energy market in a scenario where the energy mix varies in order to reduce the GHG emissions (European Commission, 2011). It is stated that it is possible to implement a European energy system until 2050, that provides energy security, clean energy and is a credible competitor on the market (European Commission, 2011). This approach is not meant to replace national policies but is rather meant to support them on a larger scale. By developing this framework, the EU strives to make these policies more effective (European Commission, 2011). As a result the document states that the European approach will result in increased “security and solidarity and lower costs compared to parallel national schemes” (European Commission, 2011, p. 3).

In December 2019, the European Council approved the additional objective of a climate-neutral EU until 2050 due to recent scientific developments (European Council, 2019). This target setting complies with the goals of the Paris Agreement and enhances the objective of the 2050 Energy Roadmap (European Council, 2019).

2.1.2.4 Energy Union

In strong relation to its strategies, the EU has developed a concept of an energy union. In 2014 this concept has been already mentioned in the 2030 energy strategy. In the following year on February 25th the European Commission adopted a framework strategy for a resilient energy union (European Commission, 2015).

The energy union target is built on the basis of incentives to be given to “EU consumers - households and businesses - secure, sustainable, competitive and affordable energy” (European Commission, 2015, p. 2). To achieve this objective, the European Commission (2015) constructed a framework based on the following dimensions:

- Energy security, solidarity and trust;

- A fully integrated European energy market;

- Energy efficiency contributing to moderation of demand;

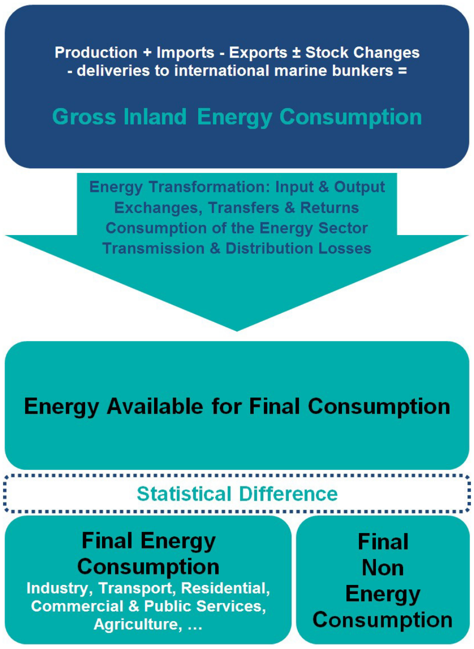
- Decarbonising the economy, and

- Research, Innovation and Competitiveness (p.4).

2.1.2.5 Energy Indicators

To measure and control the energy sector the EU developed various indicators over time. To measure the energy consumption of its member states, the European Union introduced several indicators. Figure 3 illustrates these indicators which are also part of the Eurostat's energy balance.

Figure 3: Energy Indicators in the EU


Production + Imports - Exports ± Stock Changes
- deliveries to international marine bunkers =
Gross Inland Energy Consumption
Energy Transformation: Input & Output
Exchanges, Transfers & Returns
Consumption of the Energy Sector
Transmission & Distribution Losses
Energy Available for Final Consumption
Statistical Difference
Final Energy Consumption
Industry, Transport, Residential, Commercial & Public Services, Agriculture, ...
Final Non Energy Consumption

Source: European Commission (2019).

While the gross inland energy consumption and the energy available for final consumption are computed through a top-down approach, the final energy consumption and final non energy consumption are determined through a bottom-up approach (European Commission, 2019). Top-down approaches are calculated based on the production. Bottom-up approaches instead originate from the actual consumption (European Commission, 2019). Hence, top-down energy indicators represent the supply side rather than the consumer side.

An indicator the illustration does not consider, and which should not be omitted in the scope of this paper, is the share of renewable energy in gross final energy consumption. This was introduced to measure the EU-28's progress in the energy transition and is the indicator the target values of the renewable share in the EU strategies refer to. The indicator is based on the gross final energy consumption (European Commission, 2019).

Its underlying indicator was first defined in the Official Journal of the European Union (2009) in Article 2 (f) of Directive 2009/28/EC as:

'gross final energy consumption' refers to the energy commodities delivered for energy purposes to industry, transport, households, services including public services, agriculture, forestry and fisheries, including the consumption of electricity and heat by the energy branch for electricity and heat production, and including losses of electricity and heat in distribution and transmission; (p. 27)

The directive the European Parliament and Council also defines the previously mentioned target values for the share of renewables for 2020 of each member state (see Appendix 2).

Since the EU did not outline the explicit formula in its publications, this can be derived from a combination of the gross inland energy consumption definition and the definition provided in Directive 2009/28/EC Article 2 (f). When using all information simultaneously, it can be seen that both indicators only differ in one aspect.

The gross inland energy consumption can be formulized as:

$$\text{Gross inland energy consumption} = \text{primary production} + \text{recovered products} + \text{net imports} + \text{variations of stocks} - \text{bunker} \quad (1)$$

Comparing this to the gross final energy consumption, only the transformation losses must be excluded from the formula to determine such indicator (European Commission, 2019).

2.2 Challenges of the Energy Transition

The alteration of a country's energy mix represents a rather complex, large-scaled project and this poses unavoidable problems. On a European scale, these issues are amplified by the large variety of regions and member states within the European Union. Different obstacles raise when attempting to manage and overcome a variety of challenges. To give a glimpse of the extent of the many difficulties that the transition of an energy system and an energy mix might encounter, this chapter deals with the crucial challenges the EU-28 face. These challenges can be classified based on their different characteristics. In the relevant literature, three broader perspectives on the challenges could be identified: technical, political and public challenges.

With respect to the scope of this paper, especially the transition of an energy mix towards a cleaner and sustainable mix represents one of the main concerns. Hence this chapter deals with the challenges of the renewable energy integration. It is obvious that conventional energy carriers, like oil, gas or coal, must be replaced with zero-emission energy sources to drive the energy transition forward. While all fossil fuel-fired power plants emit greenhouse gases, only two current energy types do not directly emit GHG. These emission free alternatives are nuclear and renewable energy. The change towards renewable energies implies many new uncertainties that need to be confronted. Major challenges arise when the target of the energy transition is to establish an energy mix that consists exclusively of renewables. While nuclear power plants are able to ensure an energy supply efficiently, one of the major drawbacks might be the radioactive waste that needs to be sufficiently managed (Basu & Miroshnik, 2019). The emission free energy production of renewable energy sources is clearly their main advantage. However, it would be naïve to neglect the challenges rising from this energy type and the advantages conventional energy carriers continue to have.

In recent publications various authors have dealt with the challenges of renewable energy production in detail. This literature is summarized within this chapter to provide a comprehensive overview of the difficulties the EU-28 are facing in the energy transition.

2.2.1 Feasibility due to varying Conditions

Several renewable energy carriers are reliant on specific conditions, such as climate, geography, infrastructure or resource endowments (Stram, 2016). Many countries are therefore restricted on just a few alternatives that can be integrated into their energy mix efficiently. While the renewable power generation is strongly depended on the environmental conditions, conventional power plants are rather independent from these constraints (Stram, 2016). The implementation of solar energy e.g. would not be as efficient in the Scandinavian countries as in the Mediterranean countries due to the dissimilarities in the number of sunshine hours. Thus, this variation of the conditions in Europe makes it particularly challenging to implement renewables on a large-scale in all the EU. Conventional power plants instead have fewer limitations due to the mobility of their energy carriers. This allows them to be practically independent from the geographical location (Stram, 2016).

2.2.2 Costs

Next to renewable energy sources, hydro power plants are the only conventional energy type that does not cause any direct costs by fuel. Still, the integration of renewable energies creates unprecedented expenses in the energy system and in facilities (Stram, 2016).

The energy grid is designed to handle alternating current energy, as it is produced and fed into the grid by facilities of conventional energy carriers. Renewable power plants instead produce direct current power that needs to be fed to the grid (Stram, 2016). To synchronize the plants with the grid, additional equipment must be installed to convert the energy from a direct to alternating current (Stram, 2016). In addition, further systems must be installed to increase and synchronize the energy's voltage with the grid (Stram, 2016). Stram (2016) states that these investments, which are usually expressed as dollars per megawatt, are in sum significantly higher than in facilities of conventional energy types.

The geographical extent of renewable energy plants, like in solar or wind parks, causes high expenses as well. Since the plants can be spread over a large area, unlike conventional energy plants, the integration of these facilities involves additional costs (Stram, 2016). After transporting the energy from the plant to the grid, it needs to pass further delivery nodes before it finally arrives to the ultimate consumer (Stram, 2016). As a result, high transmission costs arise due to the long-distance transmissions of solar and wind plants and the relatively low load factors (Stram, 2016). Even the power plants of the alternative emission free energy source, nuclear energy, imply higher transmission costs caused by necessary safety measures (Stram, 2016).

2.2.3 Integration of Smart Grids

The existing structure of power grids is known to be inefficient and unreliable (Buchholz & Styczynski, 2014). This together with the increasing shares of renewables and the distributed energy generation produces necessary changes in the grid system (Buchholz & Styczynski, 2014). In order to significantly increase the share of the renewable energy production in the EU's energy system also, the EU must accept this challenge and transform its energy grids (Mourshed et al., 2015).

To prepare the grid for future developments in the energy sector and to improve its efficency and reliability, the concept of smart grids is a particularly promising approach. Several definitions of smart grids were developed over time since the concept encompasses many different components and technologies. E.g. Dileep (2020) emphasizes the definition of the European technology platform, which defines a smart grid as "an electricity network that can intelligently integrate the actions of all users connected to it – generators, consumers and those that do both – in order to efficiently deliver sustainable, economic and secure electricity supplies" (p. 2590).

A smart grid hereby is the enabler to integrate different users, technologies and energy sources efficiently into the grid (Buchholz & Styczynski, 2014).

2.2.4 Bi-directional Flow of Energy

The bi-directional flow of energy is a technical condition that is closely interrelated with the smart grid concept. With the transition in the energy production also the energy flow in the grid needs to be adjusted. Through the rise of distributed energy generation whereby energy consumers can become generators as well, the grid must be able to manage both, the energy flow from distributed systems and large facilities (Buchholz & Styczynski, 2014; Nikoletatos & Tselepis, 2015). The solution that enables consumers to become suppliers as well is a bi-directional flow of energy in the grid.

This bottom-up energy flow has different applications scenarios. A basic application is the installation of photovoltaic systems in private households which feed excess energy into the grid (Nikoletatos & Tselepis, 2015). Another use case is connected to the advancing plug-in electric vehicles. In a uni-directional grid the vehicles are fed by the grid during their charging periods. The bi-directional grid moreover allows them to fed energy back to the grid if necessary e.g. for frequency and voltage regulation (Mwasilu, Justo, Kim, Do, & Jung, 2014).

2.2.5 Flexibility / Management of Energy Demand Peaks

New challenges arise with the integration of the renewable energy supply by volatile energy demands in a market. During energy peaks the demand might be significantly higher or lower than the regular level. To balance the energy demand and supply grid operators need to respond to the peaks to increase or decrease the energy supply accordingly (Stram, 2016). This makes it necessary that the energy production facilities should be sufficiently flexible.

To ensure flexibility, the energy suppliers must be able to increase and decrease their production at short notice (Stram, 2016). Considering the emission free energy types, their degree of flexibility is challenging to manage by the system operator. Among the renewables different degrees of flexibility are present. A negative example is the solar and wind energy supply (Stram, 2016). Both are mainly dependent on the meteorological conditions, so that changes in winds or solar emissions imply a certain level of unpredictability and intermittency (Stram, 2016). Accordingly, a flexible increase or reduction of produced energy is not achievable solely with this kind of technology.

To raise the flexibility of an energy system the diversification of energy sources is a relevant part of an adequate policy that should be pursued by the EU. This diversification can be implemented on a technological as well as on a geographical dimension. Stram (2016) lists geothermal, biomass and hydroelectricity as proper renewable energy types that are able to counteract peak phases.

Stram (2016) states that, regarding the generation dispatching schedule, the production of renewables is free and that these are a “must dispatch” generation source ahead of nuclear energy. If there is a low demand period e.g. on weekends, and in addition to the energy of the nuclear plants also an increase in the wind can be recorded, the energy production might exceed the demand. (Stram, 2016) In this period, renewable power plants are not able to be shut down and nuclear power plants struggle massively in their flexibility (Stram, 2016). Since the power plants need a significantly long time period to ramp up or down their load and connected systems, nuclear energy acts very inflexible on demand changes (Stram, 2016). This leads to a cost inefficient dispatching process by energy prices that lay often below the actual generation costs (Stram, 2016).

2.2.6 Interconnection of Energy Systems

An extensive interconnection on a regional, national, and international scale is a proper solution to transmit exceeding energy to neighboring systems and it would also improve the flexibility of the overall system (Nikoletatos & Tselepis, 2015). The implementation of a widely interconnected grid system also allows for a higher degree of diversification in the energy mix, by connecting different energy sources that are e.g. bound to regional conditions (Nikoletatos & Tselepis, 2015). Still, if the neighboring systems face similar problems with an exceeding energy generation simultaneously this solution is not applicable (Stram, 2016). If it is moreover considered that the European Union set its 2050 goal to a one hundred percent emission free energy production, this solution might be very challenging to combine with the EU’s vision as a stand-alone approach.

2.2.7 Storage Technology

Another solution to handle excess energy and thereby to efficiently balance supply and demand, is to store unrequired energy. The challenge current storage technologies represent is the high costs (Braff, Mueller, & Trancik, 2016; Erbach, 2016). In addition to this, energy is lost during the storage process (Erbach, 2016).

In general, there are currently two options that need to be considered if it comes to energy storage. The first option is the use of a pumped storage. To save energy, water is stored in a reservoir uphill during off-peak phases. It can then be released to drive the turbines of an electric generator downhill to serve raising energy demands (Erbach, 2016; Stram, 2016). Together with the need of having a lower and higher water reservoir this type of energy storage has the major drawback of being bound to certain conditions (Erbach, 2016). An implementation is therefore not feasible in all locations.

The second option is the storage of exceeding energy with batteries. While this type of storage is not dependent on any location and could be applied anywhere in the grid, the high costs are currently a major drawback (Erbach, 2016). However, it is an emerging option, which is driven by a scale up of production and development of the battery technologies (Erbach, 2016).

2.2.8 Energy Security

As illustrated in chapter 2.1.2 the EU aims at an increased share of renewables to reduce their GHG emissions and seeks for high energy security simultaneously. While the Energy Union defined energy targets for all member states, each member state also follows its national interests. The attitude about the influence of the energy transition on the energy security is hereby determined by varying preconditions among the EU-28 (Mata Pérez, Scholten, & Smith Stegen, 2019). As a result, Mata Pérez, Scholten, and Smith Stegen (2019) managed to identify two general groups among member states.

Countries of the first group are skeptical of their future energy security. This is a consequence of their high dependency on energy imports from non-member states, especially from Russia, and of their vulnerability to any disruption in the supply (Mata Pérez, Scholten, & Smith Stegen, 2019). Simultaneously these countries own a poor power infrastructure and a non-diverse energy supply system (Mata Pérez, Scholten, & Smith Stegen, 2019).

The second group instead promotes the Energy Union and the energy transition. The countries in this group, which are located in western continental Europe, own a diverse energy mix and reform their import dependence, to fight climate change and gain business advantages (Mata Pérez, Scholten, & Smith Stegen, 2019). However, to mitigate the risks of the import dependency they diversify the energy types, their suppliers and the supply routes (Mata Pérez, Scholten, & Smith Stegen, 2019). In general, the group members could be identified to be politically more stable and in good international relationships with their energy suppliers (Mata Pérez, Scholten, & Smith Stegen, 2019).

Guivarch and Monjon (2017) furthermore state that the goals of a GHG emission reduction and a higher energy security do not necessarily complement each other. Rather they might imply contradictions. One of the effects of the expansion of renewables could generate the need for energy sources with a high flexibility in energy peak phases like natural gas. In the recent past the change to renewables already showed, that it might lead to an even higher dependency on gas imports from Russia (Guivarch & Monjon, 2017). In this context the Nord Stream 2 project between Russia and European countries, that will be further described in chapter 2.3.3, is an exemplary case in how the goals of the energy independency and security can contradict each other. This dependency on non-EU countries might influence the energy security negatively, for instance in a scenario where the relationship between Russia and the EU becomes particularly tense. A recent example of the risks this dependency implies can be seen by the Crimea crisis that strained the relationship between the parties (Guivarch & Monjon, 2017).

However, Guivarch and Monjon (2017) believe that several studies concluded that the energy policies affect the energy security in very different ways. They conclude that the time component is hereby the critical factor for making assumptions on the effect of the policies on the energy security. As a result, the energy policies can affect the energy security differently in the short-, medium- and long- term (Guivarch & Monjon, 2017).

2.2.9 Pricing Mechanisms

Another issue is the impact of an energy transition on the price development. So does Stram (2016) point out that there is a strong positive correlation between the share of renewables (in this case solar and wind energy) and the general price level of energy. The author finds that with a rise in the renewable energy the national price level is increasing as well. Two significant challenges regarding the change of pricing mechanisms need to be carried out in the energy transition.

An adequate prospective pricing system that can be found in literature and which is also proposed by the EU in the Directive 2012/27/EU, is demand response (DR). Eid, Koliou, Valles, Reneses, and Hakvoort (2016) define DR as "the ability of the demand side to be flexible, responsive and adaptive to economic signals" (p. 15). While DR is not a new concept, the rising share of renewable energies is a driver which makes the implementation of the concept particulary necessary (Eid, Koliou, Valles, Reneses, & Hakvoort, 2016). As outlined above, the energy transition implies major system challenges regarding the flexibility and stability of the grid or the high costs. DR, with time-based pricing and especially dynamic pricing mechanisms, is an considerable option to mitigate these challenges as it is enabling the energy system to be potentially more sustainable, reliable and cost-efficient (Eid, Koliou, Valles, Reneses, & Hakvoort, 2016).

However, an implementation of a time-based pricing model is not straight forward and different issues may arise. Eid, Koliou, Valles, Reneses, and Hakvoort (2016) characterize four significant challenges: the initial technology investments, coordination problems, incumbent issues and non-sustainable side-effects of DR. Especially smart meters and other necessary devices imply significant acquisition costs, e.g. the costs of a smart meter in Europe accounted from 200 to 250 Euro on average (Eid, Koliou, Valles, Reneses, & Hakvoort, 2016). Accordingly, the European Commission has found in 2014 that until 2020 the costs for electricity and gas smart meters will account to an investment of 200 and 45 billion Euro, respectively (European Commission, 2014). Coordination problems refer to the issues that might rise when different demand adjustments are required. To manage this problematic, rulesets on how to handle cases like this need to be defined which would allow to avoid coordination problems (Eid, Koliou, Valles, Reneses, & Hakvoort, 2016). The incumbent issues are based on the need to adjust the market trading policies and a compensation mechanism that prevents a false penalization of electricity suppliers for system imbalances (Eid, Koliou, Valles, Reneses, & Hakvoort, 2016). A shift of peaks in time represents another challenge that results in increasing CO_2 emissions (Eid, Koliou, Valles, Reneses, & Hakvoort, 2016). The emissions might rise due to an increased base-load production of “dirty” energy production combined with a simultaneous reduction of a cleaner peak-load energy production (Eid, Koliou, Valles, Reneses, & Hakvoort, 2016).

The second major challenge in the price structures that might occur is negative pricing. To balance supply and demand in a case over low-demand peaks, incentives must be provided by the authorities to encourage users to consume more energy (Stram, 2016). As a consequence, negative energy prices are offered to counteract the overproduction of energy (Stram, 2016). By doing so consumers are then biased to use more energy during low demand periods. This practice is not hypothetical since, as Stram (2016) states, similar incidents already occurred during recent years.

2.3 Historical Energy Incidents

In addition to the challenges, recent history showed that different countries and regions had to face major incidents that had an impact on the energy sector. These incidents were either directly caused in production facilities or in related fields, like in raw material sourcing or in energy transmission. In general, the incidents directly led to consequences on the environment. For this reason, they raised a high awareness regarding the threats of various energy carriers and systems. Consequently, they might lead to an overthinking of the influences of the energy systems on the public side as well as on the political side.

This impact can be identified easily by examining historical energy incidents more into detail. For instance, historical incidents have led to changes in the energy mixes, in exit plans for nuclear energy and has improved the energy security and strategies to increase the share of carbon free energies. Also, the introduction or change of policies by responsible authorities can be associated with certain incidents. Regarding the public, the occurrence of certain incidents might influence their opinion and lead to a stronger interest in energy politics and support for changes in the energy system.

To get further insights into the causes and consequences this chapter summarizes some major energy related incidents and disasters, that occurred in the last decades, in chronological order. These incidents vary mainly in their geographical and technical occurrence, their causes, and their influences on the energy systems, policies, and society.

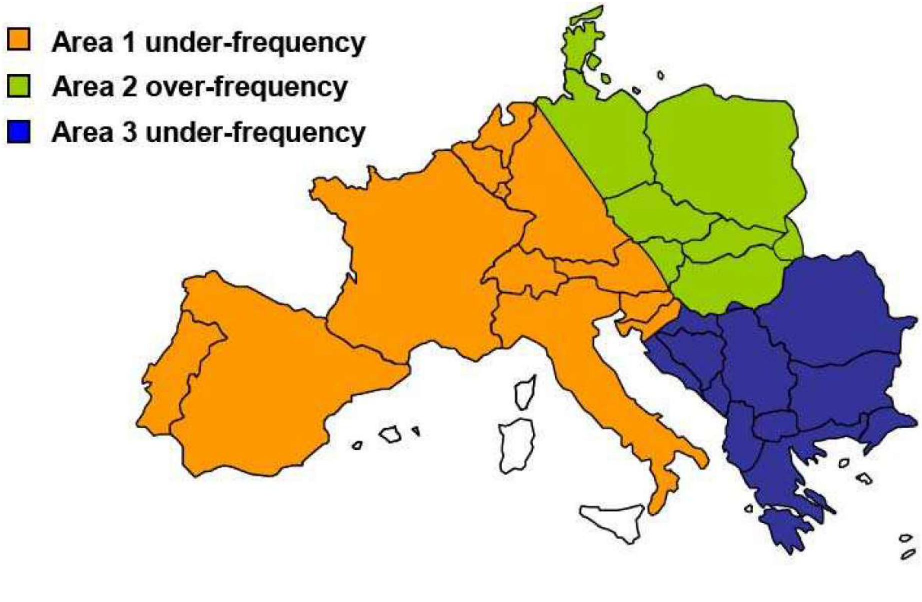
2.3.1 Power Blackouts

In the last two decades, several massive power blackouts have been recorded in different locations all around the world. To not go beyond the scope of an overview, in the following lines two exemplary blackouts are summarized.

2.3.1.1 Europe 2006

The power blackout on November 4, 2006 is a recent example of a power incident, that affected countries of the EU massively. The incident was triggered by the German electricity company EON (C. Li, Sun, & Chen, 2007). Since a cruise ship had to pass under a high voltage line of the transmission grid, the company switched off these lines in the north of Germany (C. Li, Sun, & Chen, 2007). While the switch off was communicated in advance from the shipyard with the involved TSOs and the analysis of the grid security was performed properly, the switch off was confirmed (C. Li, Sun, & Chen, 2007). As the shipyard requested an extension of the deactivation period by three hours just on the previous day the changes could not be communicated with the authorities in the Netherlands (C. Li, Sun, & Chen, 2007). This led to an insufficient power exchange between both countries and an overloading of the lines, which resulted in a division of the European grid into three parts (see Figure 4). The western part (Area 1), was low on power due to missing imports from the east, the eastern part (Area 2) instead had excessive energy, and south-eastern part (Area 3) faced just minor imbalances (C. Li, Sun, & Chen, 2007). Users in western Europe were majorly affected negatively since the low power situation caused a turn-off of consumers to balance the supply and demand (C. Li, Sun, & Chen, 2007).

Figure 4: Affected Regions of the Power Blackout 2006



Source: C. Li, Sun, and Chen (2007)

As the TSOs became aware of the situation quickly the lines were turned on again and a resynchronization process was initialized (C. Li, Sun, & Chen, 2007). However, millions of consumers in France and Germany and hundreds of thousands in Belgium, Italy, the Netherlands and Spain remained without energy for around two hours (C. Li, Sun, & Chen, 2007).

In comparison, the European blackout just occurred on a small timescale and without any permanent damages. However, it demonstrated that the European energy system is a complex construct, in which one wrong decision or mistake might have wide-reaching consequences for all member states. As a result, in January 2007 the European Commission released a statement which highlights that adequate measures on a European level are urgent (European Commission, 2007). Specifically, common security standards, an improved coordination between TSOs and higher investments in the grid must be promoted (European Commission, 2007). This power blackout demonstrates how the EU could benefit from a concept of an Energy Union to prevent future incidents, e.g. through higher energy security reached by means of an improved information flow and targeted European energy policies.

2.3.1.2 India 2012

A large-scale example of lacking energy security can be found in India in 2012. Between July 30 and July 31 major parts of northern and eastern India suffered the largest power outage in history. Two power failures affected more than 620 and 700 million people, respectively (Romero, 2012; Wu, Chang, & Hu, 2017). Investigations showed that the issues were caused initially due to grid problems. The first outage was a consequence of an overloading of a transmission double line where one line was under maintenance (Romero, 2012; Wu, Chang, & Hu, 2017). The result was a 32 GW generation shortage (Wu, Chang, & Hu, 2017). Due to the inappropriate crisis management and the following countermeasures, another system failure was caused the following day (Wu, Chang, & Hu, 2017).

As adequate improvements and prevention measures, enhancing real-time monitoring and microgrids was suggested (Romero, 2012; Wu, Chang, & Hu, 2017). Specifcally, the distributed energy generation is deemed to be an appropriate measure. The application would improve the management of future issues and the security of the functionality of essential services in particular rural areas (Romero, 2012).

2.3.2 Nuclear Disasters

When nuclear energy is examined two major incidents in history are well known: the incident at the Chernobyl power plant in 1986 and the incident at the Fukushima Daiichi power plant in 2011.

2.3.2.1 Chernobyl 1986

The first big nuclear disaster happened on April 26, 1986 in the Chernobyl power plant when the reactor block 4 exploded (Haas, Mez, & Ajanovic, 2019). When the reactor block was supposed to be shut down for maintenance, a test was carried out before the shutdown (Haas, Mez, & Ajanovic, 2019). Due to human errors in combination with technical design errors of the power plant the explosion of the reactor was caused (Haas, Mez, & Ajanovic, 2019; Yablokov, Nesterenko, & Nesterenko, 2009). The resulting graphite fire that lasted many days spread radioactive contamination parts over Europe, Asia, North America and northern Africa with major consequences for the environment, people and countries (Haas, Mez, & Ajanovic, 2019; Yablokov, Nesterenko, & Nesterenko, 2009).

