Electricity Bills

Analysis of the Differences in Price for European households

Francesco Cabras

Supervisor: Prof. Luigi Amedeo Bianchi
Co-Supervisor: Prof. Claudio Agostinelli

Department of Mathematics
University of Trento
Academic Year 2019/2020

Contents

Chapter 1

Introduction

Electricity bills are among the most important expenses for families around the world, at least for the lucky households that have access to it. According to the *Energy Information Administration*, "Electricity prices generally reflect the cost to build, finance, maintain, and operate power plants and the electricity grid [28]." The peculiarity of this good is in the fact it is produced in the moment it is consumed. In fact, electricity storage is still quite expensive, so providers usually decide to sell at a low price rather than store [10]. Thus, short-term prices are impacted by the time in consideration, day or night, morning or afternoon, but also by climate conditions: demand due to heating in the winter and cooling in the summer are the main drivers for seasonal price spikes. This is even more relevant when not only the demand but also the supply is influenced by the weather. One of the obstacles in the broad spread of renewable sources is precisely its reliance on external conditions, e.g. sun, wind or tide.

The idea of this study is to propose machine learning approaches to *price forecasting* in the electricity market. The aim however is not to fit the monthly or even daily fluctuations in the price, but the longer-term swings, typically taking place at year level. For this reason, the focus is going to be on the demand rather than on the supply, and indicators used have to do with the economic situation, the electricity market structure and the sources for energy generation.

The electricity market is in evolution: in recent years many countries around the world have gone through liberalization of their markets, with national monopolies in the electricity sector being broken down. The result is that consumers can choose among different suppliers, and may decide to modify the profile of their demand to reduce their costs [16]. This means that the demand of electricity has become more elastic. Another trend has been the increasing investment in renewable sources of energy. To weigh the effect of such changes on the good price is the main purpose of this document. Previous works in this field have traditionally focused on measuring the impact of specific variables, while the focal point here is not merely the interpretation of the effects but especially the accuracy of the final prediction. The aim is to build

models that not only allow to understand whether the choice of using a certain source of energy rather another is significantly related with a decrease in the final price, but also to use that information to call for a prediction that is as precise as possible. The goodness of the algorithms proposed is assessed by looking at error on the validation set, since some data is left for this purpose.

Traditionally, in econometric studies on electric price fluctuations, the number of factors being studied and the number of countries considered is rather limited. This is widely justifiable: the narrower the scope of the problem, the greater the confidence in the final result. This study, on the other hand, starts from different assumptions, with the purpose shifting to the identification of those economic, market and energy choice indicators that allow for accurate price predictions both in Europe and worldwide. The models proposed are unlikely to be optimal, i.e. the most accurate, for specific countries or even small geographical areas, yet understanding the exact nature of those local dependencies is not really the aim of this thesis. The geographical scope, in fact, is the whole European continent, although results are also benchmarked using data from a broader set of countries, all over the world.

The analysis starts with a presentation of the factors at stake for long-term price prediction. With this aim, *Chapter 2* summarizes country-wide differences in consumption patterns, energy generation choices and economic development. Once it is clear why differences in those indicators are expected to explain gaps in the electricity price between countries, in *Chapter 3* statistical models are proposed in order to test those expectations. Finally, *Chapter 4* provide a comparison of performances among models, also exposing the areas this analysis may be improved.

Before proceeding, credits are to be given to the developers of *R*, the software environment used for obtaining statistically valid results, *Tableau*, the platform used for elaborating the visualizations and *Github*, for the storage and versioning of the code. Also, this report was drafted with LaTeX.

Chapter 2

Elements at Stake

Figure 2.1 provides a global picture of the amounts spent by households around the world for their electricity bills. Countries are colored on the basis of their price level, measured in US dollars per kWh.

Some countries are not colored either because their price is too low, as the case is for Kuwait, with less than one cent per kWh, or because it is too high, as for Solomon Islands and Venezuela. This latter country in recent years has been suffering from exceptional inflation: the political and macro-economical frameworks are extremely important in this sense.

Looking at the pale color of countries like South Sudan or Somalia, which are among the poorest in the world, the reader may think that low electricity prices are consistently experienced in low income countries, assuming no sky-rocketing inflation is encountered. However, there are counterexamples. Consider for example Algeria and Niger, two bordering countries: the former has an electricity price that is one tenth the price of the latter (2.1 cents in Algeria, 21.3 cents in Niger) but the *GDP* per capita of the former is 10 times the *GDP* per capita of the latter (414 dollars in Niger, 4,115 in Algeria). Since inflation is not really a problem in Niger, this discrepancy suggests that the link between economic development and electricity price is not straightforward. In Niger, for example, the problem lies in the extremely low electrification rate: only one in seven Nigeriens has access to modern electricity services, and just four percent of rural residents have access through the national provider. The country's total electrical energy consumption per capita is 44 kWh. It is the second lowest energy consumption per capita in West Africa, behind Guinea Bissau, and the ninth lowest in the world.

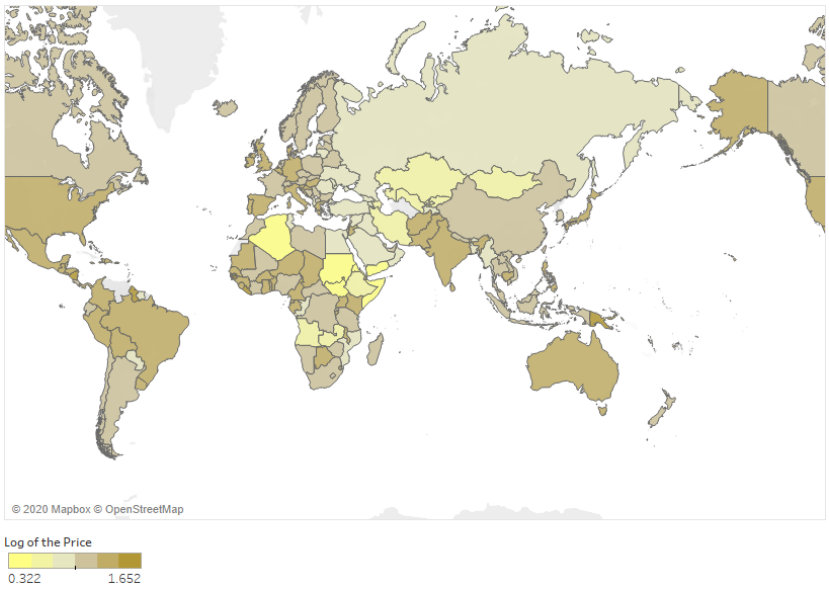

Figure 2.1: Price of Electricity around the Globe.

2.1 Access to Electricity

While the whole population of Algeria has access to the grid, this is not the case for Niger, where only part of the relatively wealthier urban population does. This explains why the price is not related with the gross domestic product per capita: since only the wealthy population can afford to have access to the grid, the prices can be kept higher. This causes an endless loop, as network costs - one of the most relevant components in the final price - are high because shared among few people. *Figure 2.2* shows how electricity price (in dollar cents per kWh) is related to the share of population having grid access, in countries with no universal electricity service. The color is used to indicate the latitude of the country: the redder, the further south, the greener, the further north.

Although countries providing little electricity access to their population also have low gross domestic product, prices are not proportionally lower when compared to wealthier and more electricity-intensive countries. Again *Figure 2.2* shows how countries with little grid access are concentrated in Africa: in recent years, the Asian continent (India in particular) has invested heavily in its electrification grid.

2.2 Production

Access to electricity is linked to production and consumption patterns inside a country. Electricity is not readily available in nature, it must be generated from natural or artificial energy sources. Apart from a slight decrease during the 2008 financial crisis,

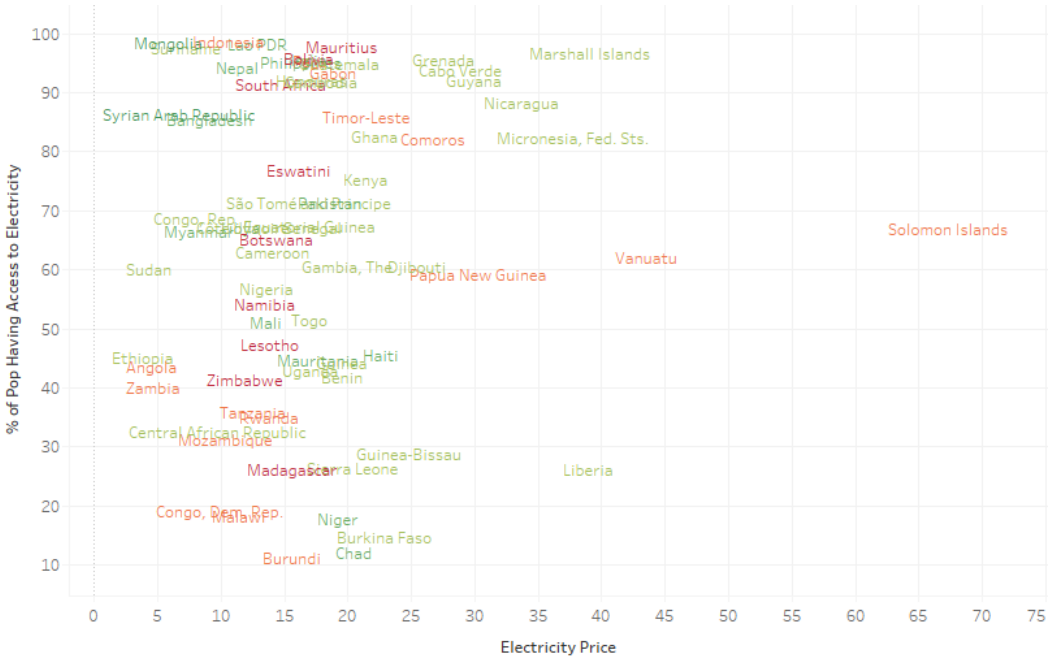

Figure 2.2: Low electrification rate does not always correspond to low consumer price.