2.3.2.2 Fukushima Daiichi 2011

A more recent nuclear disaster event materialized when Japan's coast suffered major consequences provoked by a Tsunami that resulted from a 9.0 magnitude earthquake in the Pacific sea (Funabashi & Kitazawa, 2012). The natural disaster on March 11, 2011 had a destructive impact on the Japanese coast where the Fukushima Daiichi power plant was affected significantly (Funabashi & Kitazawa, 2012).

Since the power plant was cut off from all electricity supplies, the reactors which were shut down automatically could not be cooled down anymore (Funabashi & Kitazawa, 2012). This led to a meltdown in reactor cores at unit 1, 2 and 3, hydrogen explosions at unit 3 and the exposure of the environment to radioactive materials (Funabashi & Kitazawa, 2012). This major disaster also highly impacted on the awareness of the nuclear energy production risks and failures in the EU (Funabashi & Kitazawa, 2012).

Previously to the disaster, nuclear energy represented a fundamental part of Japan's long-term energy strategy (Vivoda & Graetz, 2015). With the incident this strategy changed and many parties in Japan were expressing opposing opinions. This included the government, the public and the economic representatives. The public opinion was influenced by the incident as well, which resulted in a 70% approval rate of a nuclear phase out among the Japanese population and anti-nuclear demonstrations were omnipresent at that time (Vivoda & Graetz, 2015). Business representatives on the contrary argued that the national economy was dependent on nuclear energy for a full recovery from the disaster (Vivoda & Graetz, 2015). Consequently, the disaster led to a debate about the nuclear power policies in the Japanese government (Vivoda & Graetz, 2015).

The nuclear disaster of Fukushima also led to an overthinking of the energy mix in Europe. So did e.g. Germany that has shown an instant reaction to the disaster. On March 15, 2011 seven nuclear reactors were shut down temporarily and in June a law was passed to regulate a nuclear power phaseout by 2022 (Hake, Fischer, Venghaus, & Weckenbrock, 2015).

2.3.3 Fossil Fuel Incidents

Regarding solid fuels two recent events are further examined in this section. More in detail these are the "Deepwater Horizon oil spill" which was an incident regarding an oil rig in the Gulf of Mexico and the planning and construction of the gas pipeline "Nord Stream 2".

2.3.3.1 Deepwater Horizon Oil Spill 2010

The Deepwater Horizon incident occurred in April 2010. The oil spill was the result of an explosion that led to the sinking of an oil rig operated by BP near the coast of Louisiana, USA (Gyo Lee, Garza-Gomez, & Lee, 2018). Reports showed that the disaster was caused by violations of several parties involved in the operation of the oil rig (Gyo Lee, Garza-Gomez, & Lee, 2018).

As a result, an estimate of 3,19 million barrels oil were released in the environment and caused massive damages to the Gulf of Mexico area (Gyo Lee, Garza-Gomez, & Lee, 2018). Along with provoking environmental deterioration, the disaster caused large economic losses within the oil industry. The disaster massively damaged BP's reputation and the cleanup cost the company billions next to negative influences in the oil markets (Gyo Lee, Garza-Gomez, & Lee, 2018).

2.3.3.2 Nord Stream 2

Furthermore, measures of governments show that especially economic and political interests have a big influence on the energy politics. In this context the "Nord Stream 2" project must be mentioned. The project has the purpose of building a new gas pipeline from Russia to Germany in the North Sea and is putting a barrier to the aspired change to renewable energy carriers in Germany and the EU (Tichý, 2019). However, Nord Stream 2 might be one of the most discussed energy projects in the EU at the moment due to different interests and opinions regarding the project among several member states and EU institutions (Tichý, 2019). Political opponents argue that amongst others the project increases the dependency on Russia and leads to the isolation of Ukraine by bypassing the country (Tichý, 2019).

2.3.4 "School Strike for Climate" Movement

The "school strike for climate" is a movement originated in Sweden that rapidly expanded worldwide. The recently initiated and still ongoing movement for a change in climate policies was initially caused by natural disasters in Sweden.

In 2018, heat waves and forest fires in Sweden were the cause that the movement for comprehensive measures against the climate change started (The Economist, 2019). In the aftermath of the natural disasters, the 15-year-old Swedish girl Greta Thunberg initiated the "School Strike for Climate" movement by sitting in front of the Swedish parliament instead of attending school in August 2018 (The Economist, 2019). Hereby she showed a sign that said "Skolstrejk för klimatet", which means "school strike for climate", to raise awareness for climate change and to demand actions against it (The Economist, 2019). Her initial goal was to continue until the polices of the Swedish government are in-line with the climate goals of the Paris agreement (The Economist, 2019). Within the following months, children in many cities all over the world became aware of Greta Thunberg's actions and organized school strikes each Friday, and a movement that attracted large international attention has commenced (The Economist, 2019).

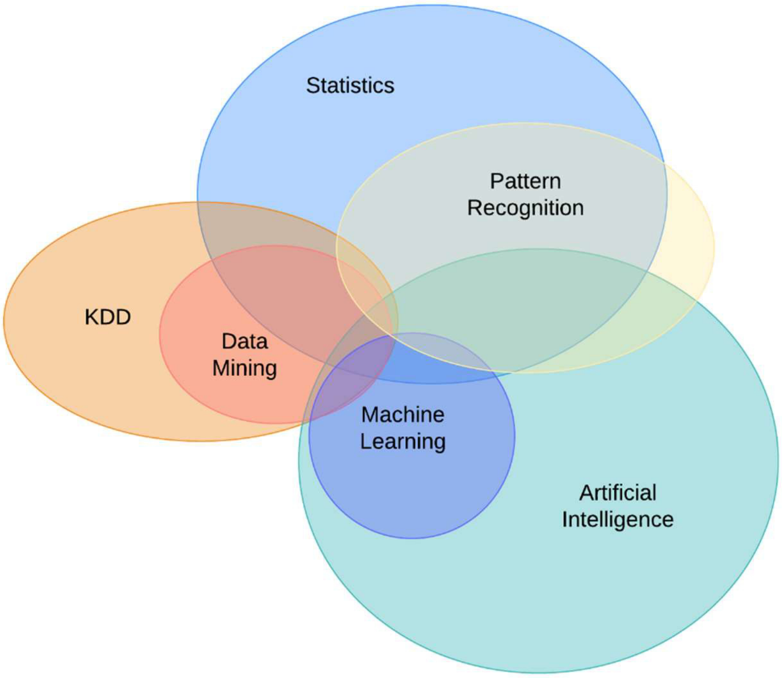
2.4 Data Mining and Machine Learning Methods

To understand the subject under analysis in the present thesis, which is based on data mining (DM) and machine learning (ML), a comprehensive introduction into both subjects is given in this chapter. For this purpose, it is important to outline the relationship of the different fields at first. After giving the reader an understanding of the basic components and principles, subsequently, a special emphasis is laid on model-based clustering algorithms and artificial neural networks. They are of particular importance and form the foundation of the following analysis of the European energy production.

2.4.1 Relationship between Data Mining and Machine Learning

To gain an understanding of the conceptual delineation between DM and ML the development of DM and its link with other disciplines must be examined first. Provided that traditional data analysis techniques were not able to handle diverse kinds of datasets the need for a new discipline was initiated (Tan, Steinbach, & Kumar, 2006). Data mining was supposed to deal with challenges like scalability, high dimensionality, heterogenous and complex data, data ownership and distribution or a non-traditional analysis (Tan, Steinbach, & Kumar, 2006). To overcome these challenges researchers with different backgrounds started to develop what is known nowadays as data mining (Tan, Steinbach, & Kumar, 2006). Hereby, DM represents a combination of methods and algorithms from various researchers' original disciplines. This is the reason why data mining intersects with known tools and methods from the fields of machine learning and pattern recognition, statistics and AI as well as database systems (Tan, Steinbach, & Kumar, 2006). As a matter of fact, this means that a clear distinction between the methods and algorithms of DM and ML is not always possible. Some of them can be both used in DM and in ML. In chapter 2.4.3 and 2.4.2 this overlapping will become more clear.

Figure 5: Degree of Overlapping of various Data Science Fields



Source: Adapted from Shobha and Rangaswamy.

2.4.2 Machine Learning

Rebala, Ravi, and Churiwala (2019) define ML as "a field of computer science that studies algorithms and techniques for automating solutions to complex problems that are hard to program using conventional programing methods" (p. 1). As Figure 5 illustrates ML is rather considered as a subfield of artificial intelligence (AI). While AI is generally focusing on the use of several approaches to make machines intelligent, ML focuses only on one approach. Namely, this approach is the creation of "models by learning from existing datasets to predict or forecast outcomes or behavior" (Rebala, Ravi, & Churiwala, 2019, p. 4). Rebala, Ravi, and Churiwala (2019) list classification, clustering and prediction as problems that can be solved by the application of ML models. The models can be classified into different categories. Therefore, it must be distinguished between supervised, unsupervised, semi-supervised and reinforcement learning models (Rebala, Ravi, & Churiwala, 2019).

To further address the characteristics of the models, the difference between labelled and unlabeled data needs to be clear upfront. Labelled data describes data that is able to answer a question. However, if the data is not able to directly answer a question unlabeled data is present (Rebala, Ravi, & Churiwala, 2019).

Based on labelled data, supervised learning algorithms can be implemented. The algorithms need to be trained by hand before being able to solve a problem. By applying a part of a given dataset as training data, the algorithm learns the key characteristics of each data point with an corresponding answer (Rebala, Ravi, & Churiwala, 2019). If properly implemented, this enables the algorithm to provide a right outcome or answer even for an unseen dataset (Rebala, Ravi, & Churiwala, 2019). Supervised learning problems can moreover be differentiated between classification and regression problems. Classification problems refer "to the problem of identifying the category to which an input belongs to among a possible set of categories" (Rebala, Ravi, & Churiwala, 2019, p. 57). On the contrary regression problems refer to models that are able to make predictions with information steaming from continuous variables (Rebala, Ravi, & Churiwala, 2019).

Whereas supervised learning uses labelled data, unlabeled datasets are sufficient to apply unsupervised learning algorithms. The algorithms are applied to identify previously unknown similarities or patterns in the dataset and assign them to groups or clusters (Rebala, Ravi, & Churiwala, 2019).

Semi-supervised learning represents a hybrid model with principles of both, supervised and unsupervised learning. In a semi-supervised learning problem a model can be applied on a dataset with only some datapoints labelled (Rebala, Ravi, & Churiwala, 2019). The algorithms first utilize unsupervised learning methods to identify groups or clusters in the dataset (Rebala, Ravi, & Churiwala, 2019). Then labels are assigned based on known labeled data points within each group (Rebala, Ravi, & Churiwala, 2019).

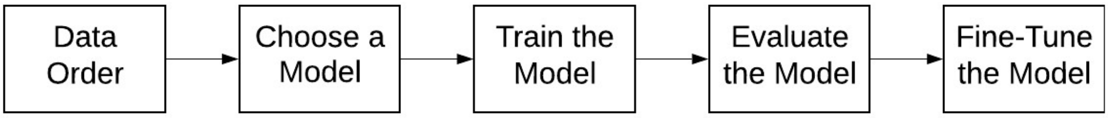
The class of reinforcement learning focuses on changing situations and huge state space. With reinforcement learning a machine is enabled to sense its "external environment and choose an action based on its own state and the external environment, with the aim of maximizing a specific predefined goal" (Rebala, Ravi, & Churiwala, 2019, p. 22).

Many algorithms belong to the field of ML and can be assigned to any of the previously introduced learning models like:

- Naïve Bayes' Algorithm
- Support Vector Machines
- K-Means Algorithm
- K-Nearest Neighbor (KNN)
- Random Forest
- Artificial Neural Networks (ANN)
- Recommender Systems
- Reinforcement Learning System

A general process of implementing machine learning and deep learning efficiently is specified by Taulli (2019) and involves five consecutive steps which are illustrated in Figure 6.

Figure 6: Machine Learning and Deep Learning Process



Source: Own work.

As a first measure it must be ensured that the data is unordered. Otherwise the algorithm might be able to detect the order as a pattern and output unwanted results (Taulli, 2019). The next step is to choose an adequate learning model, which is likely to be a trial and error process (Taulli, 2019). Subsequently, the model needs to be trained on the data to enable the prediction on unseen data. For this purpose, the dataset is split into a training and a testing set, where the testing data should represent the training data sufficiently (Taulli, 2019). Afterwards the testing data is employed to evaluate the accuracy of the model (Taulli, 2019). In a final step the model must be fine-tuned. In this context, fine tuning stands for an adjustment of the model parameters of the algorithm accordingly, so that the best possible results can be determined by the model (Taulli, 2019).

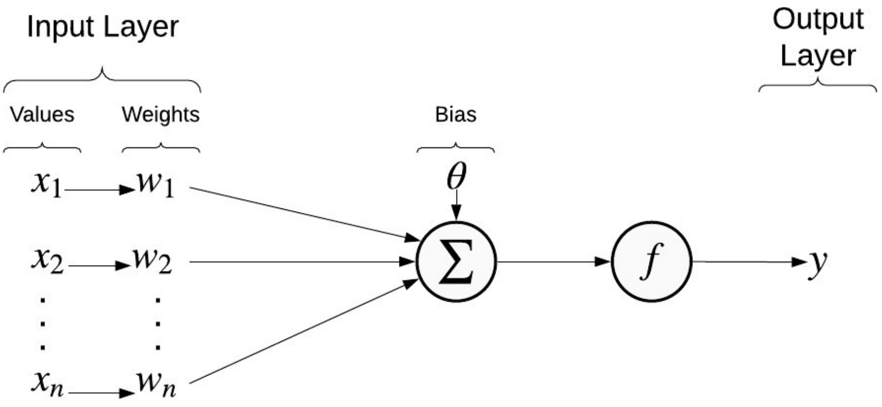
As mentioned before, in this chapter a special emphasis is laid on artificial neural networks and clustering algorithms. Both fundamental pillars of the following analysis are examined in detail in the subsequent subsections.

2.4.2.1 Artificial Neural Networks

As illustrated before, there is a close relationship between machine learning and data mining. Beyond that, also deep learning must be considered. This can be characterized as a subfield of machine learning to which also ANNs belong (Taulli, 2019). When ANNs were developed they were initially inspired by the human brain. Once scientists got a deeper understanding of the human brain, it became clear that neural networks are not reflecting the human brain functionality accurately (Taulli, 2019). However, some of the brain's fundamental functionalities are the foundation of neural networks. As a result, several different algorithms that are based on the idea of neural networks were developed over the years (Rebala, Ravi, & Churiwala, 2019; Taulli, 2019).

The simplest form of neural network is a single neuron perceptron. Figure 7 illustrates the structure of such a network.

Figure 7: Single Neuron Perceptron



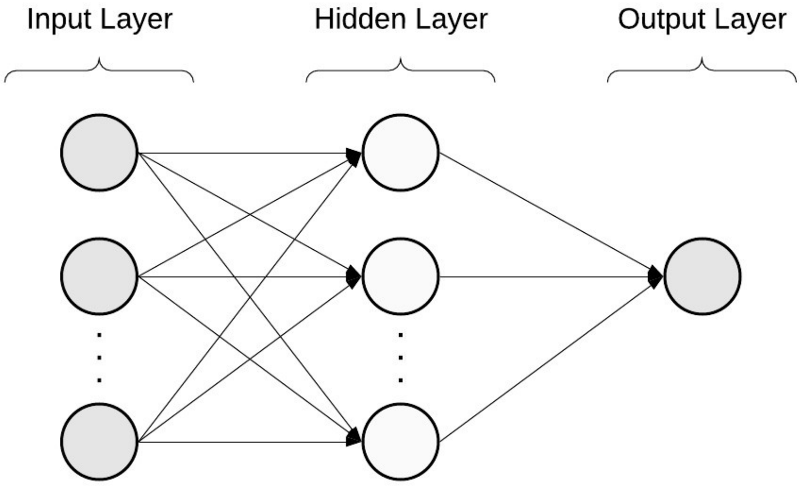
Source: Adapted from Kabir and Hasin.

The network takes a vector of values $\mathrm{x} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n\}$ and their appropriate weights $w = \{w_1, w_2, \ldots, w_n\}$ as input. The value of the weight of each input specifies the input's relative importance in the neuron (Taulli, 2019). The network then sums up the weighted input values and applies to it a predefined activation function (Taulli, 2019). In many cases a non-linear activation function is used to better reflect a real-world scenario (Taulli, 2019). An appropriate activation function can be chosen from a range of alternatives with different characteristics. In addition, the bias θ is considered in the calculation. This constant value is included to achieve smoother calculations (Taulli, 2019). The output is formalized as (Dreyfus, 2005):

$$y = (\textstyle\sum_{i=0}^{n} w_i x_i + \theta) \tag{2}$$

A major drawback of a single neuron perceptron is that it is only able to solve linearly separable problems (Livingstone, 2009). Depending on the nature of the problem, its complexity and the data's distribution, more neurons can be added in a hidden layer to enhance the power of the model (Kabir & Hasin, 2013). The hidden layer is located between the input and the output layer. Figure 8 visualizes a general model composition of a neural network with one hidden layer. Moreover, this network can be classified as a feed forward neural network. Feed forward neural networks are, as the name indicates, one-directional. The input of neurons in one layer can only be the output of a previous layer (Dreyfus, 2005; Taulli, 2019). If each neuron is connected to the neurons of a subsequent layer, the network is moreover specified as fully connected (Taulli, 2019).

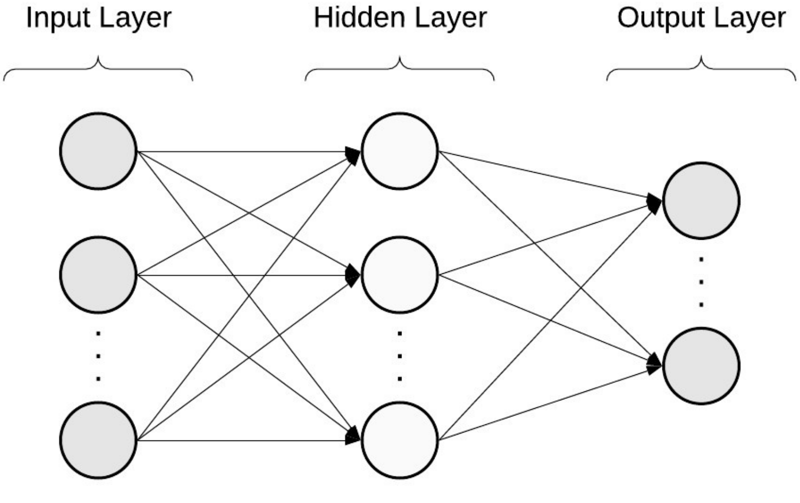
Figure 8: Neural Network with one hidden Layer



Source: Adapted from Kabir and Hasin.

For given tasks it might be necessary for the neural network to create more than one output value (Dreyfus, 2005). The structure of such a model is illustrated in Figure 9.

Figure 9: Multi-Output Feed Forward Neural Network

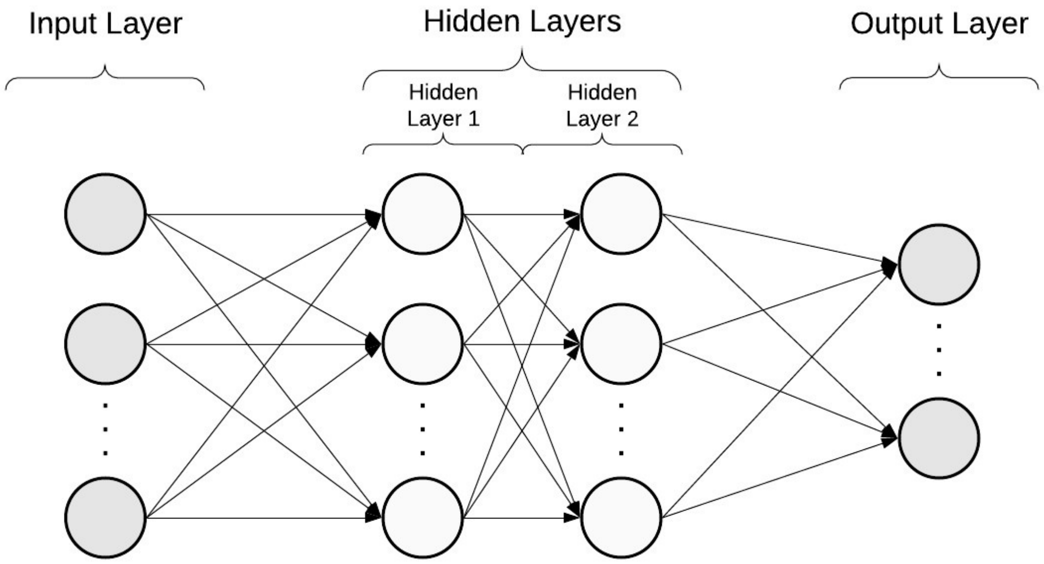


Source: Own work.

If a neural network is applied to a much more complex problem, multiple hidden layers can be stacked on each other (Bisong, 2019; Dreyfus, 2005). A model with multiple hidden layers (see Figure 10) is known as a multilayered perceptron (MLP). In addition to their high performance MLPs also promote backpropagation (Taulli, 2019). The technique, which was first introduced in the 1970s, solves one of the major drawbacks of ANNs. Still, its breakthrough came in 1986 when Rumelhart, Hinton, and Williams released the paper "Learning Representations by Back-propagating Errors" (Taulli, 2019). The application of backpropagation helps to maximize the model accuracy by training the network (Taulli, 2019). This is done by adjusting the weights of the network accordingly (Taulli, 2019). Whereas the adjustment of the weights was highly time-consuming with traditional methods, by introducing backpropagation this issue became negligible (Taulli, 2019).

Once the forward network propagation is completed, either the cost function or the error between the predicted and actual output is calculated (Bisong, 2019). Normally the output of a feed forward neural network is likely to be incorrect with a high error after the first run (Bisong, 2019; Taulli, 2019). To minimize the cost function in machine learning algorithms the gradient descent optimization algorithm is applied, specifically to minimize a model's predefined cost function (Bisong, 2019). The resulting error gradient is then successively back propagated through each layer of the network to adjust the model's weights (Bisong, 2019).

Figure 10: Multilayered Perceptron with two hidden Layers

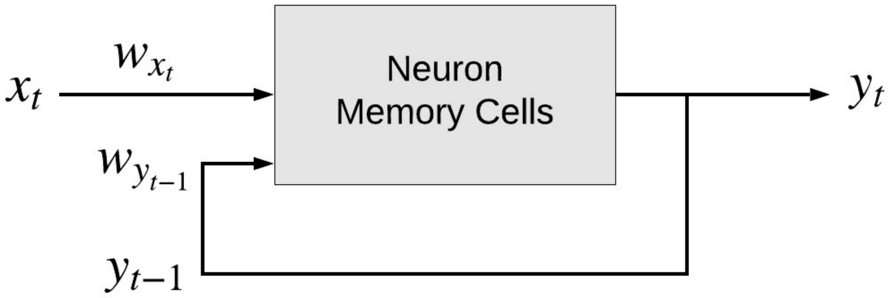


Source: Own work.

Unlike feed forward neural networks, a network can also contain cyclical connections or loops (Dreyfus, 2005). Such a network is called recurrent neural network (RNN). The networks were developed precisely to solve learning problems within time dependent data e.g. time series datasets (Bisong, 2019; Rebala, Ravi, & Churiwala, 2019). A RNN applies a looping framework, so that the output of one sequence is an additional input to the following one (Bisong, 2019). For this reason the predictions made by an RNN are not only based on the input but also on past sequences (Bisong, 2019).

When compared to a classical neural network, a neuron of an RNN has one key difference. As indicated afore, the RNN neuron has in addition to its input value x_t another input y_{t-1} that represents the output of the previous sequence (Bisong, 2019). Figure 11 illustrates a recurrent neuron and its behavior, i.e. it shows how an RNN maintains in its memory information on previous computations. In addition to the two input values, the neuron also receives fitted weights w_{x^t} and $w_{y^{t-1}}$ (Bisong, 2019).

Figure 11: Recurrent Neuron



Source: Adapted from Bisong.

By unfolding the recurrent neuron, the structure of a basic RNN can be formalized. Hereby, the output y_{t-1} of a neuron memory cell is the input of the subsequent recurrent layer at timestep t (Bisong, 2019). Next to receiving the output of the previous sequence the neuron receives its regular input value x_t at the current timestep as well (Bisong, 2019). When a dataset is employed in a RNN the number of recurrent layers depends on the sequence length of the dataset (Bisong, 2019). Subsequently, for each n layered sequence of a dataset,n layers are added to the RNN. Each recurrent layer of a network hereby consists of a number of neuron memory cells (Bisong, 2019).

RNNs are trained by backpropagation through time (BPTT), which is a modified version of the backpropagation algorithm that is adjusted to train models with recurrent structures (Bisong, 2019). This is done by unrolling the neuron primarily to applying the backpropagation algorithm to the neurons at each time step, similarly to the traditional approach of ANNs (Bisong, 2019). The major downside of RNNs is the vanishing gradient problem. This problem is most likely to occur if the model gets very large, and as a result it will be struggling to learn long-term dependencies (Bisong, 2019; Rebala, Ravi, & Churiwala, 2019; Taulli, 2019). By obtaining volatile gradient weights for the neurons, the values can become either very large or vanishing small (Bisong, 2019). If this is the case, the neurons are not able to learn anymore and the vanishing gradient problem is present (Bisong, 2019; Rebala, Ravi, & Churiwala, 2019).

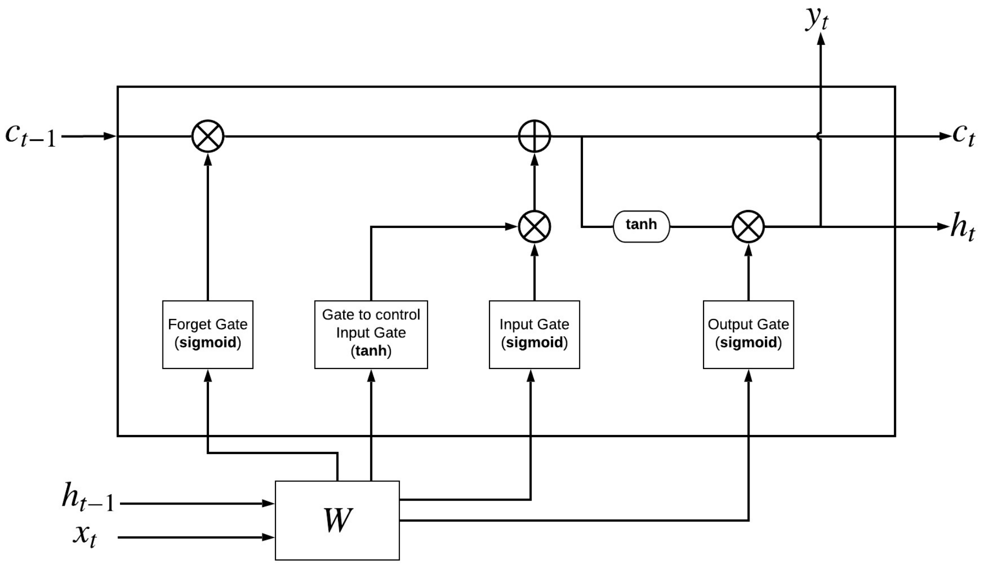
For this reason, long short-term memory (LSTM) neural networks were developed. These networks enable a very efficient handling of time series data with long-term dependencies (Bisong, 2019). Compared to classical RNNs, LSTM neural networks have additional gate components, which control the information flow of the recurrent neurons (Bisong, 2019). More precisely, these components are the memory cell, the input gate, the forget gate and the output gate.