Country	Production (2000)	Production (2019)
China	1,356	7,482
United States	4,053	4,385
India	570	1,614
Russia	878	1,122
Japan	1,068	1,013
Canada	606	649

Table 2.1: Major Electricity Producers (in KwH), Enerdata.

electricity production has been increasing constantly ever since, with China having overtaken United States as the main producer.

Table 2.1 illustrates which are the major electricity producer countries in the world, with the most recent data available. It is interesting to notice how China's production leapt forward, increasing 5 times in 20 years, by an average of 5% every year. This led China to take the US place as the world's main producers: while the US increased their generation capacity too, they could not keep the pace of the Asiatic giant. India was only the 7_{th} major producer in 2000, but was able to triple its production as part of the effort to grow its economy. Japan generation capacity, conversely, stagnated in the same time period.

Shifting the focus on European countries, one finds Germany and France to be the two major producers in the continent, followed by the United Kingdom. European situation is illustrated in *Figure 2.3*.

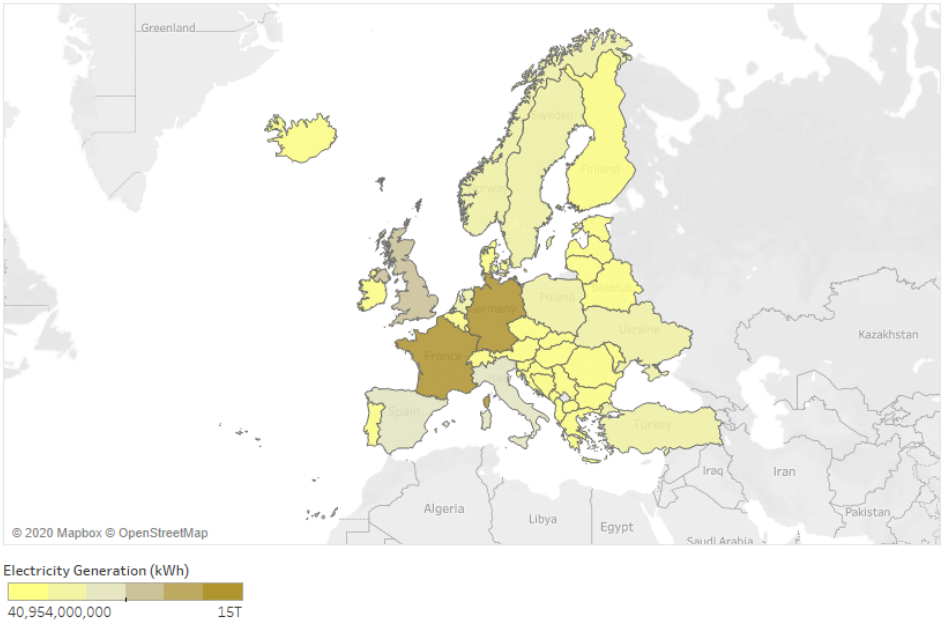

Figure 2.3: European countries and corresponding electricity generation capacity.

Country	Losses (%)
Togo	71
Libya	69.7
Haiti	60.1
Iraq	50.6
Congo, Rep.	44.5
Niger	41.8

Table 2.2: Distribution Losses (% of output), World Bank data.

2.3 Distribution

Of course producing electricity might be a problem for countries that lack the necessary resources. But an even more significant issue is often the absence of the necessary infrastructure to distribute this good. In poor regions in particular, distribution losses due to poor infrastructure represent a big concern and hinder economic development. *Table 2.2* shows which countries have the least efficient electricity distribution system.

Sadly, most of the distribution losses occur in very poor countries, where access to electricity is limited.

Country	Consumption (2000)	Consumption (2019)
China	1,138	6,510
United States	3,590	3,865
India	376	1,230
Russia	693	922
Japan	986	918
South Korea	263	553

Table 2.3: Major Electricity Producers (in KwH), Enerdata.

Country	kWh consumption per capita
Iceland	53,832
Norway	23,000
Bahrain	19,597
Kuwait	15,591
Canada	15,588
Finland	15,520

Table 2.4: Countries with highest consumption per capita values, World Bank data.

2.4 Consumption

Since the major costs associated with electricity are actually related to transportation, most of the electricity produced in one country actually remains inside that country until it is consumed. This explains why the major electricity producers are also the major electricity consumers. However, there are still marginal differences among these rankings. *Table 2.3* shows which are the principal electricity consumers in the world.

Also in this case, information regarding Asian countries turns out being the most compelling. China, India and South Korea have experienced a major boost in their electricity consumption while Japan decreased its value in the last 20 years. China is the country which consumes (and produces) the largest share of electricity because it is also the most populated. Data regarding consumption per capita is reported in *Table 2.4*.

In general the more extreme the temperatures, the more electricity is needed to heat or cool down the building and improve well-being. Consumption per capita values are particularly high for countries which are both wealthy and located in those extreme temperature areas. It is the case of Bahrain or Kuwait where massive amounts of energy are needed to cool buildings. Or, conversely, it is the case of Iceland and Norway, where temperatures are very low and during winter months sunlight is available only for a few hours. *Figure 2.4* displays how European countries with highest consumption per capita are located in the North, where temperatures are lowest.

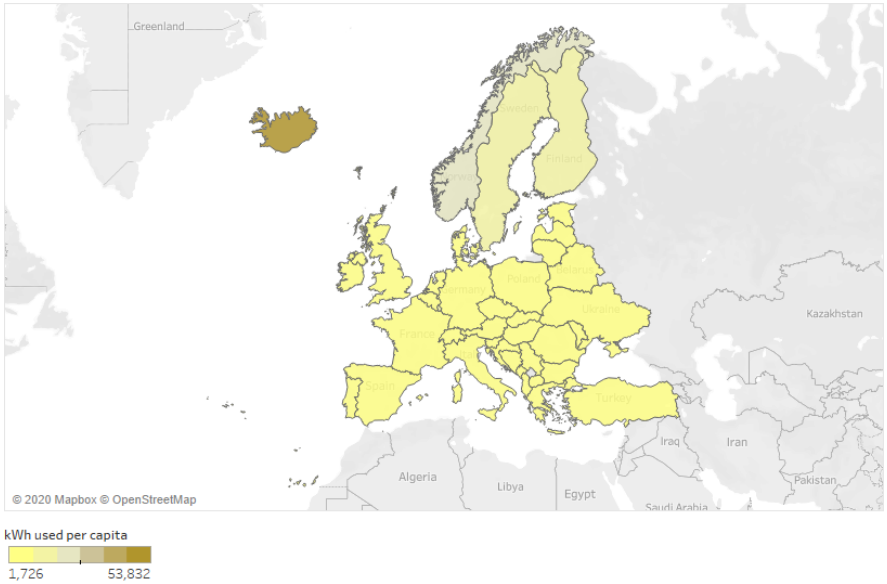

Figure 2.4: European countries and corresponding consumption per capita values, World Bank data.

2.5 European Union Electricity Market

Electricity is not an easily tradable product, as it needs legal regulations and technical standards to be agreed upon before it can become easily marketable. Electric power is, after all, nothing more than a flow of electrons inside metallic wires of a massive, interconnected network. For this reason, for decades, electricity was deemed to be an "anti-market" product, best suited to non-competitive markets like natural monopolies [11]. Furthermore, the particular characteristics of electricity, its non-storability and the necessary constant balance between demand and supply, supported the State's intervention. This monopoly predisposition resulted in efficiencies, where State subsidies have been the rule to maintain a stable industry [8]. Nevertheless, there is one region in the world that has tried to overcome these "anti-market" barriers to create an extremely vast electricity market: the *European Union*.

The EU is a unique example of a large extended market in which member countries can exchange goods. Liberalization of the economic markets and **convergence** are two of the main objectives in the Union. In particular, European electricity market liberalization represents the world's most largest-scale cross-jurisdiction reform of the electricity sector involving integration of national markets. Reaching such result required several years, for the following reasons:

- The objective of this project is to open up national monopolies' market spaces to foreigners, a radical change that inevitably leaves some parties unsatisfied. One risk of market concentration is that big incumbents try to build barriers in order to maintain their position and to limit the entrance of more efficient market

actors [23].

- In the last 25 years there has been no technological innovation disruptive enough to challenge the incumbent energy providers.
- The national arrangements that had been developed between industry players and public authorities could not be easily merged at the EU level into a common scheme.

Creating a market that is actually free, however, is not an easy task, and this explains why there are still wide differences in the price households pay for electricity in different member countries. Although the pace of the ongoing liberalization is steady, the integrated European electricity market is yet to be achieved.

2.5.1 Taxation

Taxes account for a sizable share of the final price consumers pay for energy around the EU and can have a strong impact on consumption patterns, the type of energy consumed and its use. There is disagreement among European countries on how much households should contribute in their bills: the consequence is wide differences in tax rates. Belgium, Ireland, Germany and Denmark are the member states where electricity is the most expensive. Taxes, however, are not the only explanation for the higher price: these four countries have different tax rates. In Belgium and Ireland, the price for electricity is high, even with taxes and levies excluded. In Germany, and even more in Denmark, on the other hand, taxes and levies have a huge impact on the final price. This behavior is well represented in *Figure 2.5*, where the price of the electricity service and the tax amount make up the total price paid by European consumers. Final prices are lowest in Bulgaria and Hungary, whose inhabitants pay one third of the amounts German citizens do.

The average final price paid by European households is 21.7 Euro cents per KwH including taxes and 12.8 cents per KwH excluding taxes, which sum up to 9 cents of taxes per KwH. Over time, the impact of taxes on the final bill price has grown: in 2019, the mean electricity price excluding taxes is the same as it was in 2010 but is superior by about 2-3 cents when taxes are considered. *Figure 2.6* shows how electricity prices for European citizens changed over time, on average.

2.5.2 Inflation

As all other commodities, the price of electricity tends to increase over time at a similar pace with respect to inflation. The average increase in prices in a given country is an

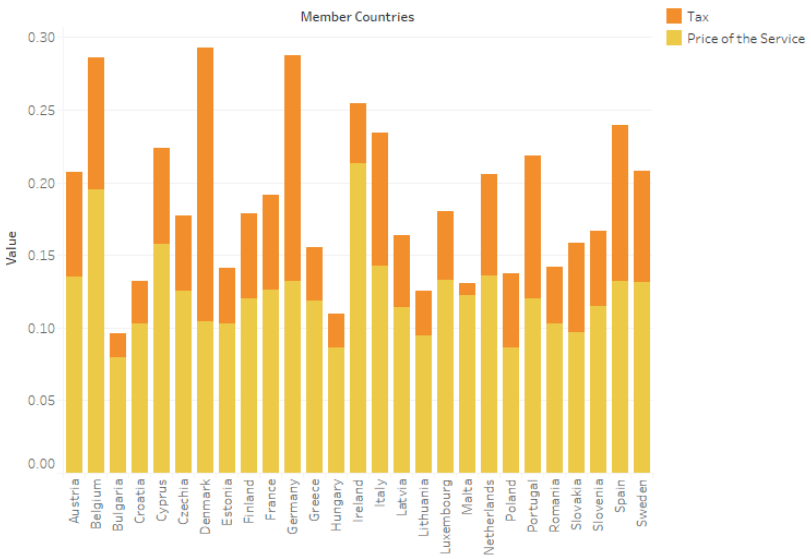

Figure 2.5: 2019 (second semester) data, electricity prices.

obvious cause for electricity price surge. In the European Union the levels of inflation are over time converging, as it should be for such a tied market, and the pace of such convergence became even steadier after the global financial crisis [5]. The body which is in charge for controlling the inflation level in the Euro area is the *European Central Bank* (ECB), and the primary objective of its monetary policy is to maintain price stability. The ECB aims at inflation rates of below, but close to, 2% over the medium term. The ECB failed to keep inflation under this level for the first years of the 21th century, before the 2008 financial crisis, but after this the inflation level dropped and never got back to such high levels. The 2% inflation target was highly debated in recent times, especially after the Federal Reserve (the United States central bank) has relaxed it.

As it is possible to discern from *Figure 2.7*, the consumer price index and the electricity price tend to show a similar trend. The average level of prices in 2015 is used as a benchmark and is denoted by a 100 in the plot. Interestingly, however, the electricity price has had its peak in 2012, while inflation continued being positive until now.

The behavior of inflation depends both on structural and on short-term factors. The inflation rate over long periods is determined by the extent to which the rate of money growth (which is controlled by the ECB or by national central banks in non-euro countries) exceeds the growth rate of real output. Short-run fluctuations are instead due to demand shocks, e.g. sizable increases in government spending, and supply shocks, e.g. sharp rises in the oil price [3].

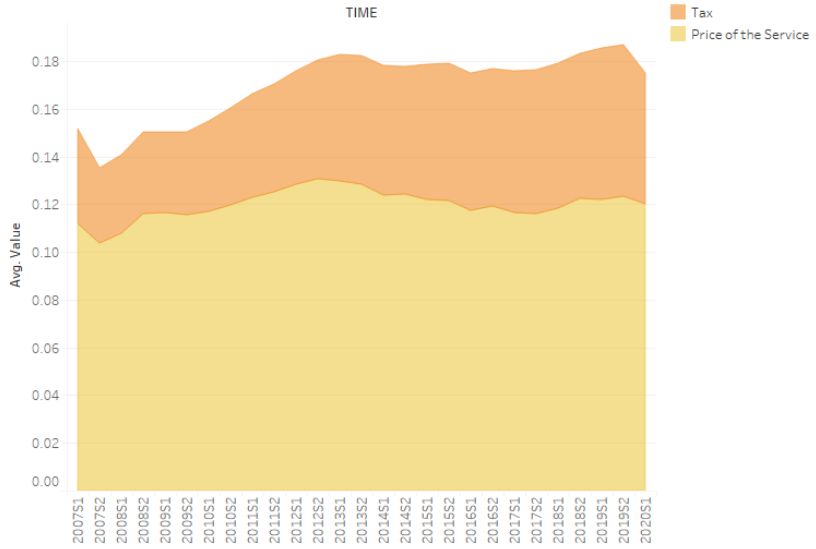

Figure 2.6: Evolution of prices over time.

2.5.3 Market Concentration

The directive 2003/54/EC requires that all non-household customers can freely choose their electricity supplier by 1 July 2004, with following full market opening including all household customers within three years. Most of the European Union electricity market is now at least formally, open to competition; this was not the case a few decades ago. However, most national electricity markets are still controlled by relatively few companies and small consumers seem quite resistant to switching supplier [13].

The European Union is fighting monopolies in the electricity sector for the simple reason that the market price is supposed to be higher with respect to a situation of perfect competition. In the 1990s the United Kingdom was the front-runner of electricity reforms, while France has often been regarded as a country averse to moving away from public monopoly [9]. In *Figure 2.8* European countries are colored according to their level of market concentration. It is possible to notice how, most notably, the market concentration levels are quite heterogeneous among countries, with France, Estonia and Croatia experiencing very different levels with respect to United Kingdom, Poland and Spain.

In any case, *Figure 2.8* also demonstrates how, although the liberalization process has led to the disintegration of national monopolies, it did not lead to a disruptive fall in concentration within the sector. With an average value of the biggest supplier market share of 50% the European situation is quite far from the internal market and open competition setting. Different national and regional markets with the presence of incumbents as main actors still persist in each electricity market. Despite liberalization, the level of concentration is hence quite high. In *Figure 2.9*, the focus shifts to the biggest European markets over the last 20 years: Germany, Spain, France and

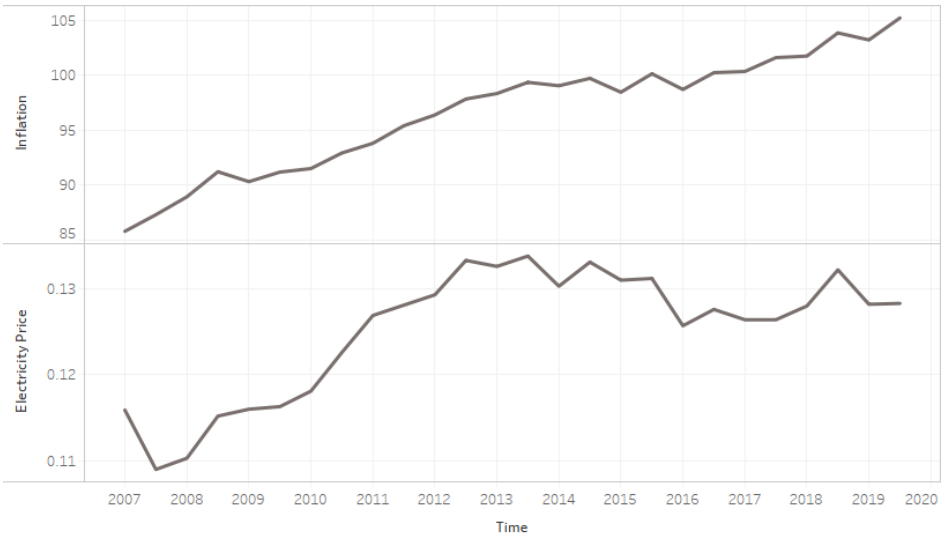

Figure 2.7: Consumer Price Index and Electricity Price in the European Union.

Italy. It is interesting to notice how they all experienced a slight decrease in market concentration, but with France starting from a value that stands out with respect to the others, with the most important supplier having a market share of 90%.

One final note of caution: in this section and later in the model discussion the market share of the biggest supplier in the market is used as a proxy for market concentration. The lower the share, the more the market is liberalized. However, looking at the largest supplier is a bit limiting and it is used because it is the only information Eurostat discloses about this matter. A market in which there are only two players sharing the same size is completely different from one in which the biggest supplier controls half of the market and the rest of it is controlled by very little providers. Yet, the largest supplier market share variable value would be the same in both settings.