Figure 12 illustrates the composition of a LSTM cell and its components, gates and the connections among them. It can be observed that the cell has three input values (Bisong, 2019):

- the cell state of the previous time instance c_{t-1}
- the hidden state of the previous time instance h_{t-1}
- and the current input value x_t

The input gate controls "what information gets stored in the long-term state or the memory cell, c" (Bisong, 2019, p. 455). In parallel another gate controls the information flow of the input gate (Bisong, 2019). Moreover, a forget gate determines which information of c_{t-1} remains over time (Bisong, 2019). The LSTM cell has two final outputs y_t and h_t. The information that flows into these outputs is determined by the output gate (Bisong, 2019).

Figure 12: LSTM Cell Architecture



Source: Adapted from Bisong.

2.4.2.2 Clustering Algorithms

Clustering algorithms are machine learning methods applied to unlabeled groups of homogenous data (Bisong, 2019; Tan, Steinbach, & Kumar, 2006). The resulting groups of data points are refered to as clusters. Clustering algorithms can be categorized based on three different criteria.

One distinction is typically made between partitional and hierarchical clustering algorithms. In partitional algorithms, each data point is assigned to one subset of the whole dataset (Tan, Steinbach, & Kumar, 2006). Hierarchical algorithms on the contrary are organized as a tree. The root of the tree represents a cluster that contains all data points, while the lowest level is represented by the leaves, that just contain a single object (Tan, Steinbach, & Kumar, 2006). Moving up the hierarchy of the tree, each cluster unites its children or sub-clusters (Tan, Steinbach, & Kumar, 2006).

The assignment of the data objects to the clusters can moreover be hard or fuzzy. In case of a hard assignment a data point can be assigned exclusively to one single cluster (Tan, Steinbach, & Kumar, 2006). Fuzzy clustering methods might assign them to more than only one cluster (Tan, Steinbach, & Kumar, 2006). The algorithms assign weights between 0 and 1 to each data point which represent the probability that a data point is a member of a certain cluster (Tan, Steinbach, & Kumar, 2006). The sum of all probabilities must account for a hundred percent or 1, accordingly (Tan, Steinbach, & Kumar, 2006).

The final distinction that can be made relates to the completeness of the clustering algorithm. A clustering method can be either complete or partial. In certain cases, not all data points are included into the clustering mechanism, e.g. due to outliers (Tan, Steinbach, & Kumar, 2006). In this case, the method called partial clustering (Tan, Steinbach, & Kumar, 2006). Whereas, in a complete clustering approach, all data points are required to be included into the clustering procedure to avoid any kind of loss of relevant data (Tan, Steinbach, & Kumar, 2006).

While many clustering algorithms exist, the EM algorithm is outlined more in detail in the subsequent paragraphs. The clustering method is of particular importance to develop the energy paths of the European energy production in the following analysis. Since the model applied will have a strong relationship to Gaussian mixture models (GMM) and maximum likelihood estimation (MLE), an introduction to these topics will be provided before outlining the full EM algorithm.

In model-based clustering it is assumed that a mode which describes the data in a comprehensive manner exists (Tan, Steinbach, & Kumar, 2006). Model-based clustering algorithms hereby aim at finding the model that best fits the data (Tan, Steinbach, & Kumar, 2006). To achieve this, the model applies a probability distribution that efficiently describes the dataset. Still, it is very likely that one distribution cannot describe all data points sufficiently (Tan, Steinbach, & Kumar, 2006). To capture different distributions of the dataset a mixture model is applied, in which each distribution of the model corresponds to a different cluster (Tan, Steinbach, & Kumar, 2006). Often multivariate normal distributions are used in mixture models (Tan, Steinbach, & Kumar, 2006). Besides delivering good results mixture models are easy to understand and to apply (Tan, Steinbach, & Kumar, 2006).

In a gaussian mixture model the data is described by a Gaussian normal distribution. Formula 2 defines the probability density function of a Gaussian normal distribution (Tan, Steinbach, & Kumar, 2006).

$$\mathrm{P}(x_i|\theta) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{3}$$

The distribution is defined by its parameters $\theta = (\mu, \sigma)$, consisting of its mean μ and its standard deviation σ. A classical approach to estimate the model parameters is the maximum likelihood estimation (Tan, Steinbach, & Kumar, 2006). If a set n of independent data points that follows the Gaussian normal distribution is assumed, Equation 3 can be extended to Equation 4 to describe the probability density function of the all data points (Tan, Steinbach, & Kumar, 2006).

$$\mathrm{P}(\mathcal{X}|\theta) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{4}$$

Given that the results of Equation 4 can get very small, to ease interpretability and computation, the logarithm (see Equation 5) is typically applied on the probability density function as below, where the product of densities becomes a summation due the properties of the logarithm applied on exponential distributions (Tan, Steinbach, & Kumar, 2006):

$$\log \mathrm{P}(\mathcal{X}|\theta) = -\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2} - 0.5\mathrm{n}\log 2\pi - n\log\sigma \tag{5}$$

To apply the probability density equations for MLE the equation must be modified. As μ and σ are representing variables and the dataset is treated as a constant now, the likelihood function (see Equation 6) and the according log likelihood function (see Equation 7) need to be derived (Tan, Steinbach, & Kumar, 2006). While the probability density function considers a random variable $\mathcal{X}$ conditional to the parameter set θ, the maximum likelihood problem defines the exact revers, i.e. it conditions the parameter set θ to a randomly extracted sample from random variable $\mathcal{X}$ (Tan, Steinbach, & Kumar, 2006). And as a consequence, the objective of MLE is to estimate the values of the parameters μ and σ (which represent θ) for which the data has the highest likelihood (Tan, Steinbach, & Kumar, 2006). This is why the goal is to find the parameters that maximize the log likelihood function (Tan, Steinbach, & Kumar, 2006).

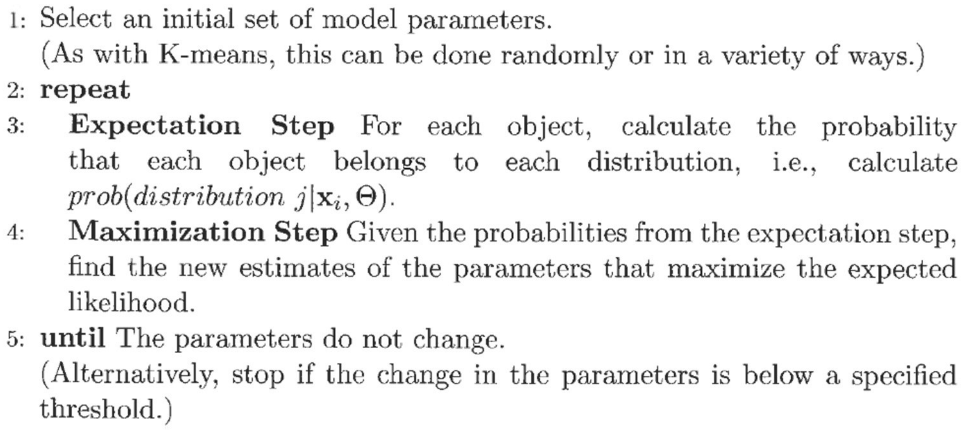
$$\mathrm{P}(\theta|\mathcal{X}) = \mathrm{L}(\theta|\mathcal{X}) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{6}$$

$$\log\mathrm{likelihood}(\theta|\mathcal{X}) = \ell(\theta|\mathcal{X}) = -\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2} - 0.5\mathrm{n}\log 2\pi - n\log\sigma \tag{7}$$

The MLE can also be applied to a calculate the parameters of a mixture model (Tan, Steinbach, & Kumar, 2006). However, if the distributions that describe the dataset are unknown, MLE is not able to directly determine the data points' probabilities and consequently the parameters (Tan, Steinbach, & Kumar, 2006). To overcome this issue, MLE can be applied as a part of the EM algorithm (Tan, Steinbach, & Kumar, 2006).

In its initial step the EM algorithm, as presented in Figure 13, assigns random values to the model parameters in θ. Afterwards the algorithm loops through the expectation and maximization step. The loop finishes either if the model parameters do not change anymore or if they fall below a previously defined threshold. In the expectation and maximization steps the actual MLE is performed iteratively (Tan, Steinbach, & Kumar, 2006). First the probability of being a member of each distributions is calculated for each object in the expectation step (Tan, Steinbach, & Kumar, 2006). To maximize the expected likelihood, in the maximization step the new parameter estimates are determined based on the probabilities of the expectation step (Tan, Steinbach, & Kumar, 2006).

Figure 13: Expectation Maximization Algorithm



```
Select an initial set of model parameters.
(As with K-means, this can be done randomly or in a variety of ways.)
repeat
    Expectation Step For each object, calculate the probability
    that each object belongs to each distribution, i.e., calculate
    $prob(distribution\ j|\mathbf{x}_i, \Theta)$.
    Maximization Step Given the probabilities from the expectation step,
    find the new estimates of the parameters that maximize the expected
    likelihood.
until The parameters do not change.
(Alternatively, stop if the change in the parameters is below a specified
threshold.)
```

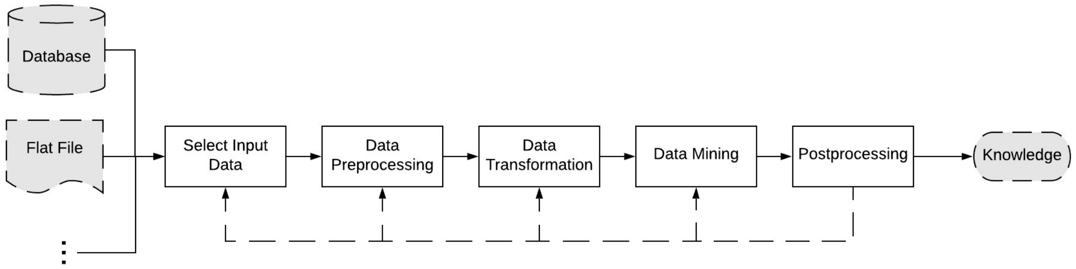
Source: Tan, Steinbach, and Kumar (2006).

2.4.3 Data Mining

Data mining represents a step within the knowledge discovery in databases (KDD) process. This procedure is a multidisciplinary activity as Fayyad, Piatetsky-Shapiro, and Smyth (1996) have outlined. Figure 5 has already clearly illustrated this overlapping between these fields. The field of data mining is hereby useful to convert raw data into useful information (Tan, Steinbach, & Kumar, 2006). KDD is specified as a combination of five basic steps. Besides the DM step, the KDD process consists of the input data selection, the data pre-processing and transformation steps and the postprocessing step (see Figure 14). KDD is commonly known to be an iterative process as loops allow for backward steps in the process (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

The first KDD process step is the selection of suitable data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). The available data can be retrieved from various data sources and can be of different data formats (Tan, Steinbach, & Kumar, 2006). Once the data selection is completed the data must be preprocessed. This step is particularly relevant since the data might be incomplete, inconsistent or noisy (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Tan, Steinbach, & Kumar, 2006).

Figure 14: Steps of the KDD Process



Source: Own work.

The process of removing these errors in the dataset is known as data cleaning (Tan, Steinbach, & Kumar, 2006). After the data is cleaned it must be transformed into the required format of the data mining step. This step usually includes feature selection and dimensionality reduction (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). The appropriate execution of data preprocessing and transformation is considered as the most important and time-consuming step, since the data is usually of many different formats and qualities (Tan, Steinbach, & Kumar, 2006).

When that the data has been prepared the actual analysis, i.e. the data mining step, can be implemented. Depending on the goal of the analysis, different data mining methods can be applied on the dataset to achieve the target. Typical methods of data mining can belong to the fields of classification, regression, clustering, summarization, association rules or anomaly detection (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Tan, Steinbach, & Kumar, 2006).

After the data mining stage is concluded, the last necessary step is the post-processing. Post-processing is the necessary step for interpretation and evaluation of the results, e.g. creating a visualization of the results (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Tan, Steinbach, & Kumar, 2006). Next to using the discovered knowledge, it can be integrated into further systems for documentation or following activities (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

In general, the toolset of DM is either of descriptive or predictive nature. Descriptive methods focus on the extraction of hidden or unexpected patterns from a dataset (Tan, Steinbach, & Kumar, 2006). In most cases descriptive methods are exploratory and a postprocessing step is needed for data validation and for an appropriate result explanation (Tan, Steinbach, & Kumar, 2006). Well known exploratory methods are e.g. clustering, association rules or anomaly detection (Tan, Steinbach, & Kumar, 2006). Predictive methods on the contrary are used to predict a specific attribute value based on other data values. For predictive methods a dependent variable needs to be determined as the attribute to be predicted (Tan, Steinbach, & Kumar, 2006). The prediction itself is then carried out based on independent variables (Tan, Steinbach, & Kumar, 2006). If the dependent variable is discrete, DM can be used to apply classification methods. If the target variable is continuous, regression methods can be applied instead (Tan, Steinbach, & Kumar, 2006).

2.4.4 Application in the Energy Sector

In relevant literature a wide range of data mining and machine learning approaches applied to research problems in the energy sector can be found. The articles differ in their implemented machine learning methods, the utilized data, the analyzed time period and in the geographical coverage of the data.

While in the past many authors used classical forecasting methods, more recently machine learning and deep learning methods have been successfully applied in the energy sector. But also the application of hybrid models is no novelty (Debnath & Mourshed, 2018). Many papers apply forecasting methods to predict short-term developments in the energy sector. Especially the examined fields of short-term forecasting vary mainly by their purpose and their geographical extent. In the energy production forecasting, it is notable that in many studies models are developed to forecast the production of renewables, in particular of solar and wind energy. This might be a popular research area due to the relative novelty of these energy sources and the need to accurately forecast their production due to their dependence on changing external conditions.

For instance, Rodat, Tantolin, Le Pivert, and Lespinats (2016) forecast the energy production of a solar power plant for the next 24 hours to establish a heat storage strategy. Another approach for solar generation short-term forecasting was developed by Bouzerdoum, Mellit, and Massi Pavan (2013). They implemented a hybrid model, by applying the seasonal auto-regressive integrated moving average (SARIMA) method and optimized the model with a support vector machine (SVM) model. Wasilewski and Baczynski (2017) instead developed a model to forecast the energy generation of two wind power plants in Poland. Based on the historical generation data for 21 months combined with historical weather data, the authors employed a MLP to forecast the generation for intra and next day power forecasting. To provide a 24 hour forecast for energy production in Trieste, two ANN models were implemented by Mellit and Pavan (2010). They successfully applied a multivariate and a univariate MLP by including air temperature and solar irradiance in the underlying dataset.

However, most research only focuses on a city or small region and very few approaches are implementing forecasts on a large scale. Mehedintu, Sterpu, and Soava (2018) compare five regression models to predict the share of renewables in the energy consumption in the EU. In detail, they have utilized data from 1995 to 2016 to carry out a forecast of the EU-28 until 2020.

Another case of a large-scale forecast that is beyond that carried out with the help of neural networks is outlined in the paper of Đozić and Gvozdenac Urošević (2019). The authors implemented a long-term forecast of GHG in the EU. The analysis specifically targets at forecasting CO_2 emissions up to 2050 by applying a hybrid ANN model. As a motivation for providing a forecast of the emissions the goals of the Energy Roadmap 2050 of the European Commission as well. The chosen input dataset of the analysis consists of the yearly totals of the EU between 1990 and 2015. Hereby, ten input variables are included into the predictive model. Specifically, the authors include the shares of the energy production mix, temperature, the gross domestic product, the average annual temperature, the energy consumption, and the population. With regards to the application of the ANN, Đozić and Gvozdenac Urošević (2019) determine the trend of the variables until 2050 with linear regression. Afterwards a cascade forward back propagation ANN with two hidden layers was applied to the dataset to carry out the forecast of the CO_2 emissions with satisfying results. To train the model, the authors used 80% of the data and the remaining 20% were used for testing. Đozić and Gvozdenac Urošević (2019) utilize the root mean squared error (RMSE) and a target range of the CO_2 emissions in 2050 for validating the results. By testing 100 models for the defined criteria, 30 models fulfilled both (Đozić & Gvozdenac Urošević, 2019).

While the application of machine learning methods in forecasting problems is by now a widespread methodology, also the application of different clustering methods in the energy sector is not an unknown methodological approach as Csereklyei, Thurner, Langer, and Küchenhoff (2017) state. However, the authors applied a model-based clustering approach to such problem which they argue is a novelty in the energy sector. They implement the approach to identify the composition of national energy mixes and test the existence of a "national-level energy ladder, energy intensity convergence and endowment lock-in effects" (Csereklyei, Thurner, Langer, & Küchenhoff, 2017, p. 442).

The "energy paths", how the authors name the development of energy mixes over time, are created by assuming that the data can be described by Gaussian mixture models (Csereklyei, Thurner, Langer, & Küchenhoff, 2017). The unit of observation is defined as "the energy mix of country "I", in year "t", or country-years" (Csereklyei, Thurner, Langer, & Küchenhoff, 2017, p. 446). For this reason, the authors generated 1025 country-year observations by merging the data of the EU-28 and 40 observation years. It should be underlined that the total number of observations is slightly reduced by the fact that the data records for Estonia, Latvia, Lithuania, Slovenia and Croatia are just observed from 1990 onwards (Csereklyei, Thurner, Langer, & Küchenhoff, 2017).

Based on this transformation the EM algorithm can be applied to fit the model (Csereklyei, Thurner, Langer, & Küchenhoff, 2017). To determine the number of clusters that are created by the model, Csereklyei, Thurner, Langer, and Küchenhoff (2017) used an adjusted (negative) Bayesian information criterion (BIC) and the Elbow criterion. As a result, the authors were able to determine clusters which "represent the same concept over the successive years" (Csereklyei, Thurner, Langer, & Küchenhoff, 2017, p. 446) for each EU member state. This concept allowed them to easily compare different years and countries with each other.

2.5 Python for Data Mining and Machine Learning

There are several software tools and programing languages used to carry out a data mining or machine learning project (Taulli, 2019). Two very common languages are R and Python (Taulli, 2019). While R is a language that focuses on statistical and data analysis, Python is a powerful programming language with many different areas of application (Taulli, 2019). The execution of Python code is platform-independent, since Python is an interpreted language (Unpingco, 2019). Among the application areas, the language has also been used to implement data mining, machine learning as well as artificial intelligence methods. The language offers a wide-reaching standard library and third-party libraries with packages for many different fields and purposes. Some of these packages that are relevant when carrying out a data mining or machine learning project will be briefly described subsequently.

Two fundamental packages to manage and process datasets are Numpy and Pandas. Numpy provides the option to process big datasets in the form of arrays and matrices. While arrays are one-dimensional, matrices are similar to arrays except for their multi-dimensionality (Unpingco, 2019). In addition to providing an architecture to integrate datasets these softwares can also process datasets with a large set of functions (Unpingco, 2019). Pandas is not only built on top of Numpy but it also expands its functionality. The package is especially applied if time series data or spreadsheet-style data is employed (Unpingco, 2019). To manage and process these datatypes Pandas provides series- and DataFrame objects. Pandas series are similar to Numpy arrays but additionally they store a corresponding index for each data value (Unpingco, 2019). Accordingly, Pandas DataFrames are the two-dimensional counterpart of Numpy matrices. Like series objects, an index is assigned to each observation in a DataFrame and each column has a heading which is called label (Unpingco, 2019). The library also offers a range of functions to access and manipulate the data. Hereby, indices and labels simplify the process of selecting and accessing specific data objects for further manipulation (Unpingco, 2019).

To visualize the data and results of later processing, different libraries are available in Python. One of the most popular and complete packages for data visualization is matplotlib (Unpingco, 2019). In addition to matplotlib also other alternatives can be helpful for more specialized visualization purposes (Unpingco, 2019).

Python also provides several libraries to apply machine learning methods. The most famous and widely used library is scikit-learn (Unpingco, 2019). It provides a wide-reaching functionality, covering typical machine learning methods and algorithms, e.g. preprocessing, classification, regression or clustering methods (Unpingco, 2019). The Keras library moreover offers the possibility of implementing common deep learning models in Python. The computations of the package are hereby executed in the backend by Tensorflow, which is a prominent deep learning framework (Unpingco, 2019).

3 ANALYSIS OF THE EUROPEAN UNION'S ENERGY PRODUCTION

3.1 Methodological Approach

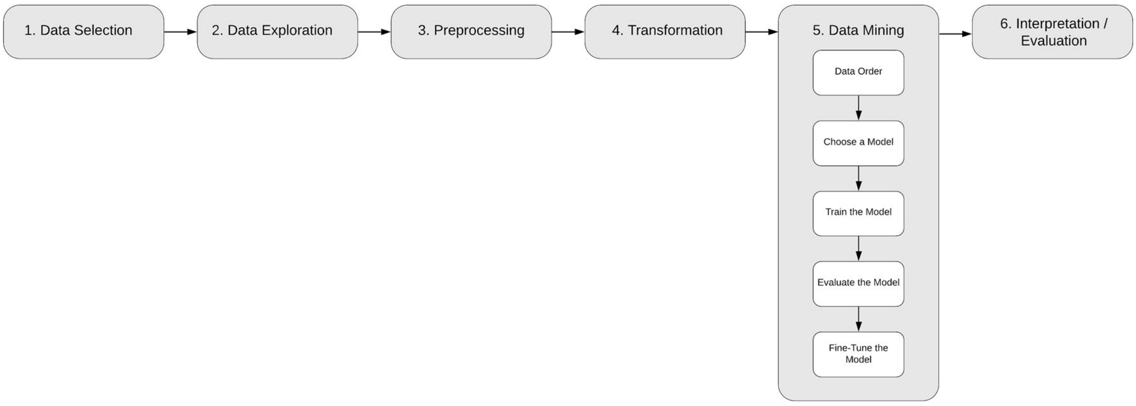
As outlined before many factors have and will still have an influence on the energy production mix and its related fields. While some events are easily predictable to a certain level, many events are hard to predict especially in long-term forecast scenarios. For this reason, the following analysis is not considering any assumptions made on future events related to the energy production in the EU. Hence, the analysis is only based on historical data, which is assumed to be rich enough to predict a likely future evolution of the energy production in the EU as precisely as possible. Moreover, the time period of the applied dataset will include events and incidents that might have already influenced the energy production in the past. In addition to these events, the comprehensive energy strategy was released by the EU recently as well. The time span of the analysis allows the forecasting model to consider both, certain past events, and possible patterns as well as trends that just started their development with recent events.

While an analysis of the EU's energy mixes might also be achieved by logical thinking and through manual analysis, in this paper data mining and machine learning methods are directly applied on the dataset instead. Since the dataset includes energy mixes for the EU-28 for a period of 28 years and 33 years will be forecasted throughout the process, a detailed manual analysis would be very time consuming. Logical reasoning also has its boundaries when looking at variable selection. If the number of variables is restricted, a decision on how to group the countries in terms of their energy mix seems plausible. But when further variables are added, the complexity increases significantly. Thus, clustering algorithms are a helpful tool to handle this complexity and discover unexpected or hidden patterns within the data. Also, an appropriate number of clusters can be identified easily with the help of data mining tools. In addition, the criteria and number of the clusters remain the same for each year, so that country paths throughout the whole time span can be identified as it was defined in the method of Csereklyei, Thurner, Langer, and Küchenhoff (2017). As the country paths describe the membership of a country to the clusters over time, they allow an easy comparison of the results from different years.

In their paper Csereklyei, Thurner, Langer, and Küchenhoff (2017) "find that countries tend to take a path towards higher quality energy mixes over time, however path inertia and dependencies arise from both infrastructure and resource endowments" (p. 442). This makes it reasonable, that due to political, environmental, or catastrophic incidents, unobvious changes in energy mixes might have occurred and that data mining methods can represent an efficient method to include these patterns in the analysis.

To follow a clear structure in the analysis of the European energy production mix a recognized framework must be applied. As outlined before a typical approach to implement a data mining project is KDD. For this reason, KDD will represent the first pillar of the methodological framework of this analysis. This will be combined with a second pillar, the machine learning process illustrated by Taulli (2019) which extends the data mining step of KDD and forms the final methodological framework in the scope of this analysis. However, as KDD does not consider any data exploration phase, it is added to the framework as the second step between the data selection and preprocessing. The scheme of this extended KDD process is illustrated in Figure 15. It represents a clearly structured, comprehensive and promising foundation to answer the research questions in a satisfactory manner.

Figure 15: Extended KDD Process Scheme

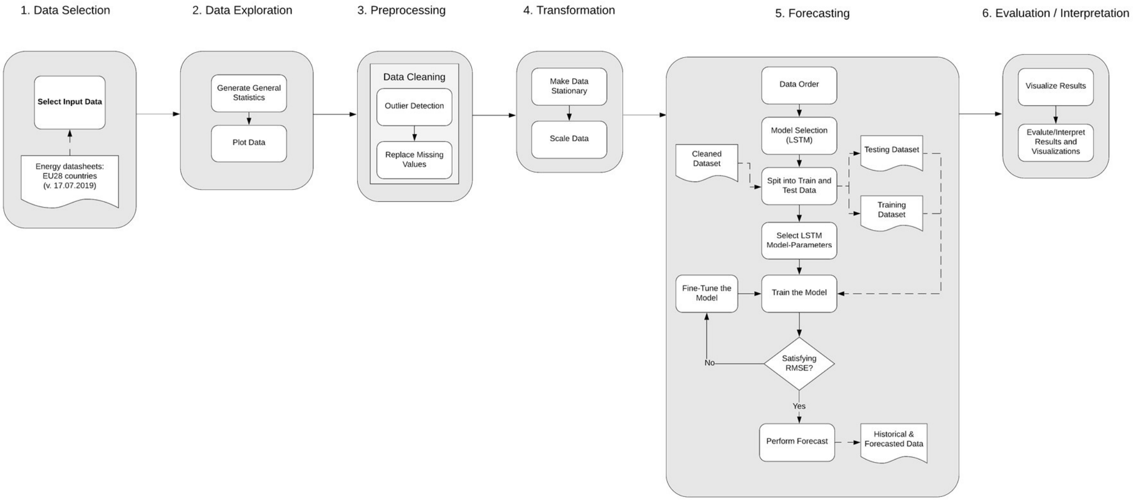


Source: Own work.