2.5.4 Gross Domestic Product

The wealth and the productive capacity of a country are closely related to the consumption patterns of its citizens. Economic development stimulates demand for electricity in the long-run: the gross domestic product has an effect primarily on the quantity of electricity consumed [15]. During times of economic hardship, many factories decide to cut back production due to a reduction of consumer demand and consequently reduce electrical demand. [2] Whether there is an effect not only on the quantity used but also on the price of the good is a more debated topic: [17] employs annual data for Malaysia from 1970 to 2008 but finds that there is no causal relationship between prices and economic growth. *Figure 2.10* displays the differences in GDP per capita for European countries [1].

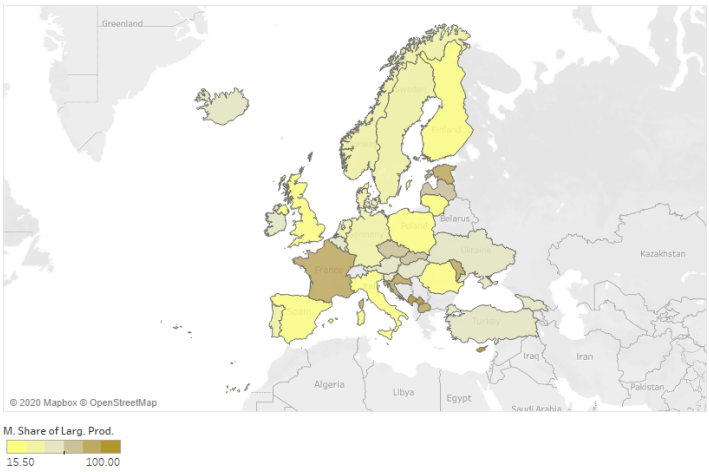

Figure 2.8: Market Share of Largest Producers for European countries, 2018 data.

Belgium, Ireland, Germany and Denmark are the European member countries for which electricity is the most expensive, and these four areas also show high GDP value. In the last 20 years Gross Domestic Product tended to increase for European countries, even if the financial crisis incurred in the end of the first decade slowed down the growth process. Also the electricity price on average grew, although at a different pace and more irregularly with respect to the Gross Domestic Product.

Figure 2.11 shows the behavior of real GDP per capita and electricity price (taxes excluded) for France and Germany, the two major economies in the European Union. In this case inflation has been taken into account and in fact the absolute GDP values are not comparable with *Figure 2.10*. Electricity is cheaper in France with respect to Germany by quite a significant amount, and Gross Domestic Product is also lower. Both variable surged over the period, but electricity price's increase has been more irregular.

Emission Intensity

Depending on how it is produced, electricity can be associated with environmental degradation. A debated question in recent times is whether cleaner energy results in higher prices or rather the opposite. The *Emissions Intensity* indicator is computed as the ratio between energy-related GHG emissions and gross inland consumption of energy. It expresses how many tonnes of CO_2 equivalents of energy-related GHGs are being emitted in a certain economy per unit of energy that is being consumed.

All European countries successfully managed to decrease their emission intensity in the last 20 years, but with wide differences. Eastern countries which base their economy on energy production and distribution such as Ukraine and Georgia saw their emission intensity levels drop by one tenth in one year. Such a result is mostly due to the

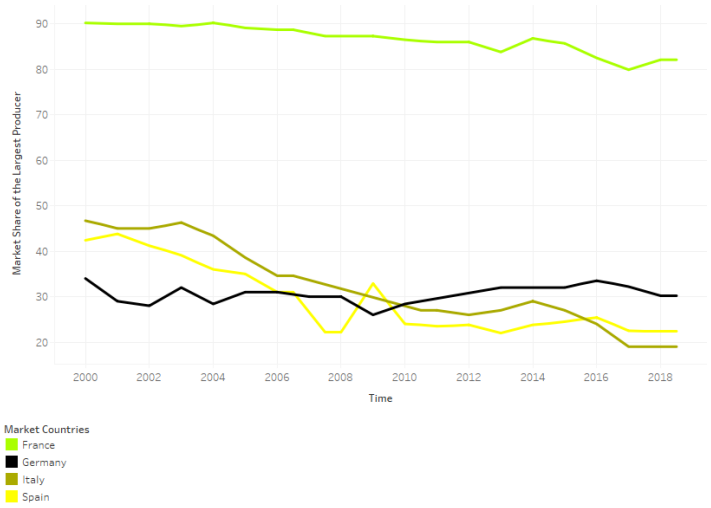

Figure 2.9: Largest supplier market share for Germany, Spain, France and Italy over time.

Country	Emission Intensity
Ukraine	90.9
Malta	90.9
Georgia	90.9
Kosovo	91.3
Moldova	91.3

Table 2.5: Countries which reduced their emission intensity the most in the last 20 years.

very low levels of efficiency these countries were experiencing in the beginning of the century. In *Table 2.5* emission intensity for most virtuous countries is summarized: the relative level in the table is benchmarked over the emission intensity level the country experienced in 2000 (i.e. level in 2000 = 100).

With *Table 2.6*, instead, a summary of the countries which experienced the smallest drop in emission intensity in the last 20 years is provided. It may not be that easy to match economic development with climate action, and progress in one field can hinder progress in the other.

Eurostat does not provide absolute values for the Emission Intensity indicator, but the World Bank provides data for CO_2 emissions per capita. *Figure 2.12* helps to visualize the countries which are the most and least CO_2 intensive. One interesting area of focus for further discussion is the Baltic. To demonstrate how political decisions can result in building two radically different economies also for countries which are neighboring and similar, it is interesting to notice how Lithuania is one of the least polluting member states in Europe and Estonia is instead the one that is polluting the most.

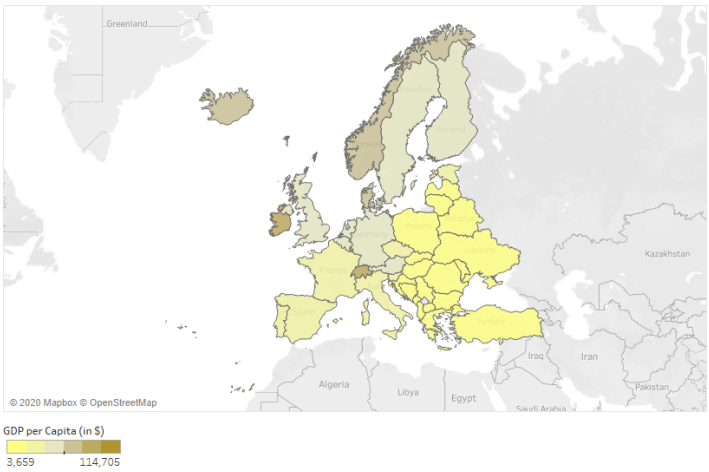

Figure 2.10: Gross Domestic Product per capita for European countries (in Dollars).

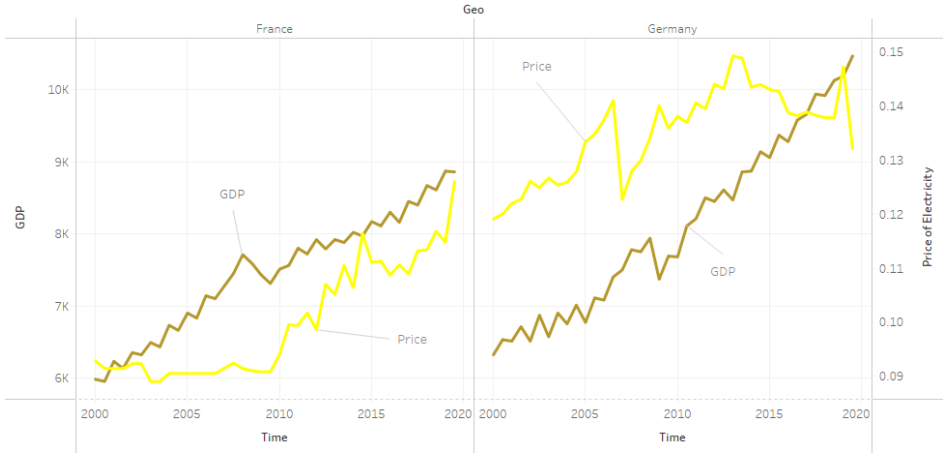

Figure 2.11: GDP per capita and Electricity Price over time for France and Germany.

Energy Mix

The detrimental effect of energy generation on the environment is reflected by the fact that 65% of emissions in the World are currently due to the use and production of energy. This percentage rises up to 80% in the European Union [19]. Energy production techniques and consequent polluting emissions vary greatly from one country or region to the other and can change significantly depending on the period. The term *energy mix* refers to the combination of the various primary energy sources used to meet energy needs in a given geographic region. Variables at stake for the final choice of the energy mix include:

- The availability of resources domestically or the possibility of importing them;
- The demand of energy which needs to be met;

Country	Emission Intensity
Bulgaria	97
Lithuania	95.6
Luxembourg	95.3
Cyprus	95.2
Estonia	95

Table 2.6: Countries which reduced their emission intensity the most in the last 20 years.