Next to applying a solid methodological framework, an appropriate software stack is necessary to implement machine learning models. Since the forecast and analysis are implemented within a Python project, Spyder was chosen as an adequate integrated development environment (IDE). Spyder comes along as a part of the Anaconda distribution, which is a platform specifically tailored for the needs of data scientists. For the implementation and execution of the project scripts, a new project environment is set up in Anaconda. Within this environment necessary toolsets can be installed. In the scope of this project Spyder is installed exclusively as an application. Furthermore, the necessary dependencies, like Pandas, scikit-learn or Keras, can be installed and kept up to date through the Anaconda distribution.

Having defined the project framework and development environment, the methodology of both the forecasting and clustering process can be characterized more into detail. Since the clustering process is built on top of the forecasting process two independent processes for forecasting and clustering must be defined. Therefore, at the beginning a forecast until 2050 for each member state will be implemented. Based on this implementation, the data will be allocated to different clusters, in order to get more insights into the data and identify how the energy sector of the EU-28 changed throughout the years and might change in the future. For this reason, the previously defined framework can be applied to each process individually. The use of the framework in the forecasting process, is illustrated in Figure 16 with a higher level of detail.

Figure 16: Flow-Diagram Forecasting Process



Source: Own work.

To begin with, the necessary data needs to be selected. To carry out the forecast of the European energy sector, the biannual energy statistical country datasheet (version of 17.07.2019) published by the European Commission is a primary data source. This dataset, which is stored as a Microsoft Excel file, contains a long-term time series, with yearly data for each European Union member state on:

- Energy Balances
- Electricity
- Main Energy Indicators
- Cogeneration
- Transport Fuels
- Greenhouse Gases Emissions

Each country datasheet contains observations for the period of 1990 to 2017. As the datasets are large, not all information in the datasheets is relevant for the analysis in this thesis. However, the main variables of the member states' energy carriers, emissions, import and exports can be acquired here. These variables of interest are the ones which are related to the main goals set in the European Union directives or might contribute to a further analysis of the research goals. By processing the data, it is structured in a 2-dimensional matrix for each country. The y-axis of the matrix describes the year and the x-axis the variables. In a final step the data of each member state is exported to a csv file.

After the data integration is concluded, the dataset undergoes the preprocessing. While a basic exploration is performed to understand the data and its characteristics better the data cleaning follows. At this point, it is crucial to identify possible errors, like missing values or outliers. Following, in the data preparation step the data must be cleaned to remove these errors and acquire a high data quality for the remaining process.

In the transformation step at first the data must be made stationary. While S. Li (2017) states that stationarity is not essential for LSTMs, she emphasizes that its presence increases the performance and learning rate of the model. Also, since the variables are measured on different scales, e.g. energy production in TWh and emissions in million tons CO_2, feature scaling is applied as another preprocessing step.

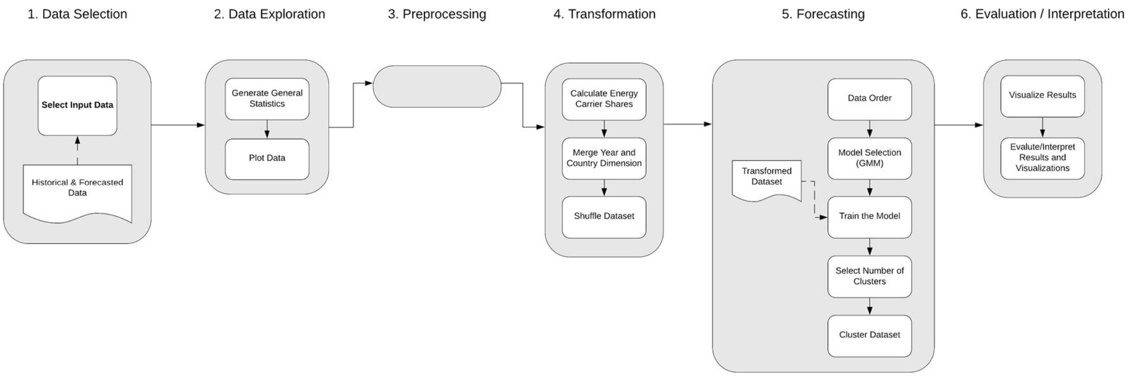
Based on the preprocessed and transformed data the actual forecast model can be developed. In this process step one model is developed for each EU member state. As described by Taulli (2019), the first step to forecast the data is to check for the data order. However, as previously defined, the model choice fell on a LSTM network. The advantages of ANNs lay in their ease of implementation, a clear model and their high performance (Đozić & Gvozdenac Urošević, 2019). Provided that LSTM networks are implemented, the definition of the data order step, as described in Chapter 2.4.2, must be inversed. By the fact that an LSTM model is dependent on an ordered dataset in the forecasting process, the time series are checked for the chronological order in this step instead.

Afterwards the dataset of each member state is split into a training and testing data to train and perform evaluation of the model accuracy. In the next step the initial parameter values of the model need to be defined. With the parameter values set, the model is ready to be trained a first time on the training set and tested on the testing dataset. As an accuracy measure, the RMSE is considered. Based on the error of the training and testing data, the model is either assessed as satisfying and can be applied to perform the forecast or the model parameters need to be fine-tuned to enhance the model quality. Setting best possible model parameters of the LSTM networks is a trial and error process and might be long-lasting.

Once the trained model outputs satisfactory results, it can be applied to execute the actual forecast until 2050. As a result, the output of the forecasts is a dataset that comprises data of each country for 61 years, where 28 years represent historical data and the remaining 33 years forecasted data. To conclude the forecasting, an initial evaluation and interpretation of the results is performed based on first visualizations and statistics.

At the same time the output dataset of the forecasting process represents the input dataset in the data selection step of the clustering process (see Figure 17). To get additional insights regarding the clustering process, which might not have been discovered at the end of the forecasting procedure, a data exploration phase is included after selecting the data. As this data is a product of the prior process further preprocessing steps are unnecessary at this point.

Figure 17: Flow-Diagram Clustering Process



Source: Own work.

Instead the step is skipped, and the data transformations are executed to prepare the data for the following clustering sub-process. To analyze the development of the energy production, the clustering method that Csereklyei, Thurner, Langer, and Küchenhoff (2017) have introduced in their study is adopted. For this reason, some adjustments need to be made to transform the data in an appropriate way. At this point the production of each energy carrier is just represent in absolute numbers, which is why the shares of each energy carrier are calculated as the initial action of this step. As the data moreover still has the dimensions country, year and energy carrier, the country and year dimension need to be merged together, as outlined by Csereklyei, Thurner, Langer, and Küchenhoff (2017). The last data transformation is the shuffling of the dataset.

By applying this measure, the fulfillment of the first clustering step, the data order step, is already ensured. As it has been defined, the clustering process is based on the methodology of Csereklyei, Thurner, Langer, and Küchenhoff (2017). Thus, a GMM algorithm is selected as the clustering method in this analysis as well. Even though the goals of the paper differ, the method applied by Csereklyei, Thurner, Langer, and Küchenhoff (2017) is a suitable approach to investigate the behavior of national energy mixes. The method is especially suitable since it enables a comparison of different years and countries, since "the clusters represent the same concept over the successive years" (Csereklyei, Thurner, Langer, & Küchenhoff, 2017, p. 446).

After the model is trained, the optimal number of clusters needs to be determined. Hereby, different methods are employed to verify which number of clusters delivers the best results.

Unlike Csereklyei, Thurner, Langer, and Küchenhoff (2017) who applied an elbow graph and an adjusted BIC (aBIC), in this analysis the silhouette coefficient and the adjusted BIC are applied.

The BIC is defined by Ahlquist and Breunig (2012) as:

$$BIC = 2\log\mathcal{L}(x, \hat{\theta}_G) - m_G \log n \tag{8}$$

In this context, $\hat{\theta}$ represents the MLE and m the free parameters in the model. The selection of the model with the best parametrization is hereby chosen by comparing the BIC scores. The model that maximizes the BIC is considered to work the best for the dataset. Still, a drawback of the BIC is that it assigns a higher penalty to more complex models and therefore simple models are privileged (Ahlquist & Breunig, 2012).

To adjust this penalization, Csereklyei, Thurner, Langer, and Küchenhoff (2017) modified the equation of Ahlquist and Breunig (2012) to:

$$aBIC = 2\log\mathcal{L}(x, \hat{\theta}_G) - 3m_G \log n \tag{9}$$

By the fact that the following clustering process is mainly based on the methodology of Csereklyei, Thurner, Langer, and Küchenhoff (2017) and since the authors already proved that the measure is applicable to the present data, the aBIC is also applied for the cluster selection in the scope of this thesis. The following steps of evaluation and fine-tuning of the model are in this case not included in the clustering process by the non-existence of a validation set.

In the clustering process the concluding step is to evaluate the results by generating meaningful visualizations and interpret the results. Within this step the data is also restructured to obtain amongst others the energy paths of the EU-28. Moreover, due to the close interrelations of both modelling processes and to provide comprehensible results, all outcomes are evaluated and discussed at the end of the chapter. Hereby, with respect to the research questions a closer look must be taken at the country paths to identify patterns or anomalies in the target years of the EU energy roadmaps: 2020, 2030 and 2050.

3.2 Data

As it has been outlined, prior to the actual data mining step, the necessary data needs to be selected, explored, prepared, and transformed to acquire plausible results. Within this chapter the implementations and gained insights on the input data are outlined in detail.

3.2.1 Data Selection

The data for this paper is obtained from the biannual energy statistical country datasheet (version of 17.07.2019) published by the European Commission. For the analysis, the data of the following variables was selected and extracted from the Excel file and saved as a comma-separated values (csv) file:

1. Country Code
2. Year
3. Coal Energy Production (in TWh)
4. Oil Energy Production (in TWh)
5. Gas Energy Production (in TWh)
6. Nuclear Energy Production (in TWh)
7. Renewable Energy Production (in TWh)
8. Waste Energy Production (in TWh)
9. Energy Imports (in Mtoe)
10. Energy Exports (in Mtoe)
11. CO_2 Emissions (in mio ton CO_2)
12. GHG Emissions (in mio ton CO_2)
13. Overall Renewable Share in Gross Final Energy Consumption (in %)

The production variables of the dataset are expressed as terawatt-hours (TWh). Based on the production variables the energy paths of each member state can be determined in a later stage. To get insights on a possible dependency on third countries, imports and exports are included in million tons of oil equivalent (Mtoe). A Mtoe corresponds to 11,63 TWh. CO_2 and GHG emissions provide information about the amount of these harmful substances that was produced and released into the environment. Both emission indices are measured in million tons CO_2. The renewable share is represented as the percentage renewable energy types have in the gross final energy consumption. The knowledge that the share of renewable energy is the indicator that is used by the EU to measure the development of its member states' renewable share and that the calculation of the gross final energy consumption is a top-down calculation approach make it reasonable to include the indicator into the analysis to get meaningful results regarding the renewable shares. This also allows a better comparison of the forecasted shares with the short-term targets as defined in the Directive 2009/28/EC (see Appendix 2) and long-term targets of the 2030 and 2050 energy strategies.

3.2.2 Data Exploration

After the selected data is acquired a first action is to control the assigned data types. This first check showed that all variables were automatically assigned as objects. To make the data actionable the data types must be changed appropriately. The years are defined as integers and the country codes as strings. As it has been elaborated, the remaining data represents numerical values. Therefore, float64 is assigned as the new data type of these variable.

Subsequently, basic statistical details can be obtained by applying the ".describe" method to the dataset to receive the output that is presented in Figure 18. To simplify the readability of the table, a color scale is applied to each statistical measurement of the energy carriers and to the count of all variables.

Regarding the count of the variables it is primarily notable that every variable except for the renewable share consists of 784 observations. For the renewable share only 392 observations are present instead. This is a clear sign of missing values. Moreover, the other takeaway is that the remaining variables do not include any missing values since the dataset contains data of 28 member states for 28 years.

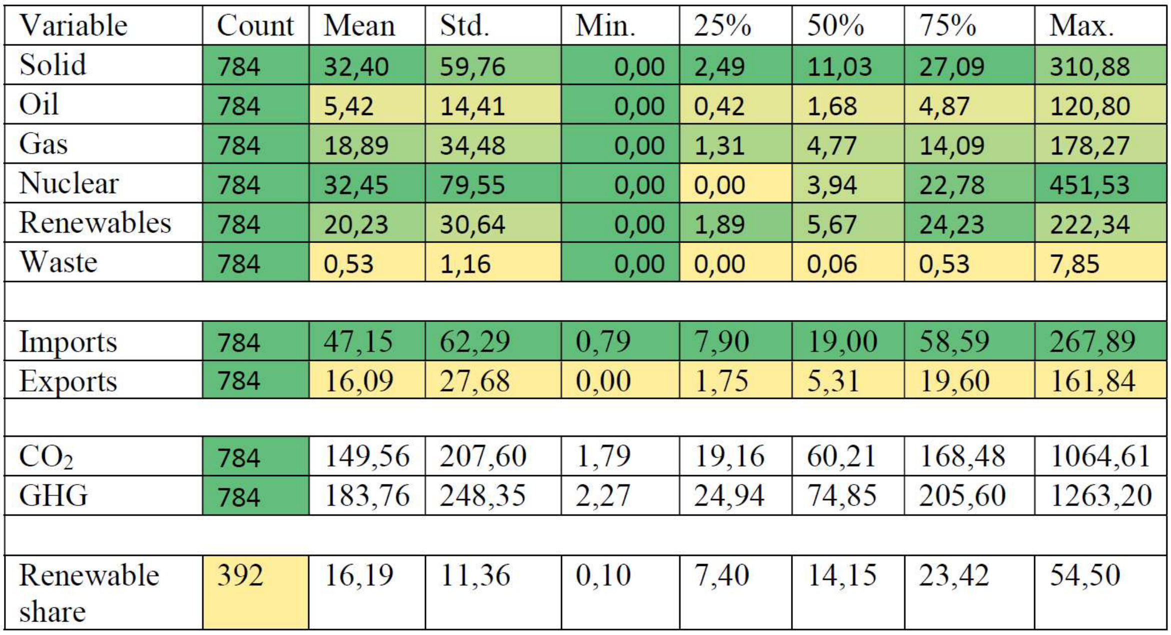
If the statistics of the energy carriers are examined in detail, it must be noted that solid and nuclear energy are the most widely used energy carriers on average among the EU-28 during the whole observation period. While their mean accounts for approx. 32 TWh, renewables follow with approx. 20 TWh and gas with 19TWh. Just oil with approx. 5 TWh and waste with 0,53 TWh remain far behind the other energy carriers.

All energy carriers show relatively high standard deviations, which indicates a high variation in the data. This is plausible since this dataset comprises the energy mixes of countries with varying sizes and energy strategies over a long time period. Besides that, each energy carrier owns at least one country-year observation in which the minimum value is zero. On the opposite the values of the maximum energy production are rather high in comparison to the average, which correlates with the high standard deviations.

The overall imports of the EU-28 are roughly three times higher than the exports with a mean value of 47,15 Mtoe. Additionally, the variation of both, imports and exports, is relatively high. While for at least one country-year observations there are no exports at all, the minimum amount of energy imports amounts to 0,79 Mtoe. Simultaneously, observations with a maximum of approx. 267 Mtoe imports and 162 Mtoe exports exist.

Regarding the emissions, Figure 18 illustrates that the mean GHG emissions (183,76 mio tons CO_2) are higher than CO_2 emissions (149,56 mio tons CO_2), which is an obvious result since CO_2 emissions are a part of the composition of GHG emissions. The standard deviation is hereby relatively high, with 207,60 (CO_2) and 248,35 (GHG). Minimum values of 1,79 (CO_2) and 2,27 (GHG) as well as maximum values of 1064,61 (CO_2) and 1263,20 (GHG) show the existence of large differences among the observations.

Figure 18: General Statistics of the Input Dataset



Variable	Count	Mean	Std.	Min.	25%	50%	75%	Max.
Solid	784	32,40	59,76	0,00	2,49	11,03	27,09	310,88
Oil	784	5,42	14,41	0,00	0,42	1,68	4,87	120,80
Gas	784	18,89	34,48	0,00	1,31	4,77	14,09	178,27
Nuclear	784	32,45	79,55	0,00	0,00	3,94	22,78	451,53
Renewables	784	20,23	30,64	0,00	1,89	5,67	24,23	222,34
Waste	784	0,53	1,16	0,00	0,00	0,06	0,53	7,85
Imports	784	47,15	62,29	0,79	7,90	19,00	58,59	267,89
Exports	784	16,09	27,68	0,00	1,75	5,31	19,60	161,84
CO_2	784	149,56	207,60	1,79	19,16	60,21	168,48	1064,61
GHG	784	183,76	248,35	2,27	24,94	74,85	205,60	1263,20
Renewable share	392	16,19	11,36	0,10	7,40	14,15	23,42	54,50

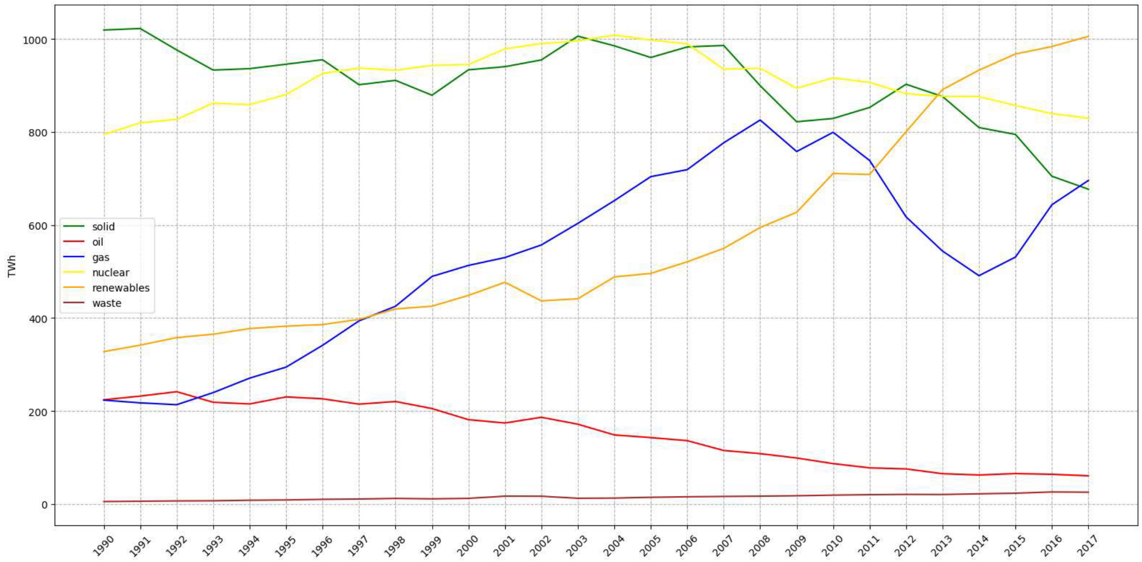
Source: Own work.

The renewable shares of the EU-28 accounts for 16,19% on average for the historical data. If the dataset is examined more in detail, it becomes clear that the renewable share was calculated by the EU from 2004 onwards. The variation of the shares is with 11,36 relatively low. In addition, the minimum share among the member states amounts to 0,1% and the maximum share to 54,5%.

If the energy production of each energy carrier is summed up by year, a brief insight into the overall energy mix of the EU can be achieved. Figure 19 represents the development of the historical European energy mix. This enables an examination of several key characteristics without the need of any deeper analysis.

The energy production through waste was always very low for the whole observation period. Also, energy produced from oil, which accounted for more than 200 TWh at the beginning of the records, is slowly tending towards zero. The gas energy production was in the early 90s' on a similar level with the energy from oil. However, unlike oil production gas became a more important energy carrier with a continuously increasing production until 2008, where the peak of around 800 TWh was reached. Afterwards a downward trend started until 2014. From then on, the production has been increasing again. Nuclear and solid resources were the dominant energy carrier for most of the time with a production varying between 800 and 1.000 TWh. Still, Figure 19 highlights that their dominance was replaced by renewable energies from 2013 on. The increase of the renewable energy production was progressing steadily and slowly until 2005. Afterwards, the production could record significant increases with a production of approx. 1.000 TWh in 2017.

Figure 19: Sum of Historical Energy Production of the EU-28 by Energy Carrier



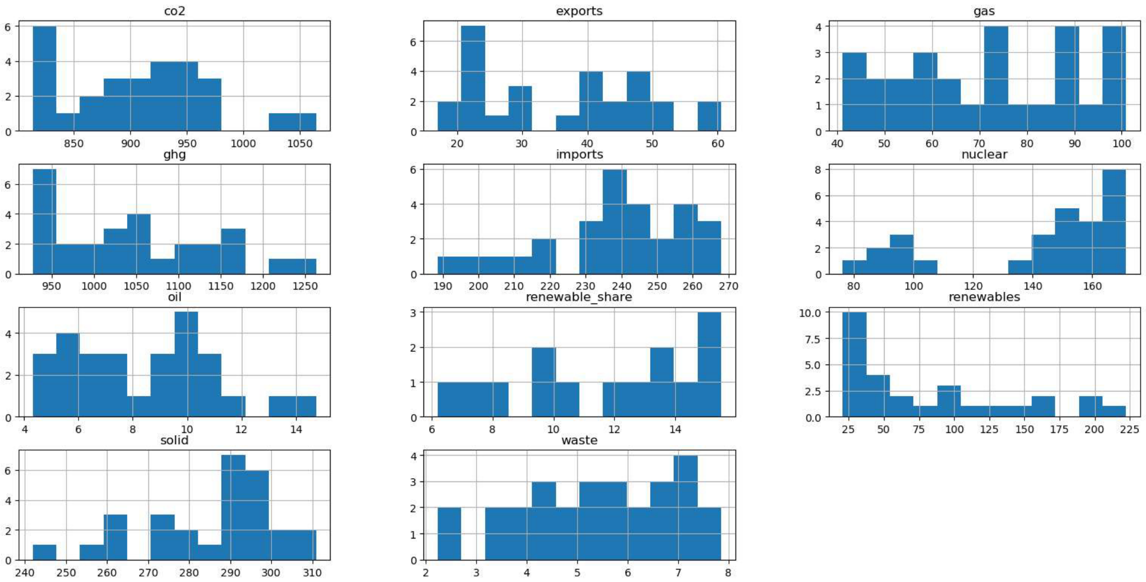
Source: Own work.

3.2.3 Data Preprocessing

To preprocess the data for the application of machine learning algorithms, it needs to be cleaned from errors in the first instance. For this purpose, the data of each country is treated separately. By visualizing the individual variables in a histogram, potential outliers can be identified. Exemplary, Figure 20 portrays the histograms of Germany. The histogram indicates that no data point lays far outside. This leads to the conclusion that the subset of Germany does not contain any outliers. Similarly, this graphical representation was created for the remaining member states to control their data for outliers.

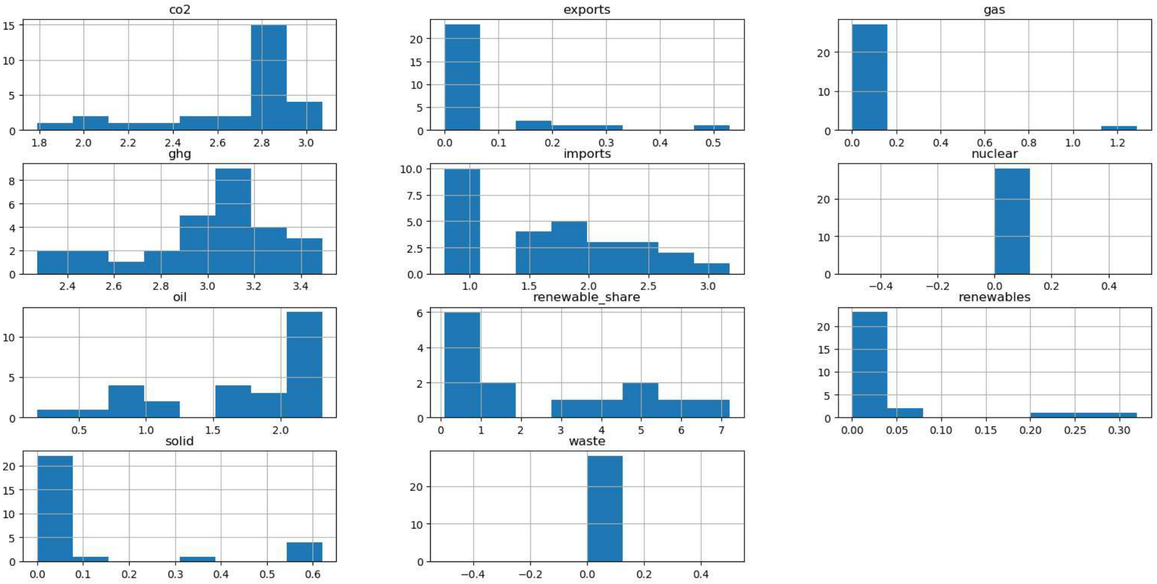
Even though, some of the histograms show a tendency for outliers. When the specific data is closely examined, the observations appear as outliers since the country was going through a general change in the composition of the energy mix at that time. To not lose this essential information these outliers are not excluded from the dataset. An exemplary and challenging case represents the dataset of Malta (see Figure 21).

Figure 20: Histogram-Plot of the Input Variables for Germany



Source: Own work.

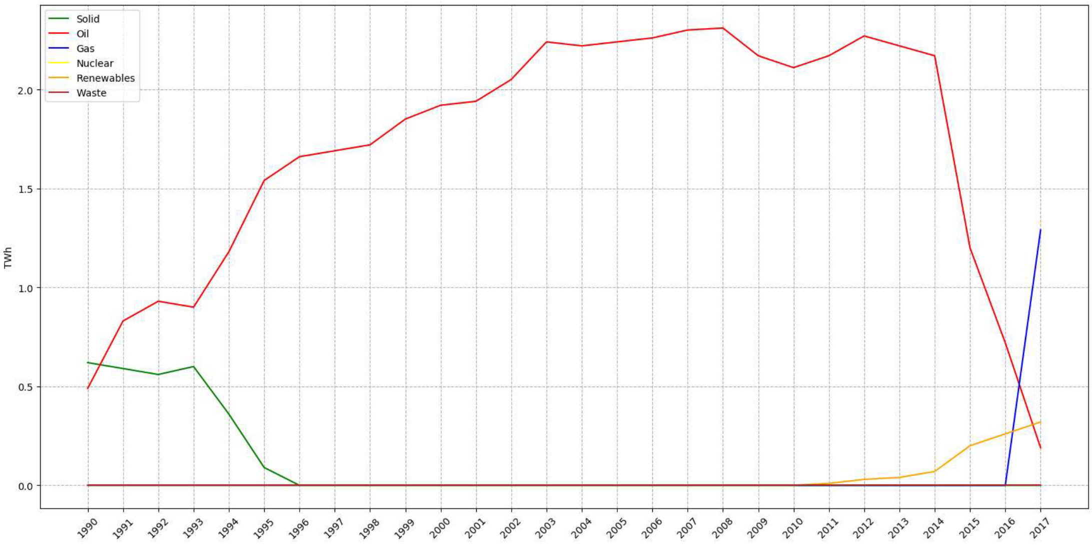
Figure 21: Histogram-Plot of the Input Variables for Malta



Source: Own work.

The line plot in Figure 22 visualizes the historical development of Malta’s energy mix. The energy mix consistently included a high share of an energy production from oil until 2014. Additionally, the energy production in Malta was solely based on oil for many years. Just at the beginning of the observation period, solid energy carriers were part of the island state’s energy mix until they fell to zero in 1996. Conversely, in the final part of the historical data series energy from gas and renewables was on the rise, whereas the production was equal to zero until 2011 and 2017, respectively. On the contrary, the energy production from oil dropped rapidly to almost zero in between 2014 and 2017. These observations clearly indicate the presence of outliers. However, if the overall picture is considered the changes in the energy mix cannot be considered as wrong since the decreasing of one energy carrier occurs in accordance with the increasing production of another one. This observed by considering the total energy production which just decreases slightly during the questionable years. Still, these minor differences can be explained by higher energy imports and constant low exports, so that even a growth of the total available energy can be observed. Nevertheless, Malta represents a special case, in which the historical energy mix was consistently similar until its changes were initiated just in more recent years.

Figure 22: Line-Plot of the historical Data of Malta



Source: Own work.

On one hand, the detected outliers contain valuable information and a replacement would lead to a significant information loss regarding possible forecasts. On the other hand, keeping the outliers might cause the forecasting model to be insufficient for Malta. To take a carefully considered decision on how to process Malta's data, this special case is examined again when the forecasting model is applied to the data later on.

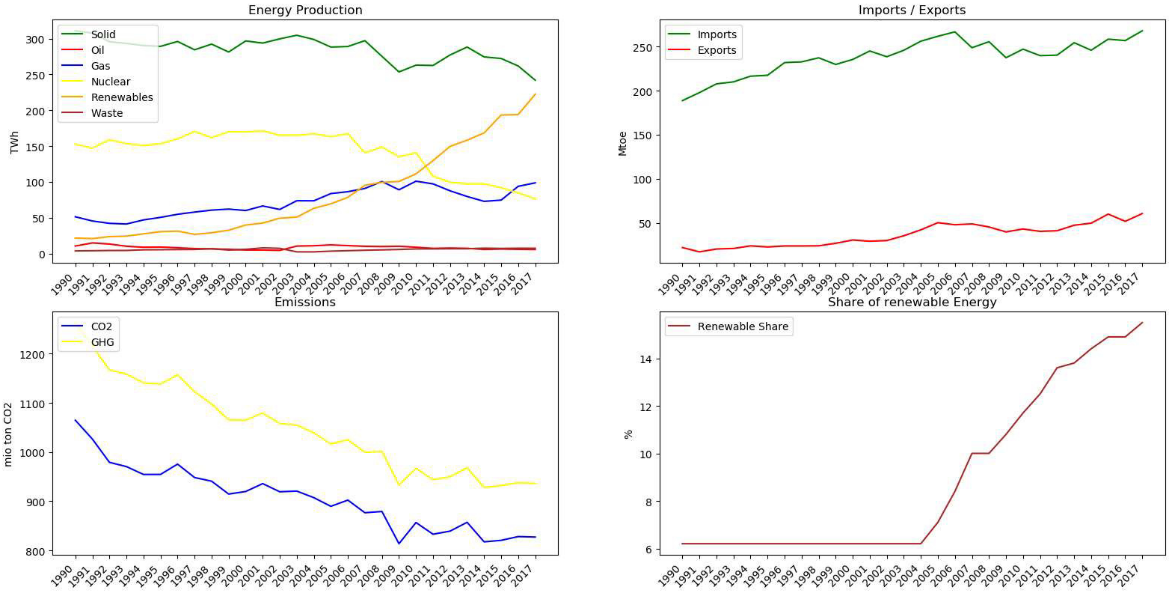
To conclude the preprocessing step, the treatment of missing values is implemented. As it was examined before, the only variable containing missing values is the overall renewable share in gross final energy consumption. The observations with missing values are not eliminated because a deletion would imply a very large information loss. As a consequence, alternative values must be imputed. The choice fell on a rather simple method by imputing the minimal values. While linear interpolation is generally a good choice for time series data with a trend, it is not quite suitable in this case as the values are missing consecutively during the first 14 observation years. Moreover, the choice of using minimal values is based on two reasons. First, renewable energy sources are on the rise since the early 2000s, so that it can be assumed that in the renewable share of the previous years was in general lower than the minimum renewable share. If the data is examined in detail it also becomes clear that the minimum value occurred usually in the first years of the available historical data. Furthermore, the fact that more recent years have a higher importance in the forecast makes it reasonable to impute minimal values is a sufficient approach in this analysis.

3.2.4 Data Transformation

To complete the data preparation steps, the data needs to be transformed to make the forecasting model efficiently applicable to each member states' subset. To achieve this, the data is made stationary and subsequently scaled.

Since time-series predictions can be implemented more efficiently with stationary data, the input dataset must be controlled for this attribute. First, it needs to be checked if a time-series shows a trend or seasonality. If this is the case the data can be characterized as non-stationary. A method to simplify the prediction problem and ensure stationarity is differencing. Resuming the previous example of Germany, a simple plot of the data subset (see Figure 23) allows us to get insights on trend and seasonality. While seasonality can be generally excluded by the fact that we work with yearly data, the data must be controlled for a trend feature only.

Figure 23: Line-Plots of the historical Data of Germany



Source: Own work.

The example of Germany with its subplots indicates that the data follows different trends and no seasonality. When applying this procedure to each country, similar results can be obtained. To ensure stationarity, all data subsets are transformed by differencing to improve the efficiency and the results of the following forecast.

The final issue that needs to be solved is the fact that the data is measured on four different scales. In time series forecasting, especially with neural networks, scaling the data prior to the modelling can lead to significant performance increases. The scikit-learn package provides the most common scaling methods to solve this issue, like z-score normalization or min-max normalization. For this specific forecasting problem, the choice fell on implementing the inbuild "MinMaxScaler" method, which is used to scale and translate each feature in a predefined range which in this case will lay between 0 and 1.

3.3 Modelling

With the preprocessed dataset, the model for the energy production forecast can be implemented and applied to the dataset. Chapter 3.3.1 describes the forecasting models and the adjusted parameters more into detail. After applying the models to the appropriate data subsets, a comprehensive dataset for the period of 1990 until 2050 will be the result. As an intermediate step, the data needs to undergo minor preprocessing steps to prepare it for the upcoming clustering process. The model based on the EM algorithm is developed and applied to cluster the energy production mix in the subsequent Chapter 3.3.3. This final measure is going to allow for an identification of the energy paths in the EU.

3.3.1 Energy Production Forecasting

As outlined before, a LSTM neural network will be utilized to carry out the projection. The package that is applied for this purpose is Keras. For the implementation of the model, the following five basic steps are necessary to achieve best possible results:

- Network Definition
- Network Compilation
- Network Fitting
- Network Evaluation
- Prediction

When defining the network, the first decision regards the application of the best fitting model type. In this case a sequential model is used. A sequential model stacks several user-defined layers whereas the first layer expects the input shape of the data to be defined. Since the input dataset consists of 13 variables, where the country code and the year act as identifiers. Hence, the remaining eleven variables need to be forecasted and these will act as input variables which in this context also take the name of "features".

In the procedure, first, the input data needs to be reshaped, to fit the model expectation of a three-dimensional shaped input data. The three dimensions are composed by samples, timesteps and features. Before defining the model itself, the initial seeds are set for the algorithm to allow a later reproducibility of the results.

While defining the model the following four crucial factors can influence the results significantly and these are the parameters which are adjusted throughout the process:

- Number of LSTM layers
- Number of Neurons
- Number of Epochs
- Percentage of the Validation Split

The parameters are adjusted by a trial and error process in the ongoing model development.

As stated above, several layers can be stacked when developing a sequential LSTM model. In this analysis a model that stacks two LSTM layers and one dense layer is deployed since it generates the best results with the present data. Adding a fully connected dense layer, which is stacked on top of the LSTM layers, is necessary to output the final prediction. Given that we deal with 28 different subsets of our dataset, it is obvious that there is no universal model for each subset. For this reason, the remaining parameters need to be adjusted for each member state individually and therefore 28 different forecasting models are the result. A detailed summary of the set model parameters for each country are listed in Appendix 3.

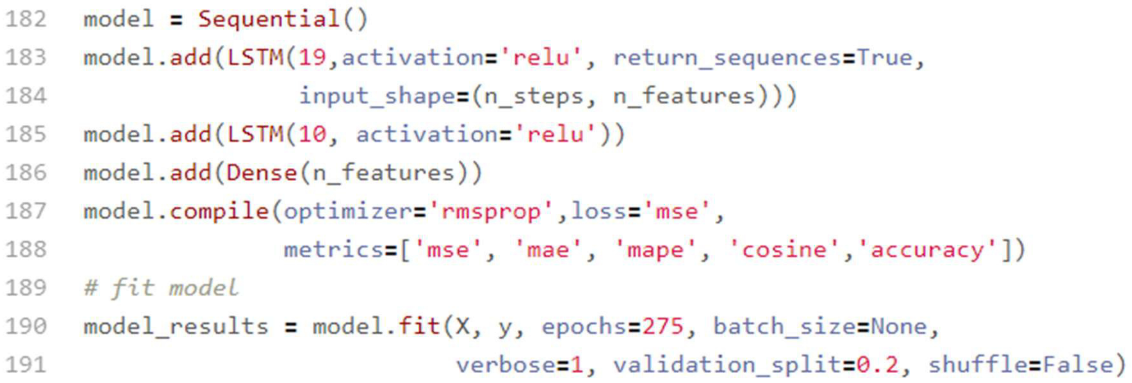
To outline the process of selecting the best LSTM model parameters, the exemplary case of Germany is resumed. To get a better understanding of the model definition the final code of Germany's LSTM model is illustrated in Figure 24 and this will be illustrated in the following lines.

As a first measure, the model is defined as sequential in code line 182. The following lines 183-186 are adding two LSTM and one dense layer to the model. The model that performed best for this specific dataset possesses 19 neurons in the first and 10 neurons in the second LSTM layer. Hereby both layers are applying the "ReLU" activation function. The first LSTM additionally needs two further arguments to be defined. By setting the return sequence to true, a sequence of vectors of 19 dimensions is returned in the next layer. Next to that, the shape of the input data is defined with the number of steps and the number of features which were predefined and saved as variables.

After completing the definition phase, the model must be compiled with lines 187-188. For the compilation, the model optimizer is set to "rmsprop" and the MSE is assigned as the loss function. Besides that, further metrics are calculated.

In the final step of the model creation, the fitting of the model is implemented. Here the input data is selected and divided into a training and testing subset. A common approach for the validation split is to use 20% of the data for testing and 80% for training. Therefore, this ratio is adopted as the default value of each model. If a dataset makes it necessary, the proportion is adjusted accordingly. In the case of Germany, the value of 20% performed well and did not force any adjustments. The code allows to define if the dataset should be shuffled. In our case it is not shuffled since we are working with time-series data that needs to remain ordered. The number of epochs that worked out the best for the dataset of Germany is 275.

Figure 24: LSTM Model of Germany

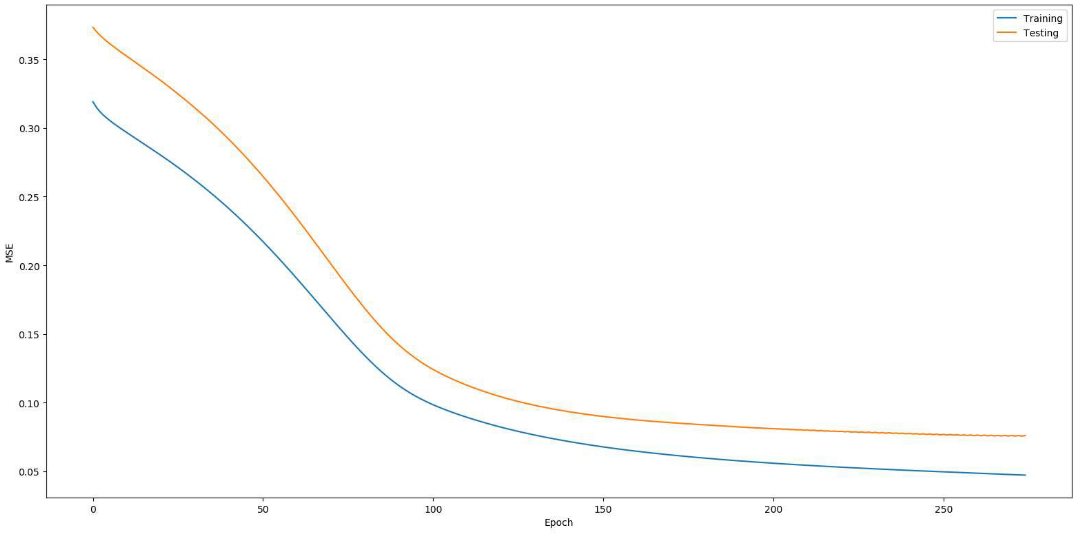


```
model = Sequential()
model.add(LSTM(19,activation='relu', return_sequences=True,
                input_shape=(n_steps, n_features)))
model.add(LSTM(10, activation='relu'))
model.add(Dense(n_features))
model.compile(optimizer='rmsprop',loss='mse',
               metrics=['mse', 'mae', 'mape', 'cosine','accuracy'])
# fit model
model_results = model.fit(X, y, epochs=275, batch_size=None,
                           verbose=1, validation_split=0.2, shuffle=False)
```

Source: Own work.

When the code is executed, the MSEs of the training and testing data are visualized through a line plot as illustrated in Figure 25. The plot shows that the learning rate was high in the approximately first 100 epochs. It slowly decreases afterwards and ends up with a MSE value of 0,0472 for training and 0,0756 for testing in the final epoch. The closer the testing error is to the training error the better it becomes to control the over- or underfitting of the model. For instance, if the line of the testing MSE would drop lower than the training MSE this would be an indicator of an overfitting model. Issues like these are resolved during the fine-tuning step of the member states' models. The final MSEs of each member state are listed in Appendix 3 as well.

Figure 25: Training and Testing Error (MSE) Plot of Germany



Source: Own work.

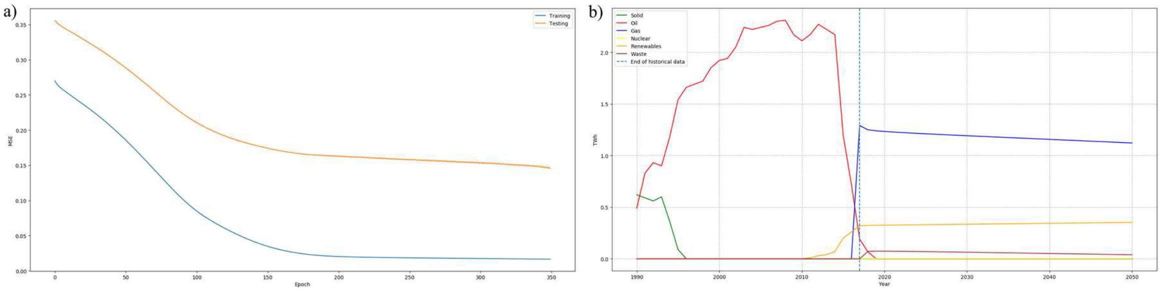
After fine-tuning the model, the last step consists of executing the actual forecast. A multi-step time series forecasting strategy needs to be chosen to support a forecast in covering more than one year only. As Bontempi, Ben Taieb, and Le Borgne (2013) state, there are four general strategies to approach a multi-step forecasting problem:

- Direct strategy
- Recursive strategy
- DirRec strategy
- Multiple output strategy

While in a direct strategy multiple models are developed to forecast multiple time-steps, the recursive strategy applies a one-step forecast model recursively for several time steps (Bontempi, Ben Taieb, & Le Borgne, 2013). If both strategies work together as a hybrid model this is called a DirRec strategy (Bontempi, Ben Taieb, & Le Borgne, 2013). Unlike the other three strategies, the application of a multiple output strategy enables a single model for direct forecast for the whole sequence (Bontempi, Ben Taieb, & Le Borgne, 2013). The choice in this forecasting problem fell on a recursive strategy as it was already employed successfully in many real-world problems which also included RNNs (Bontempi, Ben Taieb, & Le Borgne, 2013). A line plot of the historical and forecasted energy production data for Germany is visualized in Appendix 4. After proceeding similarly with each subset of all member states, the data transformations must be reversed to make the data readable again and prepare it for further processing.

One issue that needs to be elaborated again at this point is the previously introduced case of Malta. As it was afore assumed, the application of a LSTM network to forecast the data of this island state causes several problems. While fine-tuning the model, the MSE plot (see Figure 26a) indicates that the model is able to learn well on the training data with an MSE value of 0,0166. However, at this point the assumption that the rapid changes in the last years of the historical data might bias the model results can be confirmed. This means that the model is not able to minimize the testing error at a satisfying level.

Figure 26: MSE Plot (a) and Forecast (b) of Malta



Source: Own work.

The optimum that could be achieved by tuning the parameters is a testing MSE of 0,1453. From approximately epoch 200 both MSEs run in an almost straight line horizontally to the x-axis and are not approximating each other. This indicates that the training dataset is not representative of the testing dataset. Possible approaches to solve this issue could be the introduction of additional data points. This unfortunately is not possible in this analysis. An alternative is the adjustment of the train-test-split. However, when applied, this method did not improve the results and the issue remains. The final confirmation that the model does not work with the dataset of Malta is given by the forecast. Figure 26b illustrates the historical and forecasted data of each energy carrier. At first sight it can be noticed that the forecast of some energy carrier is in a clear contrast to the trend at the end of the historical time-series data. In detail, the change from oil energy carriers to gas is not represented well. While the trend of the oil energy production seems to fit the trend towards zero, the energy produced with gas, which was on the rise at the end of the historical data, just stagnates and slightly decreases. Also, an energy production with waste, that was never present in the historical data, suddenly appears in the forecast. Considering the fact that this data is clearly biased, and a representative model cannot be developed with the present methodology, Malta is excluded from the further analysis.

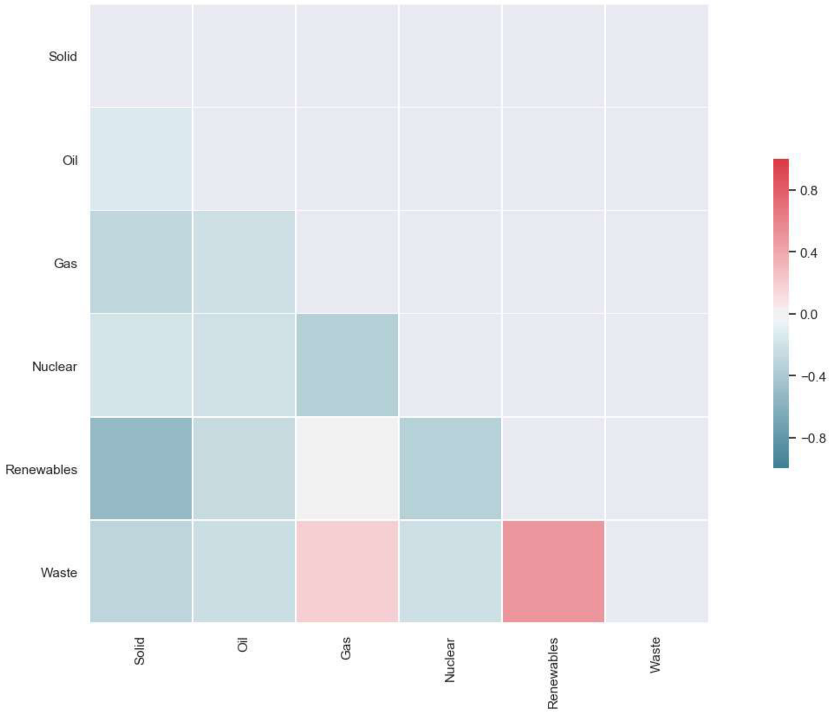
3.3.2 Intermediate Data Exploration and Transformation

After performing the forecast of each member state, the data subsets need to be merged again. As the goal is to apply an EM algorithm to cluster the energy production shares, first the variables that do not represent the energy carriers are excluded from the dataset. Subsequently the shares of the energy carriers of the total energy production are calculated. To prepare the data for clustering over countries and time, the country and year dimensions are merged. The resulting dimension is assigned as the new identifier of the dataset.

Before clustering the data, the variable correlation hypothesis is tested. On one hand, this is necessary to get a better understanding of the data and its relationships. On the other hand, if a very high correlation between any variables is observed, the application of dimension reduction might be a necessary step.

To test correlation the Pearson correlation coefficients are calculated. The results are then illustrated in a correlation matrix as illustrated in Figure 27. The matrix shows that there is neither a very high positive nor negative correlation between any of the variables. The strongest relationship exist between the renewable and solid energy carriers with a r-value of -0,52 and the renewable and waste energy carriers with 0,5. For this specific example it would mean that with an increasing renewable energy production also the waste energy production increases significantly and at the same time the solid energy production decreases. However, since no high correlation exists between all variables there is no need to apply a dimension reduction algorithm before executing the clustering process.

Figure 27: Correlation Matrix Energy Carrier Shares



Source: Own work.

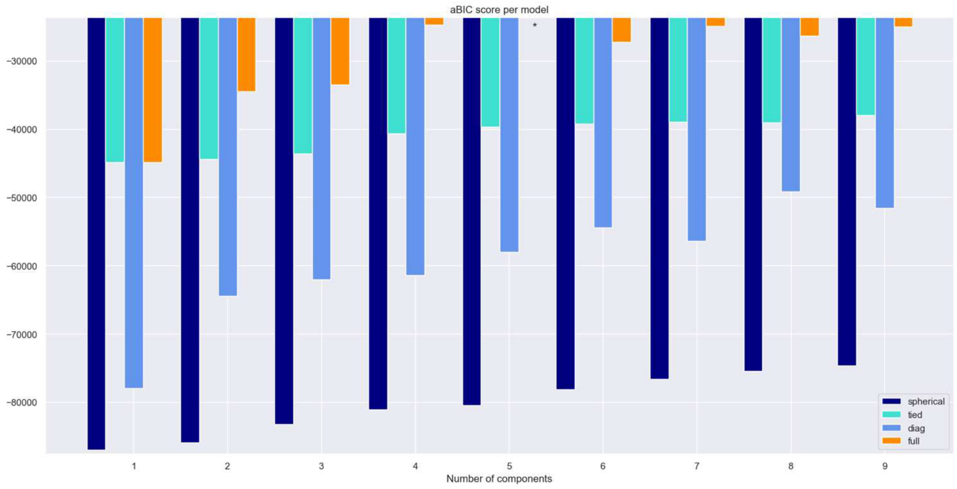
As it was mentioned initially, no extensive preprocessing steps are necessary in between the forecasting and clustering processes. The dataset is just shuffled to randomize the order of the observations. This measure enhances the probability that the order of the data does not falsify the results of the clustering algorithm.

3.3.3 Identification of Energy Paths

Once the dataset is randomized the clustering model can be implemented and executed. In an initial step, the optimal number of clusters is determined by calculating the aBIC and silhouette scores.

While Csereklyei, Thurner, Langer, and Küchenhoff (2017) employed the mclust package for R, that provides 14 different initialization procedures. In scikit-learn four different types of covariance parameters can be chosen for the model: spherical, tied, diag and full. To identify the optimal number of clusters each type is considered initially. This ensures that the model with the best performance is chosen subsequently. After setting the seeds for reproducibility, the negative aBIC scores are calculated for models with a range of 10 components and these can be visualized as seen in Figure 28.

Figure 28: Bar Graph of aBIC Scores by Intialization Methods

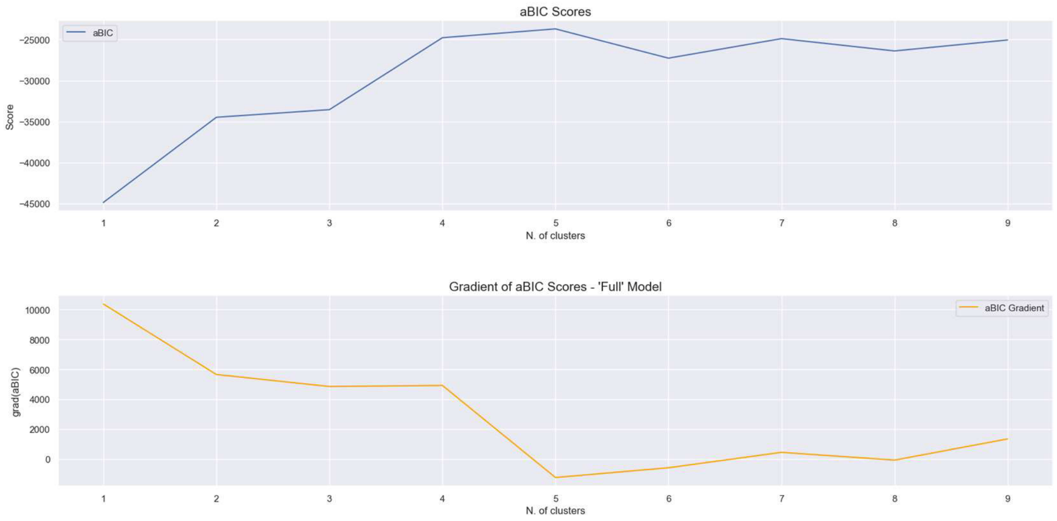


Source: Own work.

By means of the visualization, it is noteworthy to say that in general the procedure with the full covariance initialization performs best. Additionally, it is suggested that the selection of five clusters is most suitable. However, while 5 clusters maximize the negative aBIC score, there is just a minor improvement compare to a four-component model. Apart from that, the seven- and nine-component models show a similarly good score as well. Following, with the results of this graph several models can be considered as ideal. The major insight the aBIC scores provide at this point is that the model with the full initialization procedure works best. To have a better perspective, the results of the full initialization procedure are examined more in detail.

In the case that BIC scores alone do not provide sufficient information, Lavorini (2018) suggests calculating the gradients of the BIC scores. The gradients are calculated by subtracting two consecutive scores (Lavorini, 2018). If the scores are equal, the gradient is zero (Lavorini, 2018). If the scores differ, the gradient is either positive or negative (Lavorini, 2018). Since we are working with the negative aBIC, a lower second value results in a positive gradient and conversely a greater one in a negative gradient. Figure 29 shows the negative aBIC scores and their appropriate gradients of the full initialization procedure models. As stated before, this figure shows more clearly that the negative aBIC scores increase significantly up to the point where they reach the five-component model. In the following minor de- and increases follow. The gradients of the models are reflecting a similar behavior. The gradients are mainly decreasing up to the five-component solution. Beyond this point the values are stabilizing at around zero and thus no major model improvement can be observed anymore. Consequently, the gradient approach confirms the observation that a model with five clusters appeares as the most suitable.