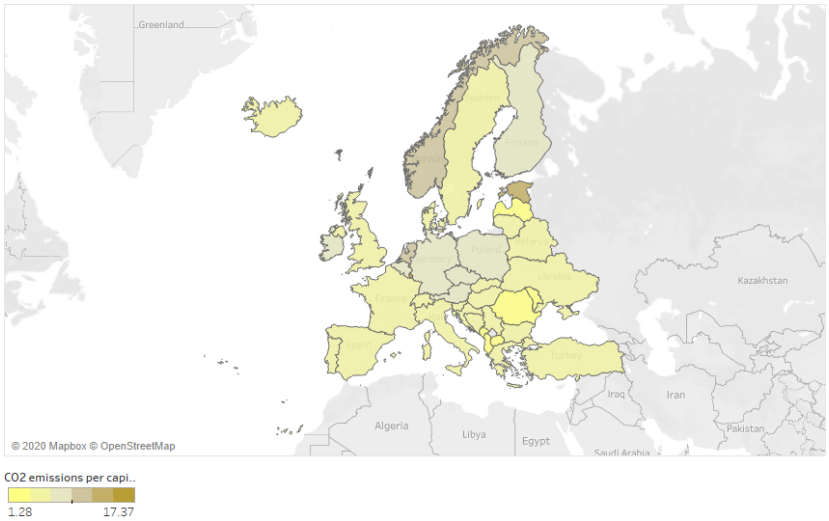

Figure 2.12: CO2 emissions (metric tons per capita).

- Policy choices determined by political, economic, environmental and geopolitical factors.

Low-income, developing countries tend to rely heavily on biomass energy and coal [25]. As countries become wealthier, new technologies enter the market, and they substitute higher quality energy sources [6].

Coal

Although the EU electricity system has innovated and become greener, it has also maintained its oldest and most polluting component: coal. As mentioned above, it is usually under-developed countries that rely heavily on coal. However, political decisions are also key in this sense. Poland has seen its economy grow steadily in recent years, but still relies heavily on coal for electricity production. European countries which lean on coal the most are listed in *Table 2.7*.

Electricity Prod. from Coal (% of total)		
Country	2015	2000
Kosovo	97.5	97.6
Poland	80.91	96.3
Serbia	72.4	62.8
Bosnia	63.4	50.7
North Macedonia	58.4	76.5

Table 2.7: Countries which rely on coal sources for electricity generation the most.

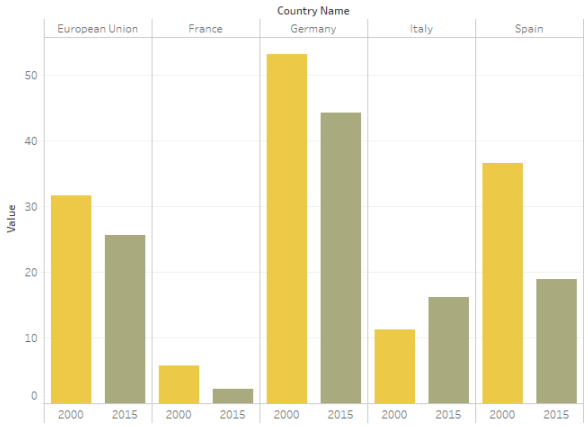

Figure 2.13: Electricity Production from Coal (% of total).

The trend in reducing the impact of coal in the energy mix has been set in all corners of Europe, and now the focus is on the timing. Even Germany, a country traditionally careful on environmental matters, in 2015 still relied on coal for 44% of its energy production. The share of fossil fuel in the EU electricity generation mix stands at 25 percent, having declined by only 5 percentage points between 2000 and 2015. This is a problem, since as [26] points out, to generate the same amount of electricity, a coal-fired power plant emits 40 percent more CO_2 than a gas-fired power plant and 20 percent more than an oil-fired power plant. *Figure 2.13* displays where major EU countries stand in the path towards decarbonization.

Oil

One traditional way of producing electricity is by fuel combustion: the raw material, oil, sits in deep underground reservoirs. Since the ultimate amount of oil is finite - and cannot be replenished once it is extracted and burnt - it is not a renewable resource. Burning oil to generate electricity produces significant air pollutants, and this may also explain why this energy source is going out of fashion: in the 70s oil combustion was one of the main techniques for electricity generation, accounting for about one fifth of the total. In following years, however, this share started to shrink constantly: in 2015

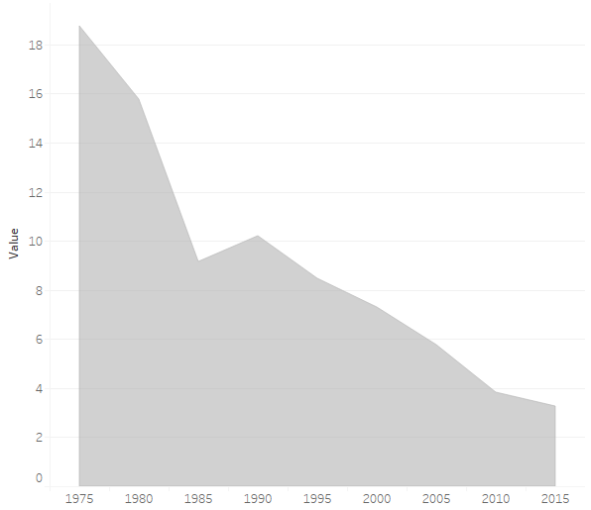

Figure 2.14: Electricity Production from Oil (% of total).

only less than 4% of electricity is generated by fuel combustion. *Figure 2.14* shows the trend in electricity generation from fuel combustion.

Since oil can be used for producing electricity, one could easily infer that the price of the former could have an effect on the price of the latter. Oil price has showed significant fluctuations in the last 20 years. From 1999 until mid 2008 the price of oil rose significant thanks to the rising oil demand in Asian countries. The 2007-2008 financial crisis corresponded to a drop in the price of crude oil, followed by a fast recovery in the following years. The world price of oil was above $125 per barrel in 2012, and remained above $100 until September 2014, after which it entered a sharp downward spiral, falling below $30 by January 2016. The cause is in the so-called *oil glut*, a serious surplus of crude oil that started in 2014 and accelerated in the next two years, with multiple causes. These include general oversupply as United States and Canadian oil production (obtained by fracking) reached critical volumes, geopolitical rivalries among oil-producing countries, falling demand across markets due to the slow down of the Chinese economy, and possible restraint of long-term demand due to environmental concerns. *Figure 2.15* represents graphically what happened in the last 20 years, with all the up and downs the price of oil experienced.

Nuclear Power

Electricity can also be produced by a process which makes use of nuclear reactions. This type of energy has one of the lowest levels of fatalities per unit of energy generated compared to other energy sources, given that electricity generation by coal and oil combustion result in air pollution [18]. However, nuclear power is going out of fashion mostly because of Chernobyl and Fukushima disasters, which have called into question its use in Europe. Germany was one of the most radical governments in abandoning its nuclear energy program, having the greatest number of permanently closed nuclear

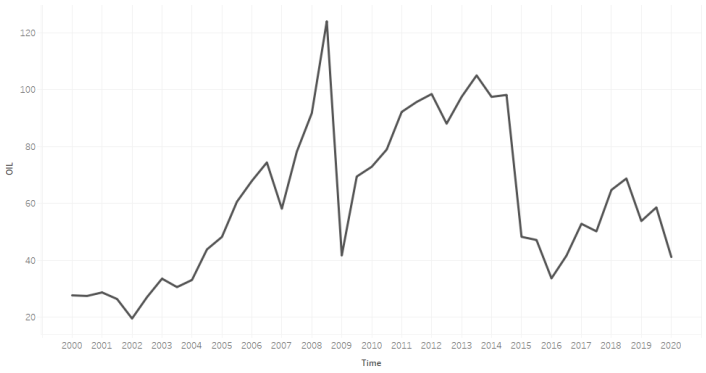

Figure 2.15: Electricity Production from Oil (% of total).

Nuclear Energy (% of total)		
Country	2015	2000
France	77.6	77.6
Slovakia	56.9	53.6
Hungary	52.2	40.3
Slovenia	38.1	35
Belgium	37.5	58.1

Table 2.8: Countries which rely on nuclear reactors for electricity generation the most.

plants among the 27 European Union member countries. In fact, it plans to shut down all remaining nuclear reactors by 2022. Having taken opposite political decisions, France still relies heavily on nuclear power generation. It has the second highest number of operable nuclear reactors worldwide and it is supposed to build an additional nuclear reactor. Leaving out France, most of the other European countries are going in the direction of a reduction in the share of energy produced by nuclear reaction. Only Slovakia is greatly investing in broadening its nuclear energy program, looking to add two further reactors to the four already in use. *Table 2.8* summarizes the trend in nuclear energy production: the share of energy produced by reactors as percentage of the total in 2000 and 2015 is displayed.