Figure 29: Plots of aBIC Scores and Gradients of the "full" Initialization Procedure Models



Source: Own work.

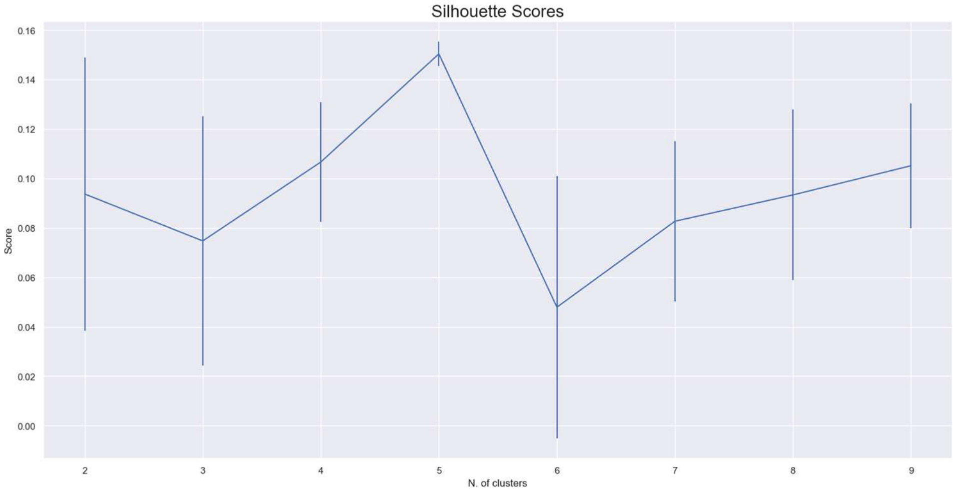
The second technique to verify the number of clusters in this analysis is the calculation of the Silhouette coefficients. The calculation of the coefficient is based on the mean intra-cluster distance of each sample (a_i) and the mean distance between each sample and the next nearest cluster (b_i) (Unpingco, 2019). The silhouette coefficient of a sample *i* is formalized in Equation 10. The mean of all samples scores then results in the mean silhouette coefficient (Unpingco, 2019). Silhouette coefficients range between -1 and 1. A high positive coefficient is desirable since it indicates that a sample has a large distance to other clusters (Unpingco, 2019). A negative score is a sign of incorrectly assigned samples and a coefficient of zero shows that a sample is in the close neighborhood to more than one cluster (Unpingco, 2019).

$$sc_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{10}$$

Figure 30 illustrates Silhouette scores for the implemented models of the present dataset. The plot reveals that the coefficients of the models are positive and relatively low in the range of 0,05 to 0,15. Consequently, the clusters in all models are relatively close to each other whereas the model with five clusters possesses the maximum Silhouette coefficient. Accordingly, by means of the Silhouette score the application of a model with five clusters is suggested as well.

After the results of both methods recommend a five-cluster solution the clustering model is applied on the data. As a result, the algorithm outputs a list of cluster labels for each observation. The list is then merged with the indices of the shuffled DataFrame. To restore the original order the shuffling is reversed by reordering the data by the country-year column in ascending order. Afterwards, the dataset is split and reshaped so that each column represents a country and each row a year (see Figure 33). The resulting DataFrame then contains the country-paths for each member state except for Malta. As a final measure, the DataFrame is exported to a .csv file and a heatmap is laid over the data to highlight the transition of the cluster assignment.

Figure 30: Line Plot of the Silhouette Scores



Source: Own work.

3.4 Results and Discussion

In addition to the country paths, several other statistics and graphs are exported from the results to get a better understanding of the data. Within this chapter all meaningful results of the analysis are discussed and linked to the objectives of this thesis to draw relevant conclusions in a final step.

Figure 31 illustrates the general statistics of the energy production for each energy carrier by including historical and forecasted data. As the table indicates the number of values amounts to 1647 for each variable now. This provides two main insights on the data, namely that there are no missing values and that the 61 observations of Malta were successfully excluded from the output.

The statistics clearly show that there exists at least one observation per energy carrier with a share of 0%. Additionally, oil is the only energy carrier with at least on observation accounting to 100%. Without any further analysis it can be observed already at this point that the dataset's maximum share in the renewable energy production accounts to 98,64%.

Regarding the target of the EU's 2050 Energy Roadmap to increase the share of renewable energy in gross final energy consumption to 100%, it can be assumed that the goal will not be reached by any country. This conclusion can be drawn from the observation that the indicator is generally lower than the actual production share. If the maximum value of the indicator is queried this assumption can be confirmed with a result of 70,1%. While also the remaining energy carriers, have a maximum share of at least 76%, the only exception is the energy production with waste whose maximum value accounts for only 3,57% with an average of 0,67% of all observations. Considering the average values, the renewable energy is the highest with 32.45% and a standard deviation of 23,9, followed by solid energy with a mean of 22,43% and a standard deviation of 24,16. Also gas and nuclear energy have a relatively high average production share with 19,85% and 18,54%, respectively. While oil is the only energy carrier that has at least one year with a maximum share of 100% the average share is relatively low with 6,06% and a high variation of 17,96.

Figure 31: General Statistics of the Energy Carriers for Historical and Forecasted Data

Energy Carrier	Count	Mean	Std.	Min.	25%	50%	75%	Max.
Solid	1647	22,43	24,16	0,00	0,91	16,07	33,66	95,63
Oil	1647	6,06	17,96	0,00	0,00	0,48	2,61	100,00
Gas	1647	19,85	17,84	0,00	4,89	14,47	32,10	76,27
Nuclear	1647	18,54	22,38	0,00	0,00	2,74	35,79	86,89
Renewables	1647	32,45	23,90	0,00	14,14	27,55	47,52	98,64
Waste	1647	0,67	0,85	0,00	0,00	0,29	1,07	3,57

Source: Own work.

When analyzing the results of the clustering process, at the beginning a closer look must be taken at the composition of the five clusters (see Figure 32). Since the EM algorithm does not sort the clusters labels following any rule, they are reorder by their average share of renewable energy production in descending order in the following step. This implies that a lower cluster number likely leads to a higher quality energy mix. Additionally, a heatmap highlights the order of the shares within each cluster. On first examination, the table shows that the average shares of waste energy are in general very low, but present in each cluster as it was underlined by the previous statistics already. The remaining five energy carriers conversely have a very high average share in each one of the clusters.

Correspondingly, the Renewable Energy cluster (0) has an average of 59,544% in the renewable energy production sector, which is the highest share among all clusters. In addition, gas with approx. 23% and nuclear energy with 15,5% complete this cluster composition along with a minor share of solid energy carriers.

The Renewable & Gas Energy cluster (1) has the second highest average share of a renewable energy production (39,632%). Also, the share of the energy production with gas is very high in this cluster with 35,405%. While the nuclear energy production just accounts for 2,662% and the waste production for 0,674%, the share of solid energy reaches a value of 17,606%. A share of 4,021% of the production with oil as an energy carrier completes the cluster.

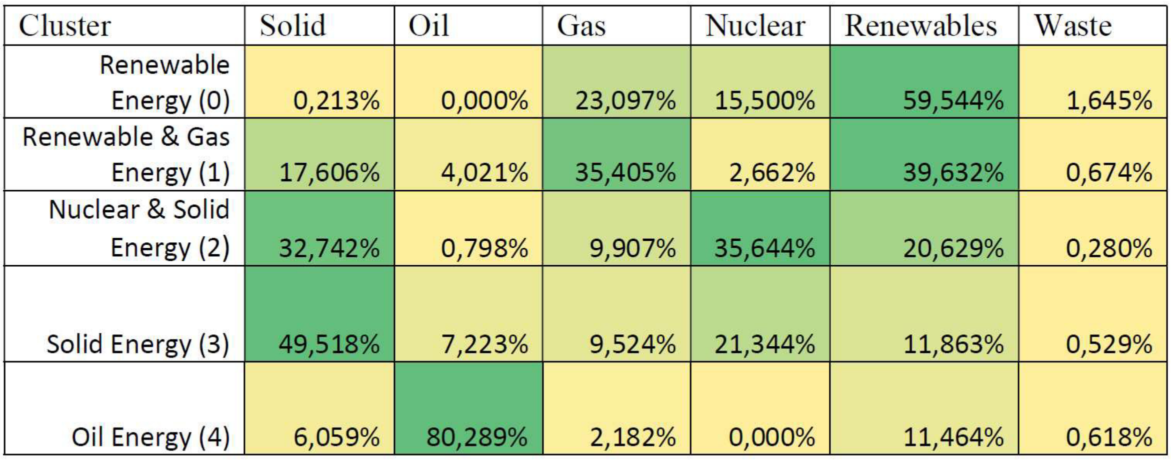
The highest average share of a nuclear energy production (35,644%) can be found in the Nuclear & Solid Energy cluster (2). Together with a high share of solid energy (32,742%) these energy carriers cover more than 2/3 of the production. The cluster is completed by renewable energy with slightly more than 20,6%, energy from gas with 9,907% as well as oil and waste energy which have shares below 1%.

The Solid Energy cluster (3) is probably the most balanced cluster as all energy carriers except for waste are present with high shares. Its significant characteristic is moreover the very strong solid energy production with 49,518% on average. Nuclear energy represents the second strongest energy carrier (21,344%). Oil (7,223%), gas (9,524%) and renewable energy (11,863%) contribute to almost all the remaining energy production. Just a small share of waste energy production (0,529%) supplements this cluster.

In the Oil Energy cluster (4) oil is by far the dominant energy carrier with a very high average share of 80,289%. The cluster is completed by relatively low average shares of 11,464% in renewable, 6,059% in solid, 2,182% in gas and 0,618% in waste energy production.

A noteworthy remark regarding the cluster compositions is that the Renewable Energy and Renewable & Gas Energy cluster, which have the highest shares of renewable energy, have the highest shares in gas as well. As stated in the literature review, the renewable energy production needs alternatives that are flexible in their energy production. Table 1 illustrated that apart from hydro and geothermal energy, the only non-renewable energy carriers with a high flexibility are oil and gas. This indicates that a renewable energy production in combination with gas as a flexible alternative is a popular approach for country-years assigned to both clusters.

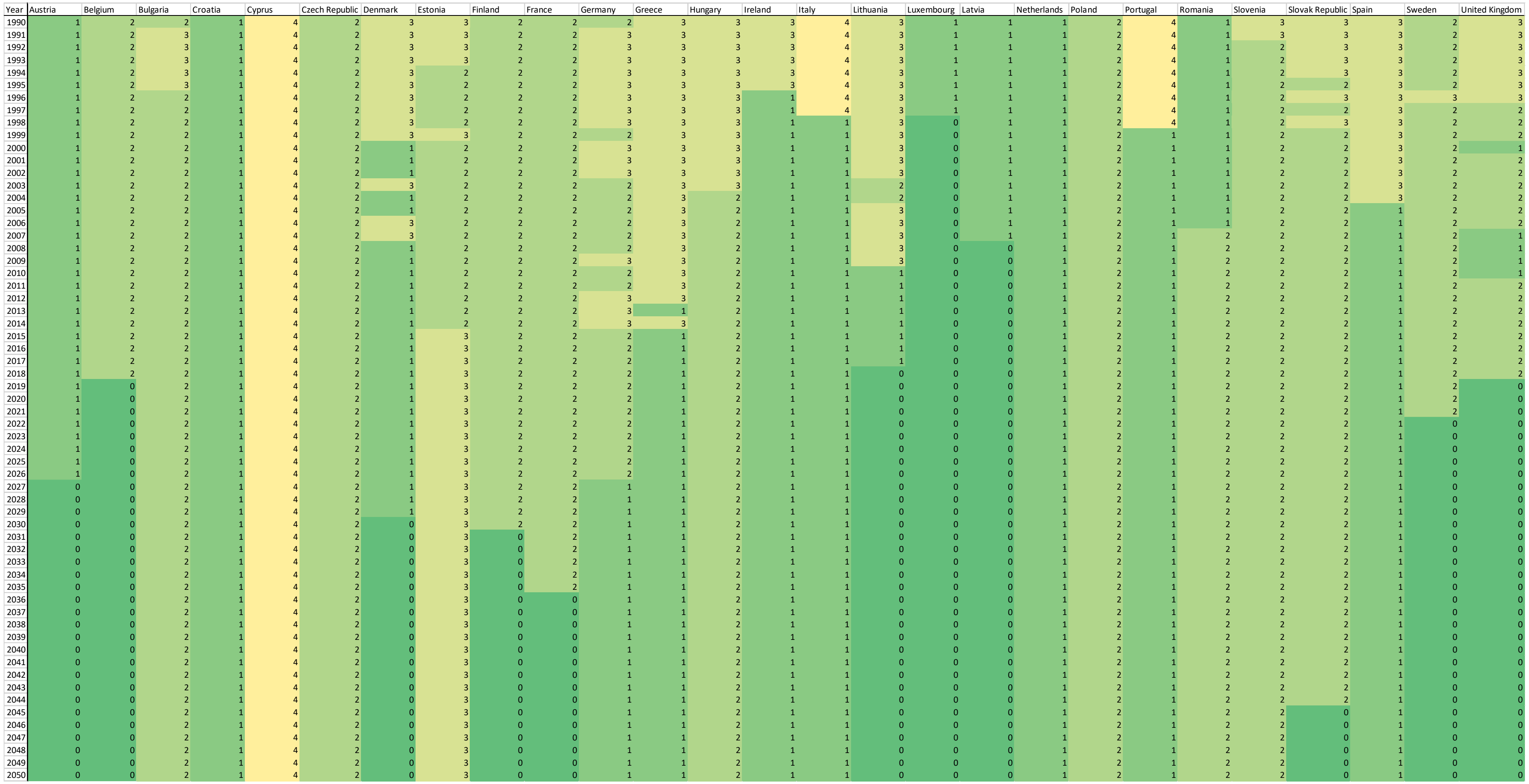
Figure 32: Cluster Compositions (Mean Values)



Cluster	Solid	Oil	Gas	Nuclear	Renewables	Waste
Renewable Energy (0)	0,213%	0,000%	23,097%	15,500%	59,544%	1,645%
Renewable & Gas Energy (1)	17,606%	4,021%	35,405%	2,662%	39,632%	0,674%
Nuclear & Solid Energy (2)	32,742%	0,798%	9,907%	35,644%	20,629%	0,280%
Solid Energy (3)	49,518%	7,223%	9,524%	21,344%	11,863%	0,529%
Oil Energy (4)	6,059%	80,289%	2,182%	0,000%	11,464%	0,618%

Source: Own work.

Figure 33: Energy Paths of the EU-28 (excl. Malta)



Year	Austria	Belgium	Bulgaria	Croatia	Cyprus	Czech Republic	Denmark	Estonia	Finland	France	Germany	Greece	Hungary	Ireland	Italy	Lithuania	Luxembourg	Latvia	Netherlands	Poland	Portugal	Romania	Slovenia	Slovak Republic	Spain	Sweden	United Kingdom
1990	1	2	2	1	4	2	3	3	2	2	2	3	3	3	4	3	1	1	1	2	4	1	3	3	3	2	3
1991	1	2	3	1	4	2	3	3	2	2	3	3	3	3	4	3	1	1	1	2	4	1	3	3	3	2	3
1992	1	2	3	1	4	2	3	3	2	2	3	3	3	3	4	3	1	1	1	2	4	1	2	3	3	2	3
1993	1	2	3	1	4	2	3	3	2	2	3	3	3	3	4	3	1	1	1	2	4	1	2	3	3	2	3
1994	1	2	3	1	4	2	3	2	2	2	3	3	3	3	4	3	1	1	1	2	4	1	2	3	3	2	3
1995	1	2	3	1	4	2	3	2	2	2	3	3	3	3	4	3	1	1	1	2	4	1	2	2	3	2	3
1996	1	2	2	1	4	2	3	2	2	2	3	3	3	1	4	3	1	1	1	2	4	1	2	3	3	3	3
1997	1	2	2	1	4	2	3	2	2	2	3	3	3	1	4	3	1	1	1	2	4	1	2	2	3	2	2
1998	1	2	2	1	4	2	3	2	2	2	3	3	3	1	1	3	0	1	1	2	4	1	2	3	3	2	2
1999	1	2	2	1	4	2	3	3	2	2	2	3	3	1	1	3	0	1	1	2	1	1	2	2	3	2	2
2000	1	2	2	1	4	2	1	2	2	2	3	3	3	1	1	3	0	1	1	2	1	1	2	2	3	2	1
2001	1	2	2	1	4	2	1	2	2	2	3	3	3	1	1	3	0	1	1	2	1	1	2	2	3	2	2
2002	1	2	2	1	4	2	1	2	2	2	3	3	3	1	1	3	0	1	1	2	1	1	2	2	3	2	2
2003	1	2	2	1	4	2	3	2	2	2	2	3	3	1	1	2	0	1	1	2	1	1	2	2	3	2	2
2004	1	2	2	1	4	2	1	2	2	2	2	3	2	1	1	2	0	1	1	2	1	1	2	2	3	2	2
2005	1	2	2	1	4	2	1	2	2	2	2	3	2	1	1	3	0	1	1	2	1	1	2	2	1	2	2
2006	1	2	2	1	4	2	3	2	2	2	2	3	2	1	1	3	0	1	1	2	1	1	2	2	1	2	2
2007	1	2	2	1	4	2	3	2	2	2	2	3	2	1	1	3	0	1	1	2	1	2	2	2	1	2	1
2008	1	2	2	1	4	2	1	2	2	2	2	3	2	1	1	3	0	0	1	2	1	2	2	2	1	2	1
2009	1	2	2	1	4	2	1	2	2	2	3	3	2	1	1	3	0	0	1	2	1	2	2	2	1	2	1
2010	1	2	2	1	4	2	1	2	2	2	2	3	2	1	1	1	0	0	1	2	1	2	2	2	1	2	1
2011	1	2	2	1	4	2	1	2	2	2	2	3	2	1	1	1	0	0	1	2	1	2	2	2	1	2	2
2012	1	2	2	1	4	2	1	2	2	2	3	3	2	1	1	1	0	0	1	2	1	2	2	2	1	2	2
2013	1	2	2	1	4	2	1	2	2	2	3	1	2	1	1	1	0	0	1	2	1	2	2	2	1	2	2
2014	1	2	2	1	4	2	1	2	2	2	3	3	2	1	1	1	0	0	1	2	1	2	2	2	1	2	2
2015	1	2	2	1	4	2	1	3	2	2	2	1	2	1	1	1	0	0	1	2	1	2	2	2	1	2	2
2016	1	2	2	1	4	2	1	3	2	2	2	1	2	1	1	1	0	0	1	2	1	2	2	2	1	2	2
2017	1	2	2	1	4	2	1	3	2	2	2	1	2	1	1	1	0	0	1	2	1	2	2	2	1	2	2
2018	1	2	2	1	4	2	1	3	2	2	2	1	2	1	1	0	0	0	1	2	1	2	2	2	1	2	2
2019	1	0	2	1	4	2	1	3	2	2	2	1	2	1	1	0	0	0	1	2	1	2	2	2	1	2	0
2020	1	0	2	1	4	2	1	3	2	2	2	1	2	1	1	0	0	0	1	2	1	2	2	2	1	2	0
2021	1	0	2	1	4	2	1	3	2	2	2	1	2	1	1	0	0	0	1	2	1	2	2	2	1	2	0
2022	1	0	2	1	4	2	1	3	2	2	2	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2023	1	0	2	1	4	2	1	3	2	2	2	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2024	1	0	2	1	4	2	1	3	2	2	2	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2025	1	0	2	1	4	2	1	3	2	2	2	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2026	1	0	2	1	4	2	1	3	2	2	2	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2027	0	0	2	1	4	2	1	3	2	2	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2028	0	0	2	1	4	2	1	3	2	2	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2029	0	0	2	1	4	2	1	3	2	2	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2030	0	0	2	1	4	2	0	3	2	2	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2031	0	0	2	1	4	2	0	3	0	2	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2032	0	0	2	1	4	2	0	3	0	2	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2033	0	0	2	1	4	2	0	3	0	2	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2034	0	0	2	1	4	2	0	3	0	2	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2035	0	0	2	1	4	2	0	3	0	2	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2036	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2037	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2038	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2039	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2040	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2041	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2042	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2043	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2044	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	2	1	0	0
2045	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	0	1	0	0
2046	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	0	1	0	0
2047	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	0	1	0	0
2048	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	0	1	0	0
2049	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	0	1	0	0
2050	0	0	2	1	4	2	0	3	0	0	1	1	2	1	1	0	0	0	1	2	1	2	2	0	1	0	0

Source: Own work.

By generating the country paths, a simple overview on the development of the energy production mixes is provided and a closer look on specific countries can be easily achieved. This ease of use is additionally supported by applying the ordered cluster labels and adding a heatmap to the country paths.

When a first look is taken at the country paths in Figure 33 a general pattern that can be observed is a low level of variation in the cluster assignment of each country. Generally, each country follows a clear path towards a specific cluster by being a member of not more than three different clusters. Without any further analysis, the application of the heatmap allows to state that the energy production mix of the EU is more and more developing towards a renewable energy production.

However, when the country paths are examined more closely, additional insights and patterns can be discovered. For instance, only four countries, namely Cyprus, the Czech Republic, Poland and the Netherlands, remain in the same cluster for the whole observation period. Furthermore, out of the 27 member states only eleven manage to change their energy mix in such a way that they become a member of the Renewable Energy cluster (0) at some point in time. In particular, Austria, Belgium, Denmark, Finland, France, Lithuania, Luxembourg, Latvia, the Slovak Republic, Sweden and the United Kingdom are part of this cluster with a high-quality energy mix.

Additionally, six countries can be identified with a potential high share of renewables in their energy production mix already in the first observation year 1990, as they are members of the Renewable & Gas Energy cluster (1). Namely these are Austria, Croatia, Luxembourg, Latvia, the Netherlands, and Romania. By being part of the cluster, the countries had already at the beginning of the records either a high share of gas, renewables, or both in their energy production mix. Since the cluster is composed of two main energy carriers each country should be individually investigated.

The energy production mix of Austria and Croatia was mainly built on hydro energy systems. Unlike Austria, Croatia had additionally relatively high shares in gas and oil in the first years. Luxembourg as well had a high share in hydro and renewables. The remaining energy was produced by almost only gas. Furthermore, most of the energy supply in Luxembourg is based on imported energy as the data shows that the small central European country has very high energy imports. The same conclusion is applicable to Latvia which is mainly depended on imports of energy. The country's energy mix is especially based on a biofuel and hydro energy production. The Netherlands is the only country with gas as its main energy carrier and only a small share of renewable energy production.

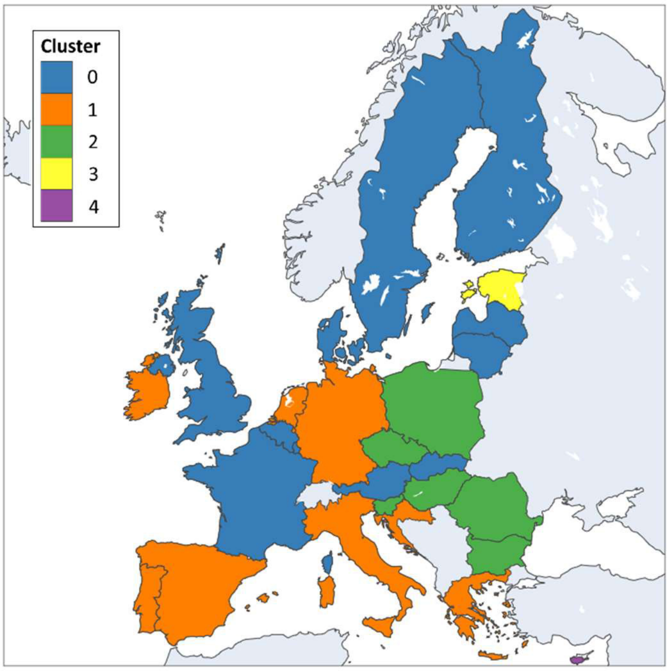
Unlike all other countries, Romania has a rather diverse energy production mix with high shares of gas, renewables, and solid fuels as well. However, the eastern European country is, together with Estonia, one of two countries that has a negative development in the long-term in its cluster membership. After 2006, Romania's cluster assignment changes from the

Renewable & Gas Energy cluster (1) to the Nuclear & Solid Energy cluster (2) where it remains until the end of the observations. This change can be explained by a transition towards an energy mix based on nuclear, solid, and renewable energy. Estonia instead is part of the Solid Energy cluster (3) in the first four years until it moves to the Nuclear & Solid Energy cluster (2) afterwards. In 1999, the country changes back to the Solid Energy cluster (3) for one year, just to return to the Nuclear & Solid Energy cluster (2) afterwards. However, this condition is not permanent and the country switches permanently back to the Solid Energy cluster (3) in 2015 with a high share of solid energy carriers in the energy mix.

For certain, the Oil Energy cluster (4) represents a special energy production mix as just three countries are assigned to it in total. While Italy and Portugal are a member of this cluster only during their first 8 and 9 observation years, Cyprus remains in cluster 4 for the whole time. This can be explained by continuous high shares of oil in its energy production mix. Partially, the data shows that these shares even account for 100%. Italy and Portugal have consistently a more diverse energy production mix by combining oil with other energy carriers. After the first years Italy and Portugal shift their energy production mix away from oil which leads to an assignment to cluster 1 afterwards, in which they remain until 2050.

If the cluster distribution is analyzed by geographical location, further major insights can be obtained. This is visualized by choropleth maps in a 10-year frequency from 1990 until 2050. The maps for the decades up to 2040 are illustrated in Appendix 5. The long-term cluster assignment for 2050 is visualized in Figure 34 below. The choropleth maps support the drawing of conclusions on regional dependencies in the energy mixes and aids result interpretation. Whereas in some years, specific regions have a similar cluster assignment, e.g. the Iberian Peninsula or eastern European countries, it cannot be generalized that the cluster membership of the countries in these regions stays the same over time. In the first three decades the cluster membership among the EU-28 is very diverse and countries tend to change the clusters more often. However, also the choropleth maps clearly illustrate the tendency towards energy mixes with a higher share of renewables. It is notable that especially many Eastern European countries remain in the Nuclear & Solid Energy cluster (2) from 2010 to 2050. This includes Poland, Czech Republic, Slovenia, Hungary, Romania and Bulgaria. The Slovak Republic shifts the cluster membership from the Nuclear & Solid Energy (2) to the Renewable & Gas Energy cluster (1) within the last analyzed decade. With the exception of Cyprus (Oil Energy cluster (4)) and Estonia (Solid Energy cluster (3)), the remaining European member states become a part of the Renewable & Gas Energy (1) or Renewable Energy cluster (0) in the long term.