Renewable Sources

The use of renewable energy for producing electricity has several potential benefits, including a reduction in greenhouse gas emission and a reduced dependency on fossil fuel markets (in particular, oil and gas). Additionally, it is considered safer with respect to nuclear production, due to the disasters experienced in the last 40 years. The increase in electricity produced through renewable sources has been steady in the last 20 years, going from a share of 9.6% of the total in 2004 to 18.9% in 2004. While the EU as a whole is on course to meet its 2020 targets, some member states still need to make additional efforts to meet their obligations. *Figure 2.16* displays the renewable

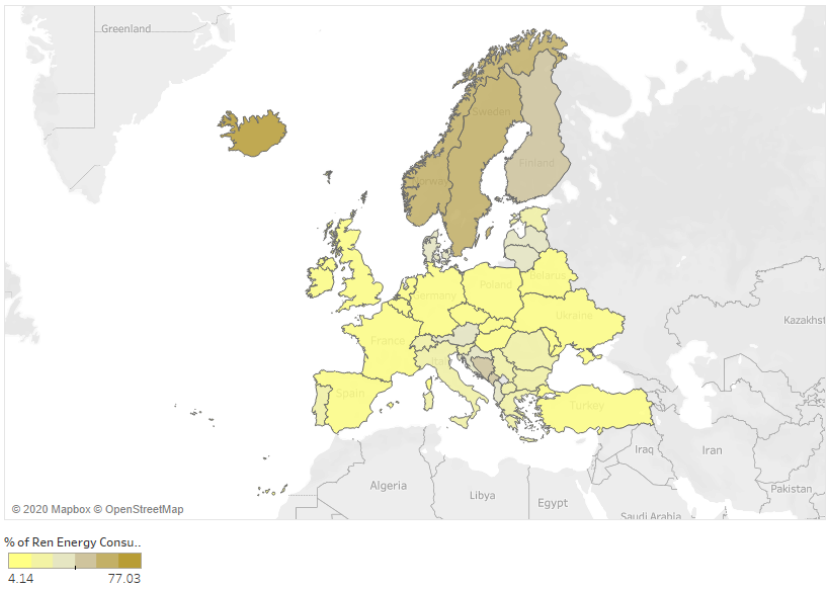

Figure 2.16: Renewable Energy Consumed (% of total).

energy consumption levels for European countries.

Northern Europe countries have invested significantly in renewable energy sources and the result is that their achievements in the path towards sustainability are particularly robust. Iceland is particularly strong in geothermal energy while Sweden and Norway make use of their considerable supply of moving water and biomass. In fact, "renewable energy sources" is a broad categorization including:

Hydro Power, which is the most relevant renewable energy source. Since water is so dense, even a slow flowing stream can result in a sizable amount of energy.

Geothermal Power, which is obtained from thermal energy stored in the ground.

Biomass, which can either be used directly via combustion to produce heat, or indirectly by converting it into biofuel.

Wind Power, which is preferred in areas where winds are strong and constant, as in the case of high-altitude sites.

Solar Power, which is having a rapid increase in importance. Italy has the largest proportion of solar electricity in the world (7.7% in 2015).

The cost of the different renewable energy generation techniques is uneven, so these variables should in principle be considered independently. However, the data sources on which this study this is counting do not disclose information with this much level of detail, using instead the aggregate "renewable" identifier.

Electricity Dependence (%)		
Country	2010	2018
Austria	62.8	64.3
Belgium	77.9	24.3
Bulgaria	40.1	82.3
Cyprus	100	92.5
Czech Rep	25.3	36.7
Germany	60	63.6
Denmark	-16	23.7
Estonia	15.5	0.7
Greece	68.6	70.7
Spain	77.1	73.3
Finland	48.8	44.9
France	48.7	46.6
Croatia	46.7	52.7
Hungary	56.9	58.1
Ireland	87.1	67.4
Lithuania	79	74.2
Luxembourg	97	95.1
Latvia	45.5	44.3
Malta	99	97.8
Netherlands	28.3	59.7
Poland	31.6	44.8
Portugal	75.2	75.6
Romania	21.4	24.3
Sweden	37.8	29.2
Slovenia	49.5	51.3
Slovakia	64.4	63.7

Table 2.9: Dependency on electricity imports.

2.5.5 Dependency from Imports

European Union member states import a significant amount of electricity from other countries, such as Norway and Russia. In total, the import dependency rate corresponded to 58% in 2018, which means that more than half of the EU's energy needs were met by net imports. The result is that Europe heavily relies on them for its supply and to a certain extent has to conform to others' prices. *Table 2.9* summarizes dependency from import rates for European countries. Malta, which lacks the necessary resources for producing electricity and never consistently invested in renewable energy, is entirely depending on abroad sources for its supply. At the moment none on the European Union countries is self-sufficient, although Denmark in the beginning of the century was, thanks to the North Sea production of oil and gas.

Chapter 3

Methodology

Many authors have tried to measure the effect of market liberalization and green energy investment, starting from [20], in which the analysis is performed on European Union member countries' markets. The authors from [22] take a broader perspective, looking at more countries but with less recent data. A strictly econometric approach is used in these studies: to validate the results algorithms are not tested on a separate data set.

The main focus here is on predicting how prices for electricity are going to evolve in the future, and not merely on understanding the causal relationship. In this sense, while the European Union is the main market on which the analysis is based, also *World Bank* data is used as a source of information for countries all over the world. This section is structured as follows: first, models are built on top of *Eurostat* data sets, and then to benchmark those findings, global patterns are investigated.

3.1 Static European Models

The variables in consideration for the European market models seek to resemble the factors listed in the introduction. The response variable, electricity price itself, is expressed in *Euros* and involves households who consume between 2,500 and 5,000 kWh. This means that the focus is on a certain category of consumers: families and not industries. For a panel data analysis on non-household consumers please instead refer to [7]. The price used in the models as response variable does not include taxes. Most remarkably, electricity prices' information is recorded twice a year, so in the models year variability is accounted. The time range chosen for the analysis is the last 20 years, from 2000 to the first semester of 2020. This makes the data at hand and the consequent results extremely up-to-date. Minor adjustments had to be carried out because of inconsistencies in the way of recording information between years prior to 2007 and years that follow. The electricity price is going to serve as the outcome variable in the later discussion, and referred as **PRI**, for short.

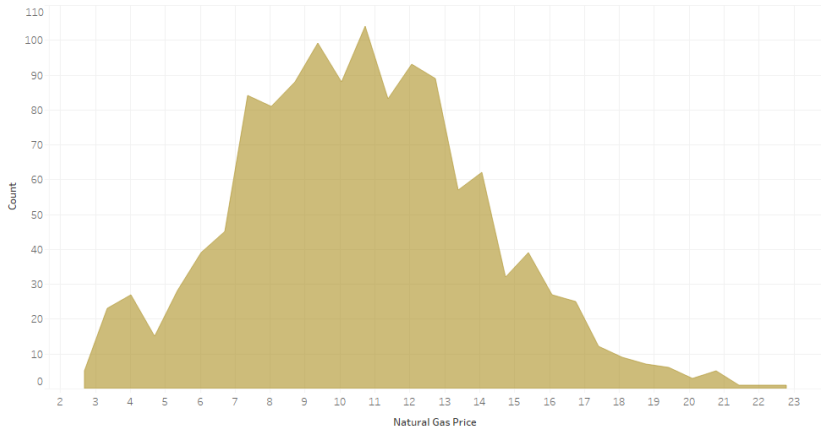

Figure 3.1: Natural Gas Price Distribution.

The two main bills households receive contemplate electric power and **natural gas**, for cooking, heating, lighting and using domestic appliances inside the building. The two goods are strictly related, because electric power can be produced from gas, and because electricity can represent a substitute of gas (e.g. for households that use induction cooktops). Data is taken from *Eurostat* and is also in this case bi-annual and measured in *Euros*. Taxes and levies are not considered: the price is the sum of the raw material value and the cost of transportation and distribution. *Figure 3.1* shows the distribution of the gas price, with the median being around 10 - 11 euros per mega joule. In the later model discussion this variable is going to be appear under the name **GAS**, for short.

The idea is to use the price of the resources from which electricity is produced to have an idea of what the price of the good could be. The choice of having the natural gas price as a variable in the data set goes in this sense. It would be reasonable to actually do the same for oil, but *Eurostat* does not disclose any information regarding this good price. Options would be either to use the US dollar price per barrel or the pump price for gasoline. The choice falls on the latter, being considered as a good proxy for the price of oil in the country. The data is available from the World Bank, and these information is disclosed once every two years. In the later models it is going to be referred as **OIL**, for short.

Electricity, as discussed in the introduction, is not only produced by burning coal, oil or natural gas, but also by using **renewable resources**. The share of electricity produced using sustainable practices on the total of electricity produced by each single country is also included in the data set, and named as **REN**, for short. This information is recorded once a year and available from *Eurostat*. However, some years the information is incomplete, specifically records prior to 2004, so those missing values are topped up with data from *World Bank*. The sets from the two different sources describe the same information, e.g. for data regarding 2004, numbers are almost equivalent. However, just pasting data could be dangerous so normalization of the World Bank information is carried out by multiplying each country value for 2000-2003 by a factor that makes

records for 2004 match for both sources.

Another variable that takes in consideration the level of sustainability of the electricity produced is the emission **intensity**. This indicator is used to monitor progress towards United Nations' goal on climate action and on affordable and clean energy and, as anticipated in the introduction, it is calculated as the ratio between energy-related GHG emissions and gross inland consumption of energy. Information is normalized with respect to year 2000, which is used as a benchmark. This variable hence expresses how fast that particular economy is in its path towards decarbonization. Data is recorded annually and sourced from *Eurostat*: in the later discussion it is going to be referred as **EMI**, for short.

It is also important to consider the supply side in this context: how much each country is consuming in terms of energy. *Eurostat* provides information regarding the quantity of oil equivalent (in kilograms), and this data is recorded annually. This indicator measures how much energy every citizen consumes at home excluding energy used for transportation. Since the indicator refers to final energy **consumption**, only energy used by end consumers is considered. The related consumption of the energy sector itself is excluded. This variable is going to be shortened as **CON** in the following model discussion.