Figure 34: Choropleth Map of Cluster Assignment in 2050

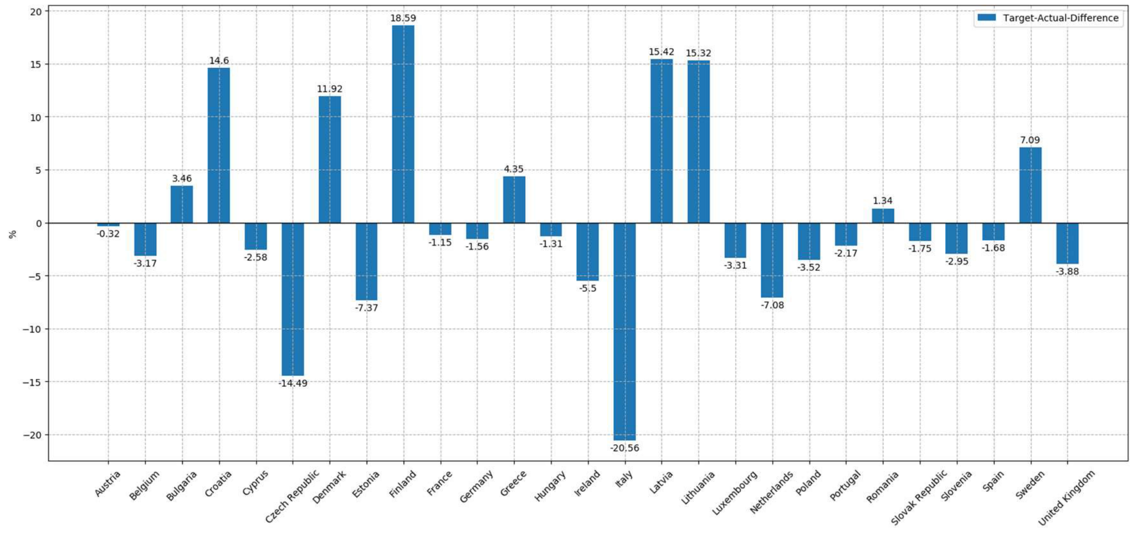


Source: Own work.

It must moreover be noted that in some cases an energy mix of a country or region could be intuitively identified. However, it is not a reliable approach since exceptions always exist and a data-driven approach allows for a better identification of patterns. Especially the choropleth maps highlight that there are many changes in the energy production mixes over time on a geographical scale, which would probably not be instinctively considered. Instead, clustering the data provides a straightforward and efficient method to identify similarities and other patterns easily among the countries and European regions.

To analyze the EU targets that address the energy transition towards renewable energy carriers, these targets need to be compared to the shares of renewable energy in gross final energy consumption. For the short-term targets in 2020 a detailed analysis can be carried out due to country specific targets as introduced in chapter 2.1.2. When comparing the target shares defined by the EU with the forecasts of each member state a clear prognosis on a possible accomplishment of the goals can be made. The specific shares are visualized in a bar graph in Appendix 6. In this illustration the shares are grouped by country to simplify the comparison between target and forecasted value. To better evaluate the results, the target-forecasted-differences of the renewable shares are calculated by subtracting the forecasted share from the target ratio. The results are visualized as a bar chart in Figure 35. A positive value indicates that the target value will be exceeded whereas a negative value shows that the target will not be met. The value of zero implies that the target value is met exactly. Correspondingly, the illustration shows that nine countries, namely Bulgaria, Croatia, Denmark, Finland, Greece, Latvia, Lithuania, Romania and Sweden, will meet their target in 2020. Unlike Bulgaria, Greece and Romania whose share is slightly higher than the target value, the remaining countries will exceed their target significantly. While Finland will have the highest exceedance of 18,59%, Latvia, Lithuania and Croatia will have a share that is approximately 15% higher than their individual target value.

Figure 35: Target-Forecasted-Difference of the Share of Renewable Energy in Gross Final Energy Consumption in 2020



Source: Own work.

In contrast, the majority of the member states can be expected not to meet the EU targets. Especially Italy (-20,56%) and the Czech Republic (-14,49%) stand out from the countries by clearly missing their targets. Belgium (-11,1 %) completes the three countries that will even have an actual share which will be more than 10% behind the targets. Still, many countries will just miss their targets with small diffences. Austria, France, Germany, Hungary, the Slovak Republic and Spain have a deficit of less than 2%. The remaining countries, that were not mentioned afore, are lacking between 2% and 10%.

As the EU also set goals on the output of GHG emissions, these also need to be controlled for their feasibility based on the forecast. In the strategies, the target emission reduction is set in comparison to the levels of 1990. Hence, the emissions in the target years are summed up and the percentual difference to the emissions sum in 1990 can be calculated. This is formalized in equation 11.

$$GHG\ emission\ reduction = -\frac{GHG\ emissions_{target\ year} - GHG\ emissions_{1990}}{GHG\ emissions_{1990}} \tag{11}$$

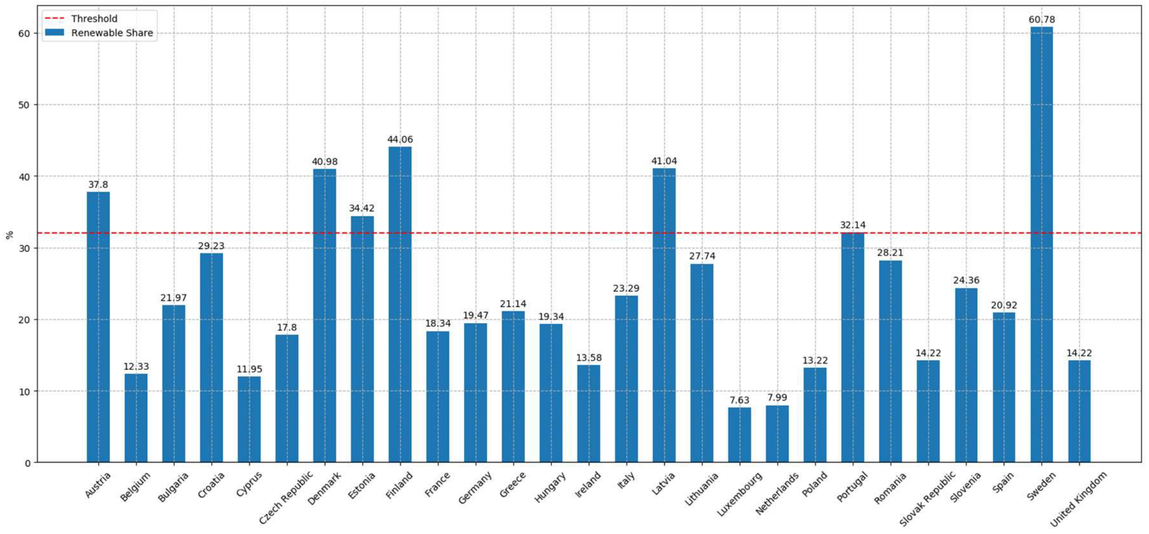
The goal in 2020 is to reduce the GHG emissions by at least 20% compared to the levels of 1990. In the target year the GHG emissions amounts to 4321,08 mio tons of which 3550,16 mio tons are CO_2 emissions. By applying the above formula, the results show that the overall reduction in the EU will account for 24,46% in 2020. This implies an achievement of the short-term objective.

Since the 2030 climate and energy framework includes medium-term targets and the 2050 Energy Roadmap long-term targets, these are specified more broadly and thus an in-depth analysis becomes more challenging. In general, the forecast over a long time period might be more uncertain due to many influences in the energy sector and to the energy mixes. However, in combination with the forecast a general prognosis on the feasibility of the EU targets can be given.

For 2030 the EU defined the target of a minimum share of renewables of 32% overall. Since each member state has a different overall energy production, a weighted average would be needed to calculate the overall indicator of the EU-28. While we forecasted the indicators for each member state, to achieve the overall share it would then be necessary to forecast all the required variables of the indicator which were defined in Section 2.1.2. Considering the fact that the data for all variables is not accessible and that higher number of variables would have made the forecast very complex and inaccurate these were not included initially.

However, to get further insights on the development of the renewable energy in gross final energy consumption up to 2030, the shares of each member state should be examined individually. Therefore, the target value of 32% is taken as a reference value. This value should be ideally reached by each country. If this can be proven to be true, it would imply that also the overall European share exceeds 32%.

Figure 36: 2030 Share of Renewable Energy in Gross Final Energy Consumption by Country



Source: Own work.

Considering the forecasted shares of renewable energy in gross final energy consumption on a national level only seven member states will exceed the target of 32%, as illustrated in Figure 36. These are namely Austria, Denmark, Estonia, Finland, Latvia, Portugal and Sweden. Sweden possesses by far the highest share with 60,78%. Together with Finland

(44,06%), Latvia (41,04%) and Denmark (40,98%) three other countries have a share higher than 40%. While Austria (37,8%) and Estonia (34,42%) will also exceed the threshold of 32% significantly and Portugal only slightly exceeds the threshold with a share of 32,14%.

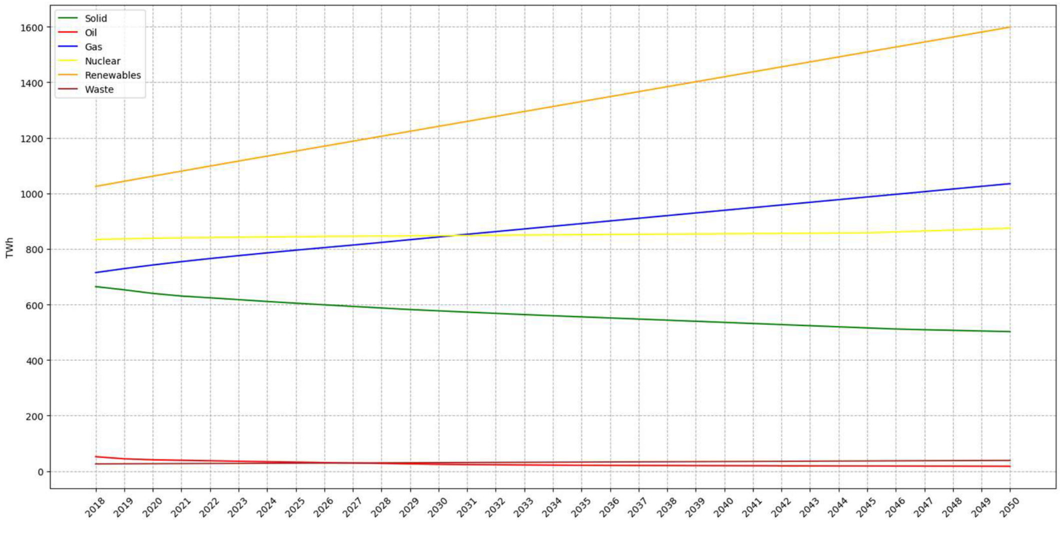
Luxembourg (7,63%) and the Netherlands (7,99%) will be the countries with the lowest shares and they are the only member states with a share below 10% in 2030. In contrast, Croatia (29,23%), Romania (28,21%) and Lithuania (27,74%) will only be lacking some percent points to reach the 32% mark. The remaining member states will all widely miss the 2030 target.

Regarding the GHG emissions, the data reveals that the objective of a 40% GHG emission reduction is slightly missed in 2030 with a percentage amounting to 35,81%. This is a total decrease of the GHG emissions to 3.671,84 and of the CO_2 emissions to 3.054,71 mio tons CO_2.

The 2050 Energy Roadmap does not specify a goal for the share of renewable energy in gross final energy consumption, instead the EU initially specified a target of 80-95% GHG emission reduction. This already ambitious target was revised and changed later to a climate neutral EU until 2050. Based on the forecast the target will clearly not be achieved. Based on the forecast results, the overall European emissions can be expected to be reduced by 56,71% in 2050. This means that the member states will still produce 2.476,56 mio tons GHG and 2.134,12 mio tons CO_2 emissions.

As the historical overall energy production mix of the EU has already been analyzed, the future development should be examined as well. The related energy production mix for the time span from 2018 to 2050 is visualized in Figure 37.

Figure 37: Sum of Forecasted Energy Production of the EU-28 (excl. Malta) by Energy Carrier



Source: Own work.

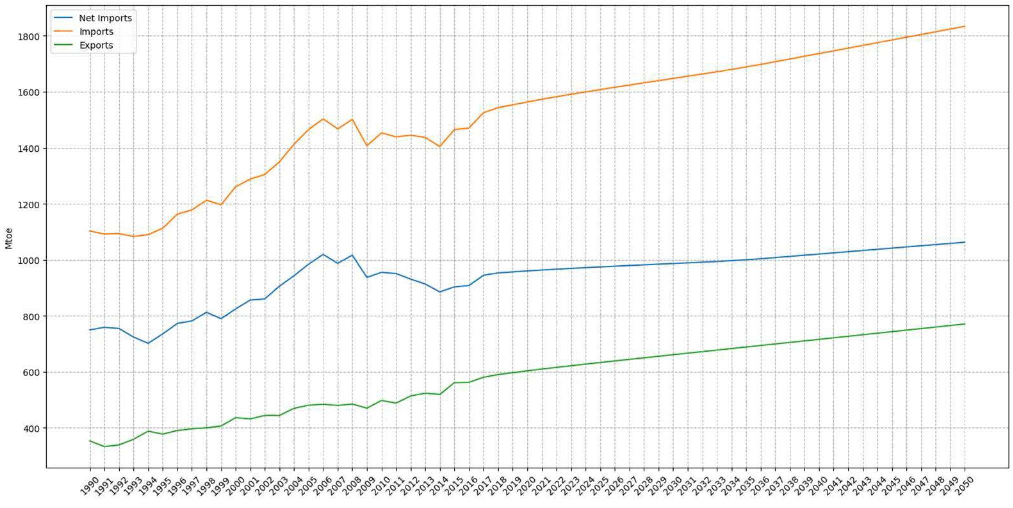
The figure underlines that also from 2018 on renewable energy carriers remain the strongest resource of the European energy production. They reach their peak in 2050 with a production of 1.598,51 TWh. The production of nuclear energy remains rather constant with a minimum growth from 834,1 TWh in 2018 to 874,7 TWh in 2050. As it has been outlined before a higher share of renewables is usually in line with a higher share of flexible energy carriers in the production and consequently it is not remarkable that the energy production from gas is also trending upwards. In 2018, the production accounted to 714,77 TWh. Between 2030 and 2031 the production already outnumbers the nuclear energy production and stays the second strongest energy carrier in the overall production mix until 2050. In 2050 the energy production will be growing to 1.034,91 TWh. It can be observed that the production with only two energy carriers is declining, namely solid and oil energy carriers. While the production of solid energy carriers, accounting for 664,3 TWh in 2018, falls to 502,22 TWh in 2050. The oil energy production, which is already relatively low, decreases even more from 51,95 TWh in 2018 to 17,42 TWh in 2050.

The energy strategies also demand a stronger unity of the EU-28 and a lower dependency on non-European energy resources. This demand is further elaborated with the specifications of the Energy Union. In this context and considering on the imports and exports of the EU a better perception can be obtained. Applying both measures, the net imports can be calculated to get further insights on the dependency on non-European countries. The net imports are formalized as:

$$Net\ Imports = \text{Imports} - \text{Exports} \tag{12}$$

Given the fact that the net imports include the overall imports and exports of the EU a positive value represents the minimum energy that is imported from non-European countries. Accordingly, a negative value stands for exceeding exports to non-European countries.

Figure 38: Imports, Exports and Net Imports of the EU-28 (excl. Malta) from 1990 - 2050



Source: Own work.

Figure 38 illustrates the development of the overall imports, exports and net imports of the EU between 1990 and 2050. This graph shows that the historical imports are increasing until year 2006 to 1.503,16 Mtoe. Afterwards the growth stagnates, and in this phase the imports even fall to a lower value of 1.404,36 Mtoe in 2014. Subsequently, the imports are increasing again until the end of the time span. The net imports of the EU show similar characteristics due to a relatively stable growth of the exports during the observation period. This is why the net imports increased from 749,42 Mtoe in 1990 to 1.062,73 in 2050. Therefore, the dependence on imported energy from non-European suppliers will probably not be reducing but rather increasing in the future. This behavior reflects e.g. the deployment of the Nord Stream 2 pipeline that ensures a better supply of Russian gas for Europe.

CONCLUSION

This thesis could prove that LSTM networks can be applied successfully to the EU energy production sector to forecast future developments. However, forecasting problems can arise if sudden changes in the energy mix appear and the model is not trained on this unknown energy mix, as the case of Malta has shown. Generally, also a larger dataset with a high granularity (like daily or monthly data) is likely to lead to improved results. It has been underlined that the energy transition is influenced by many factors of different nature, e.g. technological, political, geographical and historical challenges. In the long-term particularly, the impact of these challenges remains unknown and cannot be taken into account in this forecasting problem straightforwardly. However, this analysis by means of data mining and machine learning methods could provide a general understanding on how a possible future in the European energy production looks based on historical data and events.

The analysis highlights that based on the historical development the overall tendency in the EU goes towards a higher quality energy mix, in which a renewable energy production is dominant. From 2030 onwards this will combine with the very flexible gas as the energy carrier with the second largest share, followed by nuclear energy. As it was covered in section 2.2 also the current storage technologies are not sufficiently affordable and efficient enough to make the application of conventional energy carriers unnecessary. While it could be shown that the dependency on solid energy carriers is decreasing in the future, without a major breakthrough in the area of storage technologies, the composition of future energy mixes will still remain dependent on conventional fuels, especially on gas. Considering the geographical location of the member states, it became clear that particularly Eastern European countries tend to implement the energy transition slower than the remaining states. A notable exception from all member states is the island of Cyprus, which will continuously depend on high imports and oil as its main energy carrier.

Regarding the European energy strategies, it could be demonstrated that the specific targets set in the short-term for 2020 were not specified carefully enough. Many member states are lacking some percentage points to achieve their specific targets in the share of renewable energy in gross final energy consumption. On the contrary, some countries will exceed their targets significantly. Overall, this results in a reduction of the GHG emissions by 24,46% compared to the levels of 1990, which means that the 20% target is clearly met. Still, based on the forecast the middle- and long-term GHG emission targets will most likely not be met. Only some countries will manage to implement a higher quality energy mix in such a way that they can comply with the EU's objectives for 2030 and 2050.

The analysis was also capable of underlining that the EU will still be dependent on non-European partners to import energy in the future. This might be a result of resource endowments restrictions among the member states. This argument is supported by the controversial Nord Stream 2 project that will ensure a supply of gas from Russia in the years to come. Following the trend, the net imports of the EU will even increase overall.

Concludingly, it can be stated that the EU's strategies are very ambitious and their objectives will therefore most likely not be achieved based on the recent developments. This accounts especially for the middle- and long-term goals. While there are many challenges and influences on the European energy sector, the EU needs to overcome these challenges and significantly drive the energy transition forward to enable a successful European target implementation.

REFERENCE LIST

1. Ahlquist, J. S., & Breunig, C. (2012). Model-based Clustering and Typologies in the Social Sciences. *Political Analysis*, *20*(1), 92–112.
2. Basu, D., & Miroshnik, V. W. (2019). *The Political Economy of Nuclear Energy*. Cham: Springer International Publishing.
3. Bisong, E. (2019). Building Machine Learning and Deep Learning Models on Google Cloud Platform. Berkeley, CA: Apress.
4. Blume, S. W. (2017). *Electric Power System Basics for the Nonelectrical Professional*. Hoboken, NJ, USA: John Wiley & Sons, Inc.
5. Bontempi, G., Ben Taieb, S., & Le Borgne, Y.-A. (2013). Machine Learning Strategies for Time Series Forecasting. In W. van der Aalst, J. Mylopoulos, M. Rosemann, M. J. Shaw, C. Szyperski, M.-A. Aufaure, & E. Zimányi (Eds.), *Lecture Notes in Business Information Processing. Business Intelligence* (Vol. 138, pp. 62–77). Berlin, Heidelberg: Springer Berlin Heidelberg.
6. Bouzerdoum, M., Mellit, A., & Massi Pavan, A. (2013). A hybrid model (SARIMA–SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant. *Solar Energy*, *98*, 226–235.
7. Braff, W. A., Mueller, J. M., & Trancik, J. E. (2016). Value of storage technologies for wind and solar energy. *Nature Climate Change*, *6*(10), 964–969.
8. Buchholz, B. M., & Styczynski, Z. (2014). *Smart Grids – Fundamentals and Technologies in Electricity Networks*. Berlin, Heidelberg: Springer Berlin Heidelberg.

9. Csereklyei, Z., Thurner, P. W., Langer, J., & Küchenhoff, H. (2017). Energy paths in the European Union: A model-based clustering approach. *Energy Economics*, *65*, 442–457.
10. Dannecker, L. (2015). The European Electricity Market: A Market Study. In L. Dannecker (Ed.), *Energy Time Series Forecasting* (pp. 11–47). Wiesbaden: Springer Fachmedien Wiesbaden.
11. Debnath, K. B., & Mourshed, M. (2018). Forecasting methods in energy planning models. *Renewable and Sustainable Energy Reviews*, *88*, 297–325.
12. Dileep, G. (2020). A survey on smart grid technologies and applications. *Renewable Energy*, *146*, 2589–2625.
13. Dorsman, A., Westerman, W., Karan, M. B., & Arslan, Ö. (Eds.). (2011). *Financial Aspects in Energy*. Berlin, Heidelberg: Springer Berlin Heidelberg.
14. Đozić, D. J., & Gvozdenac Urošević, B. d. (2019). Application of artificial neural networks for testing long-term energy policy targets. *Energy*, *174*, 488–496.
15. Dreyfus, G. (2005). *Neural Networks*. Berlin/Heidelberg: Springer-Verlag.
16. Dutta, G., & Mitra, K. (2017). A literature review on dynamic pricing of electricity. *Journal of the Operational Research Society*, *68*(10), 1131–1145.
17. The Economist (2018, August 7). *Why Nord Stream 2 is the world's most controversial energy project.* Retrieved April 11, 2019, from https://www.economist.com/the-economist-explains/2018/08/07/why-nord-stream-2-is-the-worlds-most-controversial-energy-project.
18. The Economist (2019, May 14). *What are the school climate strikes?* Retrieved April 11, 2019, from https://www.economist.com/the-economist-explains/2019/03/14/what-are-the-school-climate-strikes.
19. Eid, C., Koliou, E., Valles, M., Reneses, J., & Hakvoort, R. (2016). Time-based pricing and electricity demand response: Existing barriers and next steps. *Utilities Policy*, *40*, 15–25.
20. Erbach, G. (2016). *Understanding electricity markets in the EU.* Retrieved July 25, 2019, from http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/593519/EPRS_BRI(2016)593519_EN.pdf.
21. European Commission (2007). *Blackout of November 2006: important lessons to be drawn.* Retrieved August 05, 2019, from https://europa.eu/rapid/press-release_IP-07-110_en.htm?locale=en.
22. European Commission (2010). *Energy 2020:* A strategy for competitive, sustainable and secure energy. Retrieved March 05, 2020, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52010DC0639&from=EN.
23. European Commission (2011). *Energy Roadmap 2050.* Retrieved March 05, 2020, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52011DC0885&from=EN.
24. European Commission (2014). *Smart grids and meters.* Retrieved October 31, 2019, from https://ec.europa.eu/energy/en/topics/markets-and-consumers/smart-grids-and-meters/overview#content-heading-4.

25. European Commission (2015). *A Framework Strategy for a Resilient Energy Union with a Forward-Looking Climate Change Policy.* Retrieved March 05, 2020, from https://eur-lex.europa.eu/resource.html?uri=cellar:1bd46c90-bdd4-11e4-bbe1-01aa75ed71a1.0001.03/DOC_1&format=PDF.
26. European Commission (2019). *Calculation methodologies for the share of renewables in energy consumption.* Retrieved July 09, 2019, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Calculation_methodologies_for_the_share_of_renewables_in_energy_consumption.
27. European Council (2014). *European Council (23 and 24 October 2014) – Conclusions.* Retrieved March 05, 2020, from https://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ec/145397.pdf.
28. European Council (2019). *European Council meeting (12 December 2019) – Conclusions.* Retrieved March 05, 2020, from https://www.consilium.europa.eu/media/41768/12-euco-final-conclusions-en.pdf.
29. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37–54.
30. Funabashi, Y., & Kitazawa, K. (2012). Fukushima in review: A complex disaster, a disastrous response. *Bulletin of the Atomic Scientists*, *68*(2), 9–21.
31. Guivarch, C., & Monjon, S. (2017). Identifying the main uncertainty drivers of energy security in a low-carbon world: The case of Europe. *Energy Economics*, *64*, 530–541.
32. Gyo Lee, Y., Garza-Gomez, X., & Lee, R. (2018). Ultimate Costs of the Disaster: Seven Years After the Deepwater Horizon Oil Spill. *Journal of Corporate Accounting & Finance*, *29*(1), 69–79.
33. Haas, R., Mez, L., & Ajanovic, A. (2019). *The Technological and Economic Future of Nuclear Power*. Wiesbaden: Springer Fachmedien Wiesbaden.
34. Hake, J.-F., Fischer, W., Venghaus, S., & Weckenbrock, C. (2015). The German Energiewende – History and status quo. *Energy*, *92*, 532–546.
35. IPCC (2014). *Climate Change 2014 Synthesis Report:* Summary for Policymakers. Retrieved March 05, 2020, from https://www.ipcc.ch/site/assets/uploads/2018/05/SYR_AR5_FINAL_full_wcover.pdf.
36. Kabir, G., & Hasin, M. A. A. (2013). Comparative Analysis of Artificial Neural Networks and Neuro-Fuzzy Models for Multicriteria Demand Forecasting. *International Journal of Fuzzy System Applications*, *3*(1), 1–24.
37. Kirschen, D. S., Strbac, G., Cumperayot, P., & Paiva Mendes, D. de. (2000). Factoring the elasticity of demand in electricity prices. *IEEE Transactions on Power Systems*, *15*(2), 612–617.
38. Lavorini, V. (2018). *Gaussian Mixture Model clustering: how to select the number of components (clusters).* Retrieved November 28, 2019, from https://towardsdatascience.com/gaussian-mixture-model-clusterization-how-to-select-the-number-of-components-clusters-553bef45f6e4.