Many European countries do not really have the capacity to produce enough electricity in order to be self-sufficient. The **dependency** variable indicates the share of total energy needs of a country met by imports from other countries. It is computed from energy balances as net imports divided by the gross available energy. The formula for getting the dependency is:

$$\frac{(imports – exports)}{gross\ available\ energy}$$

A negative value indicates a net exporter: a country that produces more energy than the one consumed in its internal market. This is true only for *Norway*, among European countries. Values higher than 100 generally refer to the accumulation of stocks (increase of fuel in stock), yet might also be the result of statistical discrepancies in raw data. Import dependency information is available from *Eurostat* and recorded once a year. This variable is going to be shortened as **DEP** in the models discussed later.

In the last 20 years the major change in the way electricity is sold has been the **liberalization** forced to member countries by the European Union authorities. The only information regarding the structure of the markets disclosed by *Eurostat* indicates the market share of the largest electricity generator. This information is disclosed annually, and in the following models it is going to appear as **MAR**, for short. As also discussed in the Introduction, to rely merely on the share of the largest generator may be inadequate to have a thorough idea of the structure of the market. The best indicator for market concentration is the HHI (Herfindahl-Hirschman Index) [27], and it is computed by summing the square root of the market share of each individual firm in the industry.

The electricity price is also, and maybe most importantly, dependent on how wealthy that particular country is. *Eurostat* provides the value of the **real gross domestic product** per capita, with data being published every quarter of year. "Real" in this context means deflated, so leaving prices unchanged. This is key: while the gross domestic product tends to increase every year (given that the country is an in inflation regime), the real GDP might be decreasing if on average the citizens have a lower purchasing power. The reader could wonder why to exclude the average increase in prices from the consideration, although there is a clear correlation between price increase and electricity price: the higher the increase in prices the higher the expected increase in the electricity bill. The answer lies in the fact that the **inflation** variable *hides* the electricity price change variable itself: the average increase in prices is computed on a basket of goods electricity is part of, so it would be too easy to actually use this variable in the models. The outcome variable exogeneity assumption would not be met. In the following models, the real Gross Domestic Product per capita variable is going to be called **GDP**, for short.

3.1.1 Features

In the predictions with the static setting, most of the variables are first being reshaped through a logarithmic transformation before entering the models. The only variables that are not transformed are the ones measured on a percentage scale (i.e. *REN*, *EMI*, *DEP* and *MAR*). The logarithmic transformation is carried out in order to make the variables' distribution as normal as possible so that the resulting statistical analysis becomes more valid. The reshaping lessens the influence of the potential outliers and reduces the skewness of the original data. Since the transformation is also applied to the outcome variable, this ensures that the final output is not negative. This is true because applying the sequence of logarithmic and exponential transformation guarantees the final electricity price to be strictly positive. For these first static models the variable *EMI* is going to be excluded from the analysis.

3.1.2 Countries

Eurostat provides information regarding the variables above mentioned for the members of the Union, of course, but also for some neighboring countries, which are included in the study. The idea is to understand whether the membership to the Union does have an effect when trying to predict the final price for electricity. *Figure 3.2* shows which are the European countries included in the study.

As discussed in the previous subsection, most information data is not disclosed twice a year but only once: instead of dropping those rows where information data is not available, the average between the previous and the following semester is computed and inserted in the data set. Other missing data substitution techniques are more sophis-

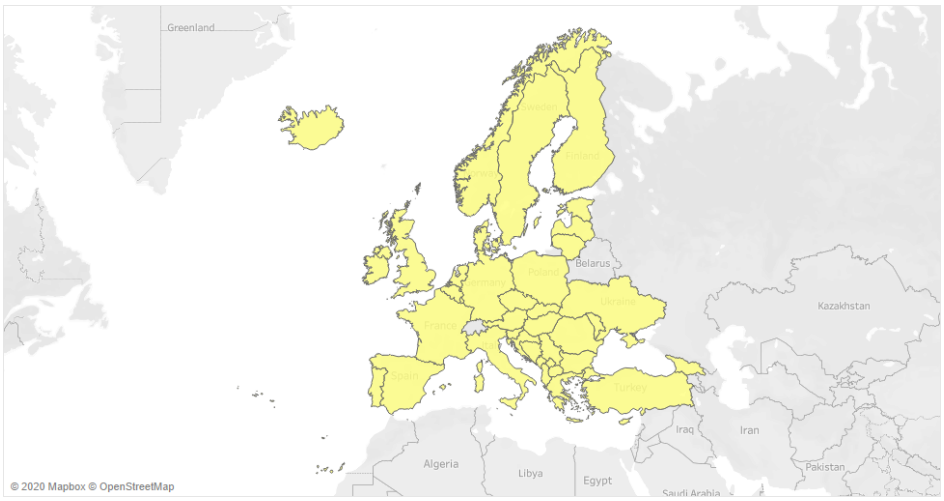

Figure 3.2: Countries which are included in the study.

ticated: for instance, Netherlands does not disclose information regarding electricity market concentration levels. To have an idea of the retail market for this country, an interesting reference is [21]. The way out of this found for not discarding Netherlands from the analysis is to substitute those Not Availables with average values from markets with a similar structure, which are Germany, Denmark, Luxembourg and Belgium.

The final data set is incomplete: many countries (e.g. Ukraine) started disclosing information regarding electricity price only in recent years. This means that for several states there are not going to be 20 observations, one for each year, but less.

3.1.3 Random Effects

Statistical models always describe variation in observed variables in terms of systematic and unsystematic components. In econometrics, a random-effects model is a statistical model in which some of the parameters (effects) that define systematic components of the model exhibit some form of random variation [24]. In the context of the problem at hand, the parameters (effects) that shall be taken as random factors are the country identifiers. For the time being, the assumption is that, given certain values for market concentration, share of renewable energy consumed and so on, no country is different from any other. This is not such an unlikely hypothesis: after all, the countries in the analysis all belong to the European continent. The summary of the random effects model is displayed below, in *Table 3.1*.

The model at hand captures 60% of the variability of the data points. All variables are significant at the 95% confidence level: GDP, natural gas price, oil price and dependency from imports are all positively related with the electricity price. Interestingly enough, the coefficient for share of renewable energy produced is negative, meaning that countries with more sustainable energy generation practices are expected to have

Coefficients				
	Estimate	Std Error	T-Value	5% Sign
(Intercept)	-4.1	0.09	-44	*
log(GAS)	0.2	0.02	9	*
DEP	1E-4	6E-5	2	*
log(CON)	-0.08	0.01	-6	*
REN	-5E-3	5E-4	-10	*
log(GDP)	0.2	0.01	22	*
log(OIL)	0.3	0.03	11	*
MAR	-1E-3	2E-4	-5	*

Global Performance			
R-Squared	0.597	Adj R-Squared	0.595

Table 3.1: Random Effects (RE) model summary.

Analysis of Variance Table						
	Res. DF	RSS	Df	Sum of Sq	F	Pr($> F$)
1	1231	21.176				
2	1271	55.337	-40	-34.16	49.6	2.2E-16

Table 3.2: Anova test to compare Model *FE* and *RE* models.

a lower electricity price, leaving everything else equal.

3.1.4 Fixed Effects

The Random Effects model, as discussed before, does not consider geographic information: the effect of country variables is assumed to be accidental. It is now time to check whether that assumption is fair or not. The Fixed Effects model implemented in this section considers all the variables already considered in the Random Effects model but also includes the geographical identifier. An Analysis of Variance test is a way to find out whether experiment results are significant. In this case the null hypothesis is that there is no difference in significance between a model with the geographic identifier and one without, the alternative hypothesis is that there is a difference. *Table 3.2* displays the R output when the *anova* function is called on the two models.

40 degrees of freedom are lost by including the country identifier, but there is a corresponding drop in the Residual Sum of Squares. In this case, the *anova* test suggests that the geographical information should be included in the model, since the p-value is very close to zero. *Table 3.3* summarizes the fixed effects model listing its coefficients and corresponding confidence levels.

Coefficients			
	Estimate	sError	p
(Intercept)	-4.37	0.20	***
log(GAS)	0.29	0.02	***
DEP	-0.00	0.00	
log(CON)	-0.05	0.01	***
REN	0.00	0.00	
log(GDP)	0.19	0.03	***
log(OIL)	0.25	0.02	***
MAR	-0.00	0.00	*

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Global Performance			
R-Squared	0.8458	Adj R-Squared	0.8399

Table 3.3: Fixed Effects (FE) model performance summary, geographical dummies not included in the table.

The *FE Model* captures around 85% of the variability of the data points. The coefficients for the geographical identifier dummy variables provide an indication whether the single countries pay more than they should given their wealth, market concentration, consumption levels and so on. To provide an example, Belgian households pay for electricity bills an amount which is significantly higher than what would be expected, while the opposite happens for Serbia.

It is interesting to notice how the coefficients of some of the variables of the *RE model* change when moving to *FE model*. For instance, the already discussed Renewable Energy Share this time shows a negative coefficient, even though it is not significant at the 95% confidence level. This means that, opposite to what said before, a higher share of energy produced sustainably leads to a decrease in the price of electricity, everything else left equal. The reason for this discrepancy is that countries with a high share of renewable energy production do have a lower geographical dummy variable (e.g. Denmark and Norway coefficient is -0.11, Sweden coefficient is -0.1).