39. Lee, T., Markowitz, E. M., Howe, P. d., Ko, C.-Y., & Leiserowitz, A. A. (2015). Predictors of public climate change awareness and risk perception around the world. *Nature Climate Change*, *5*(11), 1014–1020.
40. Li, C., Sun, Y., & Chen, X. (2007). Analysis of the blackout in Europe on November 4, 2006. In *International Power Engineering Conference, 2007: IPEC 2007 ; Singapore, 3 - 6 Dec. 2007* (pp. 939–944). Piscataway: IEEE Service Center.
41. Li, S. (2017). *Time Series Analysis, Visualization & Forecasting with LSTM:* Statistics normality test, Dickey–Fuller test for stationarity, Long short-term memory, from https://towardsdatascience.com/time-series-analysis-visualization-forecasting-with-lstm-77a905180eba.
42. Livingstone, D. J. (2009). *Artificial Neural Networks: Methods and Applications*. *SpringerLink Bücher: Vol. 458*. Totowa, NJ: Humana Press. Retrieved from http://dx.doi.org/10.1007/978-1-60327-101-1
43. Mata Pérez, M. d. L. E., Scholten, D., & Smith Stegen, K. (2019). The multi-speed energy transition in Europe: Opportunities and challenges for EU energy security. *Energy Strategy Reviews*, *26*, 100415.
44. Mehedintu, A., Sterpu, M., & Soava, G. (2018). Estimation and Forecasts for the Share of Renewable Energy Consumption in Final Energy Consumption by 2020 in the European Union. *Sustainability*, *10*(5), 1515.
45. Mellit, Adel, & Pavan, A. M. (2010). Performance prediction of 20kW$_P$ grid-connected photovoltaic plant at Trieste (Italy) using artificial neural network. *Energy Conversion and Management*, *51*(12), 2431–2441.
46. Mourshed, M., Robert, S., Ranalli, A., Messervey, T., Reforgiato, D., Contreau, R., . . . Lennard, Z. (2015). Smart Grid Futures: Perspectives on the Integration of Energy and ICT Services. *Energy Procedia*, *75*, 1132–1137.
47. Muntean, M., Guizzardi, D., Schaaf, E., Crippa, M., Solazzo, E., Olivier, J.G.J., & Vignati, E. (2018). *Fossil CO2 emissions of all world countries - 2018 Report*. Luxembourg.
48. Mwasilu, F., Justo, J. J., Kim, E.-K., Do, T. D., & Jung, J.-W. (2014). Electric vehicles and smart grid interaction: A review on vehicle to grid and renewable energy sources integration. *Renewable and Sustainable Energy Reviews*, *34*, 501–516.
49. Nikoletatos, J., & Tselepis, S. (2015). *Renewable Energy Integration in Power Grids:* Technology Brief. Retrieved June 21, 2019, from https://europeanpowertogas.com/wp-content/uploads/2018/05/Ngg1uITu.pdf.
50. Official Journal of the European Union (2009). *DIRECTIVE 2009/28/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL.* Retrieved November 29, 2019, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32009L0028&from=EN.
51. Official Journal of the European Union (2018a). *DIRECTIVE (EU) 2018/2001 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL.* Retrieved February 05, 2020, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018L2001&from=de.

52. Official Journal of the European Union (2018b). *DIRECTIVE (EU) 2018/2002 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL.* Retrieved February 05, 2020, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018L2002&from=EN.
53. Rebala, G., Ravi, A., & Churiwala, S. (2019). *An Introduction to Machine Learning*. Cham: Springer International Publishing.
54. Rodat, S., Tantolin, C., Le Pivert, X., & Lespinats, S. (2016). Daily forecast of solar thermal energy production for heat storage management. *Journal of Cleaner Production*, *139*, 86–98.
55. Romero, J. (2012). Blackouts illuminate India's power problems. *IEEE Spectrum*, *49*(10), 11–12.
56. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.
57. Shobha, G., & Rangaswamy, S. (2018). Machine Learning. In Handbook of Statistics. Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications (Vol. 38, pp. 197–228). Elsevier.
58. Stram, B. N. (2016). Key challenges to expanding renewable energy. *Energy Policy*, *96*, 728–734.
59. Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. *Pearson international Edition*. Boston, Mass.: Pearson/Addison-Wesley.
60. Taulli, T. (2019). *Artificial Intelligence Basics*. Berkeley, CA: Apress.
61. Tichý, L. (2019). *EU-Russia Energy Relations*. Cham: Springer International Publishing.
62. United Nations (n.d.). *Paris Agreement - Status of Ratification.* Retrieved May 20, 2019, from https://unfccc.int/process/the-paris-agreement/status-of-ratification.
63. United Nations (2015). *Paris Agreement,* from https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement.
64. Unpingco, J. (2019). *Python for Probability, Statistics, and Machine Learning*. Cham: Springer International Publishing.
65. Vivoda, V., & Graetz, G. (2015). Nuclear Policy and Regulation in Japan after Fukushima: Navigating the Crisis. *Journal of Contemporary Asia*, *45*(3), 490–509.
66. Wasilewski, J., & Baczynski, D. (2017). Short-term electric energy production forecasting at wind power plants in pareto-optimality context. *Renewable and Sustainable Energy Reviews*, *69*, 177–187.
67. Wu, Y.-K., Chang, S. M., & Hu, Y.-L. (2017). Literature Review of Power System Blackouts. *Energy Procedia*, *141*, 428–431.
68. Yablokov, A. V., Nesterenko, V., & Nesterenko, A. v. (Eds.). (2009). *Annals of the New York Academy of Sciences: Vol. 1181. Chernobyl: Consequences of the catastrophe for people and the environment*. Boston, Mass.: Wiley Blackwell.

APPENDICES

Appendix 1: Summary in Slovene language

Podnebne spremembe so ena največjih groženj za zemeljski ekosistem, ki jih v glavnem lahko pripišemo posledicam človeških dejanj. K podnebnim spremembam še posebej veliko prispeva energetski sektor. Magistrska naloga se v povezavi s tem na naboru zgodovinskih podatkov proizvodnje energije ukvarja z analizo kombinacij virov energije posameznih držav (angl. energy mix) v EU. Za ta namen so uporabljene metode podatkovnega rudarjenja in strojnega učenja. Cilj naloge je torej podati odgovor na vprašanje, ali je mogoče pri kombinacijah virov energije opaziti določene vzorce ter ali je mogoče opaziti tendenco v premiku proti določenim kombinacijam virov energije oziroma samim virom energije (angl. energy carrier). Eden izmed ciljev zaključnega dela je tudi primerjava ugotovitev analize in energetskih programov EU z namenom podati sklep o njihovi uresničljivosti. Programi z načrtovanimi roki za izvedbo v letih 2020, 2030 in 2050 se osredotočajo na rast proizvodnje energije iz obnovljivih virov, povečevanje energetske učinkovitosti in zmanjševanje izpustov pri proizvodnji energije. EU je specifične cilje programov sestavila na tak način, da so merljivi, kar omogoča neposredno primerjavo z izsledki analize v tej magistrski nalogi.

Za analizo je bilo z namenom sledenja strukturiranemu pristopu uporabljeno ogrodje, ki temelji na procesu odkrivanja znanja v podatkovnih bazah (angl. Knowledge Discovery from Data - KDD) ter procesu strojnega učenja (angl. Machine Learning) in globokega učenja (angl. Deep Learning), ki ga je v letu 2019 predstavil Taulli. Zgodovinske vrednosti podatkov, ki so uporabljene za napovedovanje in primerjane s ciljnimi metrikami v programih EU, so predhodno izbrane iz obrazca energetskih statističnih podatkov po državah EU, ki pokriva obdobje od leta 1990 do leta 2017. Napoved do leta 2050 je na podlagi zgodovinskih podatkov narejena s pomočjo nevronskih mrež z dolgim kratkoročnim spominom (angl. Long Short-Term Memory - LSTM). Pridobljeni podatki analize so za nadaljnjo podrobnejšo analizo združeni s pomočjo pristopa razvrščanja v skupine na podlagi modela (angl. Model-Based Clustering), ki omogoča napovedovanje kombinacij virov energije tudi za prihodnost.

Magistrska naloga ugotavlja, da bo na splošno v EU tendenca verjetno premik proti kombinaciji virov energije višje kakovosti s prevlado proizvodnje energije iz obnovljivih virov. Obnovljive vire energije bo od leta 2030 naprej najpogosteje spremljal zemeljski plin, ki velja za zelo prilagodljiv vir energije. Jedrska energija bo ostala tretji najpogostejši vir energije. Med primerjavo rezultatov analize s ciljnimi vrednostmi energetskih programov lahko opazimo, da kratkoročnim ciljem države ne posvečajo veliko pozornosti, saj jih glede na napovedi večina individualnih ciljev bodisi ne bo dosegla bodisi jih bodo presegle. Srednjeročne in dolgoročne vrednosti programov se poleg tega lahko izkažejo za še posebej preveč ambiciozno zastavljene oziroma nedosegljive. Nedavni potek dogodkov v evropskem energetskem sektorju in napovedi, ki iz tega poteka izhajajo, z gotovostjo kažejo na to, da večina držav članic ne bo zmožna doseči vrednosti, ki so bile zanje postavljene v programih EU.

Appendix 2: EU-28 Share of Renewables in 2020 by Country

Table A 1: EU Target Shares of Gross Final Energy Consumption by Country 2020

Country	Target for share of energy from renewable sources in gross final consumption of energy, 2020 (S_{2020})
Austria	34%
Belgium	13%
Bulgaria	16%
Croatia	13%
Cyprus	13%
Czech Republic	30%
Denmark	25%
Estonia	38%
Finland	23%
France	18%
Germany	18%
Greece	13%
Hungary	16%
Ireland	17%
Italy	40%
Latvia	23%
Lithuania	11%
Luxembourg	10%
Malta	10%
Netherlands	14%
Poland	15%
Portugal	31%
Romania	24%
Slovak Republic	14%
Slovenia	25%
Spain	20%
Sweden	49%
United Kingdom	15%

Source: Adapted from Official Journal of the European Union (2009).

Appendix 3: Forecasting Model Characteristics

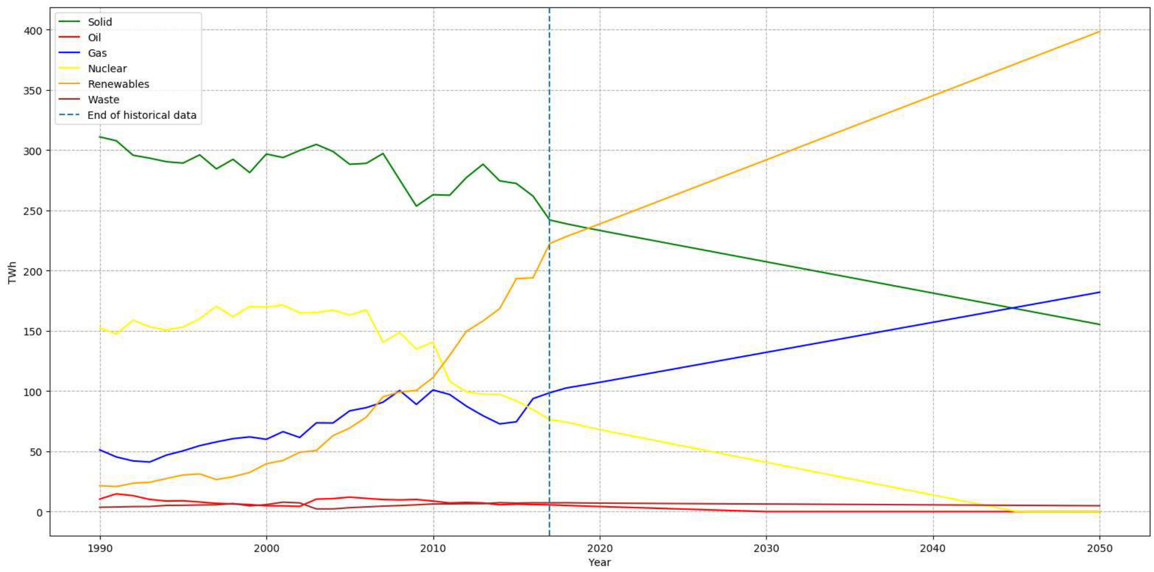
Table A 2: Parameter Configurations and MSEs of each Member State

Country Code	Epochs	Neurons Layer 1	Neurons Layer 2	MSE Training	MSE Testing	Validation Split
AT	275	19	10	0,0466	0,0484	0,15
BE	275	19	10	0,0489	0,0653	0,15
BG	275	19	10	0,0369	0,049	0,2
CY	225	22	15	0,0235	0,0681	0,17
CZ	275	19	10	0,0311	0,0504	0,2
DE	275	19	10	0,0472	0,0756	0,2
DK	350	19	10	0,0368	0,0546	0,2
EE	300	20	10	0,031	0,0398	0,15
EL	225	20	14	0,0403	0,0649	0,15
ES	250	19	10	0,0498	0,0694	0,17
FI	275	20	12	0,0497	0,0616	0,17
FR	275	20	11	0,0402	0,0564	0,17
HR	275	20	10	0,0282	0,0421	0,2
HU	275	19	10	0,0392	0,0455	0,2
IE	300	20	10	0,0394	0,0958	0,2
IT	180	20	12	0,0461	0,071	0,17
LT	300	16	9	0,0319	0,0534	0,2
LU	250	19	10	0,0434	0,0545	0,2
LV	300	17	9	0,0297	0,0327	0,2
MT	350	25	20	0,0166	0,1453	0,2
NL	300	26	14	0,0353	0,0802	0,2
PL	350	16	8	0,0365	0,0767	0,15
PT	275	20	10	0,0438	0,0731	0,15
RO	300	17	10	0,0307	0,0366	0,3
SE	250	19	10	0,0453	0,0705	0,2
SI	275	20	10	0,0327	0,071	0,2
SK	275	20	9	0,0325	0,0471	0,15
UK	200	13	8	0,0559	0,0816	0,07

Source: Own work.

Appendix 4: Overall Development of Energy Carrier Production

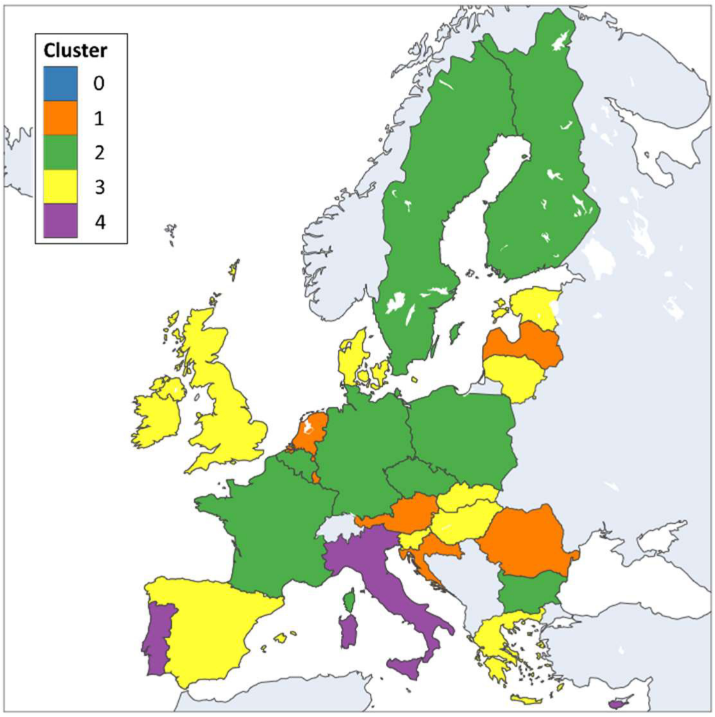
Figure A 1: Line Plot of historical and forecasted overall Energy Production by Energy Carrier



Source: Own work.

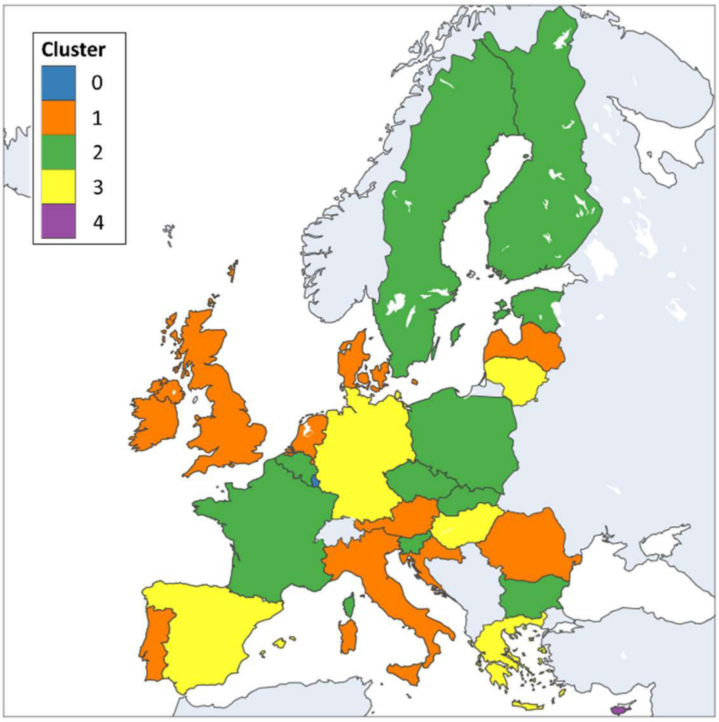
Appendix 5: Choropleth Maps of Cluster Assignment (1990-2040)

Figure A 2: Cluster Assignment 1990



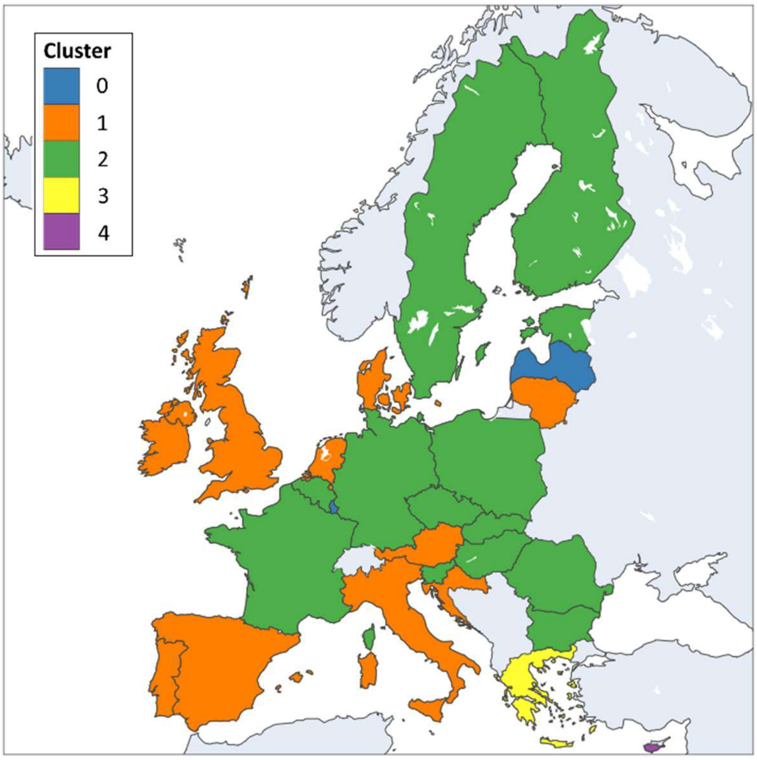
Source: Own work.

Figure A 3: Cluster Assignment 2000



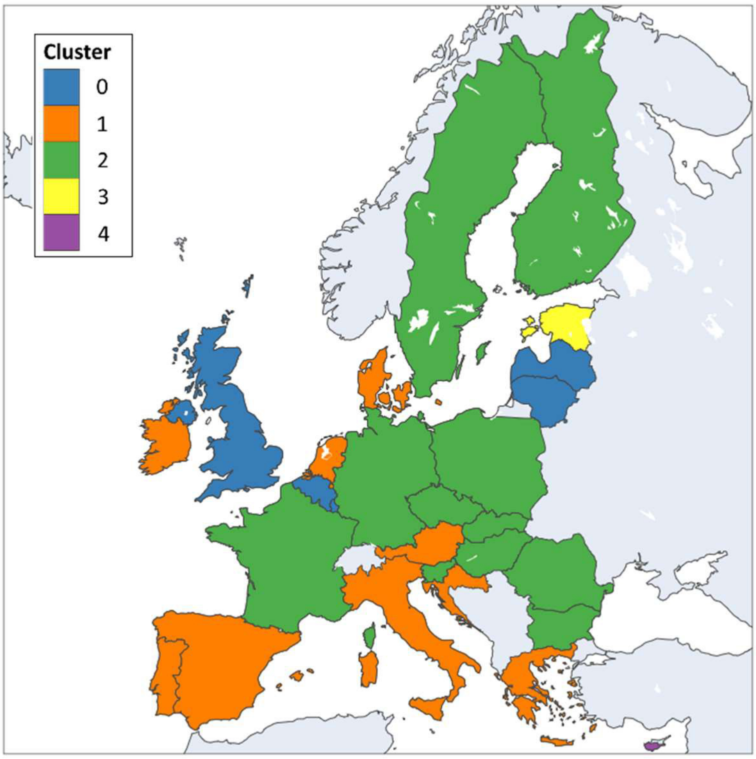
Source: Own work.

Figure A 4: Cluster Assignment 2010



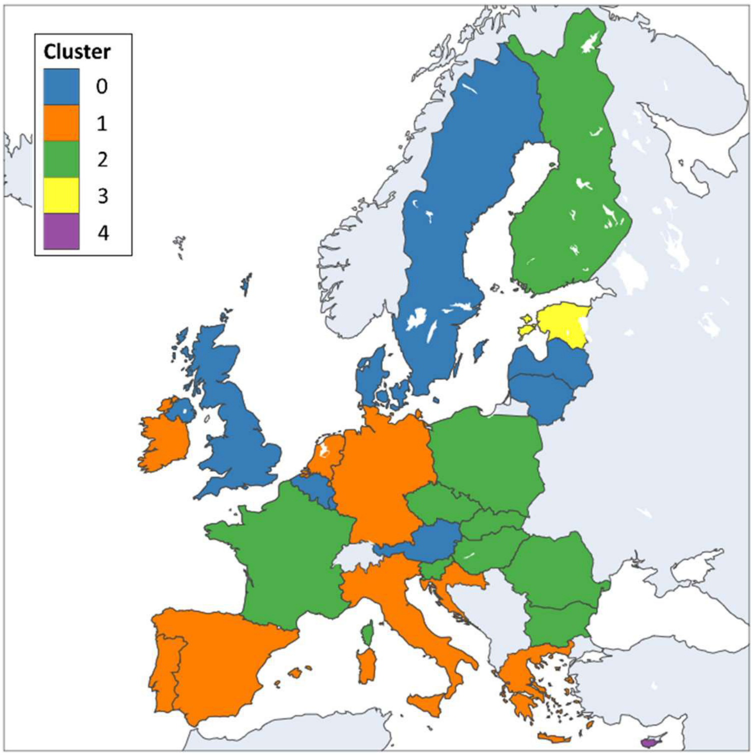
Source: Own work.

Figure A 5: Cluster Assignment 2020



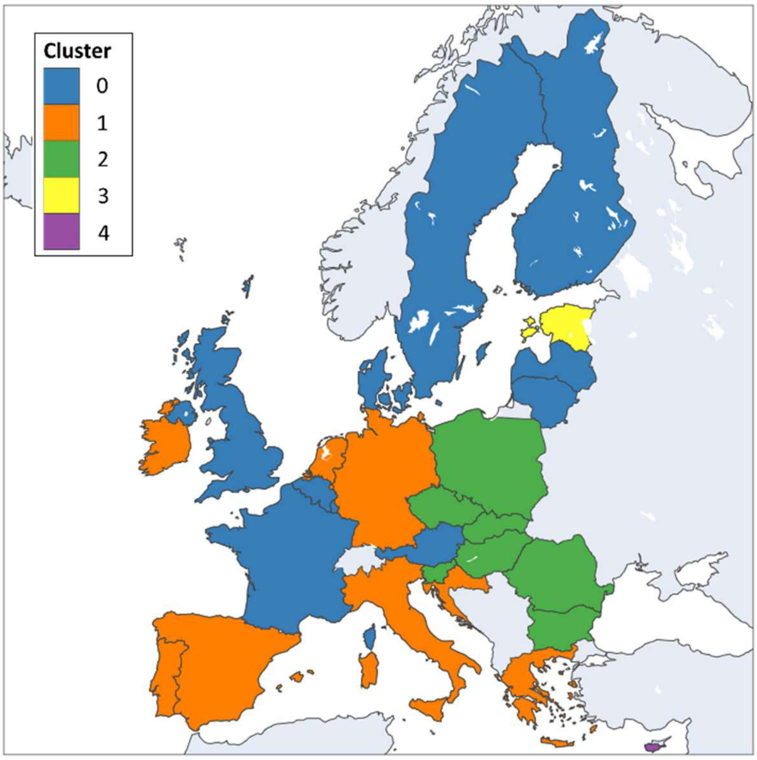
Source: Own work.

Figure A 6: Cluster Assignment 2030



Source: Own work.

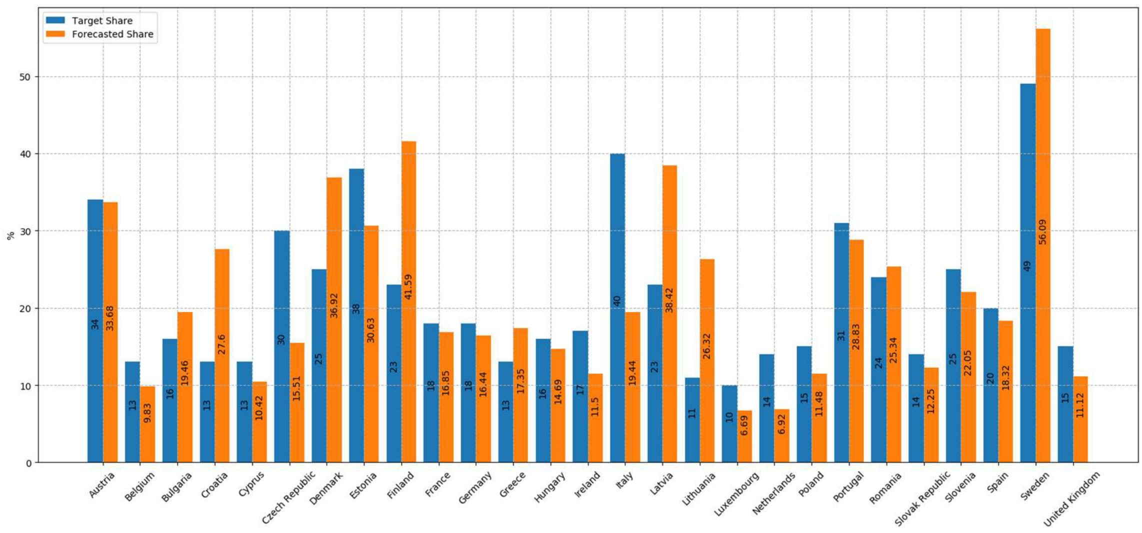
Figure A 7: Cluster Assignment 2040



Source: Own work.

Appendix 6: Target and Forecasted Share of Renewables in 2020

Figure A 8: Target and Forecasted Share of Renewable Energy in Gross Final Energy Consumption in 2020



Source: Own work.