Creating a 0-1 dummy variable for each one of the countries in the data set, fixing the geographical effects, reduces the degrees of freedom, and it is also too aggressive in simplifying the regression problem. For instance, the *FE model* suggests that Northern Europe households pay less than they should for their electricity, but any information regarding why this is the case is not available. Is it because, as was suggested in the *RE model*, the **REN** variable is higher, is it because these countries are closer to Russia, which is a strong Natural Gas provider or for some other reason? Not enough information is available yet to a take a firm stand on this.

An alternative solution, rather than relying on so many geographical identifiers or on

Coefficients			
	Estimate	Std. Error	p
(Intercept)	-3.81	0.10	***
log(GAS)	0.21	0.02	***
DEP	0.00	0.00	*
log(CON)	-0.10	0.01	***
REN	-0.00	0.00	***
log(GDP)	0.17	0.01	***
log(OIL)	0.32	0.03	***
MAR	-0.00	0.00	***
EA	0.23	0.02	***
EFTA	0.24	0.04	***
EU	0.20	0.02	***

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Global Performance			
R-Squared	0.6272	Adj R-Squared	0.6242

Table 3.4: EU model summary.

none of them, would be to include information regarding the membership to the European Union. Countries which use the Euro as their currency belong to the Economic Area group (**EA** dummy), other member states are referred to as **EU** countries (examples are Sweden, Denmark, which use their currency). Among the non EU-member states, an **EFTA** (European Free Trade Association) dummy variable is introduced to refer to countries as Norway that still have access to the common market. The set of commands to obtain the expanded data set with the dummy variables is displaced in the appendix. *Table 3.4* summarizes the *EU model* output.

Table 3.4 summarizes coefficients and performance of the *EU model*, which includes the European membership identifier among its variables. Countries which have some kind of bond with the European Union, both members and free trade associates, result having a higher intercept in the regression problem: their citizens are expected to pay a higher price for electricity with respect to citizens from outside countries.

3.1.5 Checking Assumptions

The relationship between the predictors and the outcome variable is assumed to be linear. The Residuals vs Fitted plot is used to check the linear relationship assumption. A horizontal line, without distinct patterns is an indication of a linear relationship. In *Figure 3.3*, data points are plot in the Residual vs Fitted space, so with the predicted value on the x-axis and the residual value on the y-axis.

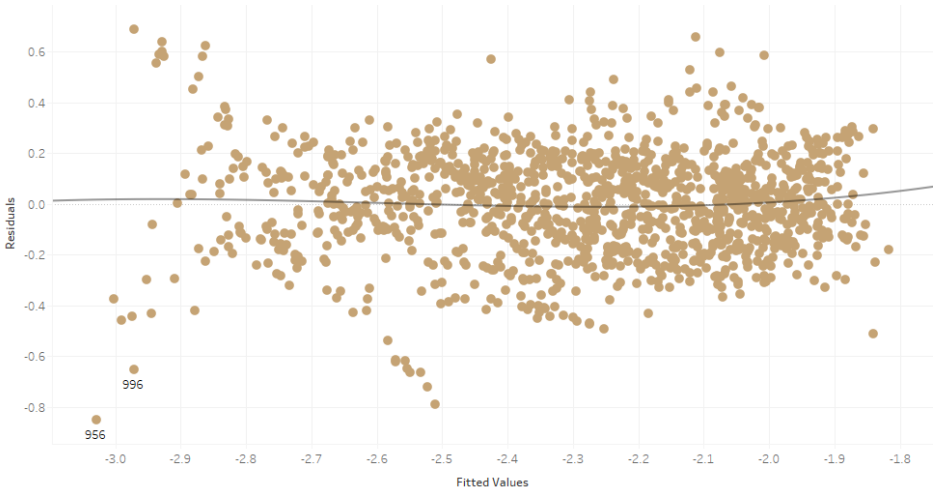

Figure 3.3: Residuals vs Fitted plot.

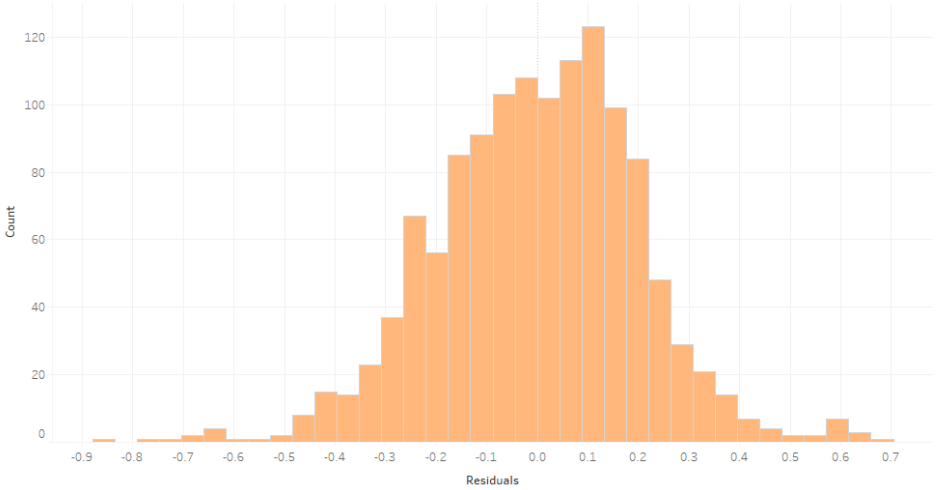

Figure 3.4: Distribution of the Residuals.

Since no distinct pattern is recognizable, the relationship between the predictors and the outcome variable is assumed to be close to linear. Points 956 and 996 seem to positioned quite far from the others, on the bottom left: they are two observations from Ukraine, a country which produces electricity at a low price thanks to its natural resources. Hence, the hypothesis that those are mistaken records is discarded. *Figure 3.4* shows how the residuals are distributed.

Residuals show a distribution which is slightly skewed to the left: it happens more often that the simple linear model underestimates the real value with respect to a no residual situation. However, also given the large number of data points, this small deviation from perfect normality is not too alarming. One note: as already discussed, the logarithmic transformation has been applied to the outcome variable at hand, so a 0.1 residual does not correspond to a 10 cents of a dollar estimation error.

Variance Inflation Factor			
	GVIF	Df	$GVIF^{\frac{1}{2Df}}$
log(GAS)	2.1	1	1.4
DEP	2.4	1	1.5
log(CON)	1.4	1	1.2
REN	1.8	1	1.3
log(GDP)	2.6	1	1.6
log(OIL)	1.5	1	1.2
MAR	1.3	1	1.1
EU	4.9	3	1.3

Table 3.5: VIF test.

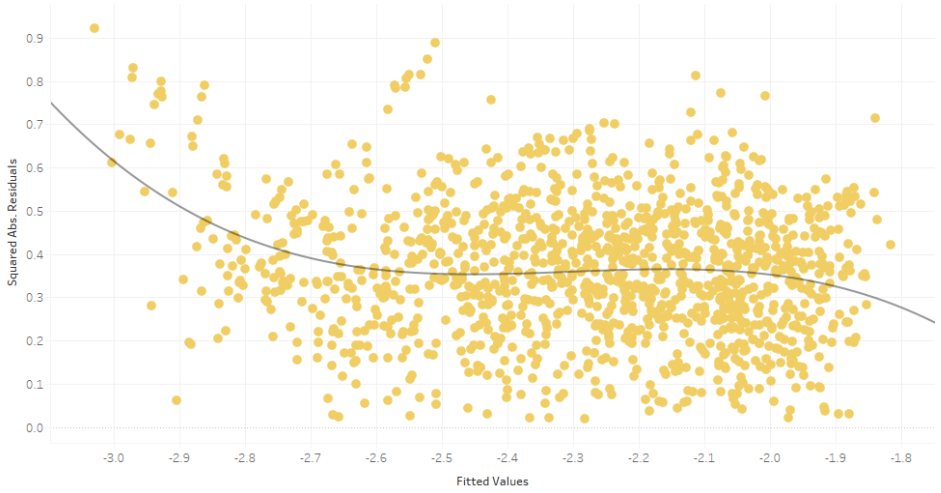

Figure 3.5: Scale-Location plot.

Another problem that can pop up in regression models is the collinearity or the multi-collinearity of the predictor variables. In the presence of collinearity, the solution of the regression model becomes unstable. This hurdle can be assessed by computing a score called the variance inflation factor (or **VIF**), which measures how much the variance of a regression coefficient is inflated due to multi-collinearity in the model. *Table 3.5* displays the result of running VIF on the variables from *EU model*.

Since GVIF and adjusted GVIF values are rather low, well under the critical 10 threshold, multi-collinearity can be safely considered as not an issue in this case.

Figure 3.5 displays the scale-location plot, which is used to check homoscedasticity, i.e. the homogeneity of variance of the residuals. On the horizontal axis the fitted values are reported, as in *Figure 3.3*, while on the vertical axis the residual value is taken in its absolute value and then transformed using a square root operation.

Equally spread points and consequent horizontal line of fit is a good indication of

BP Test		
BP	df	p-value
179	10	<2E-16

Table 3.6: BP test.

BP Test		
BP	df	p-value
114	10	<2E-16

Table 3.7: BP test after having applied BC transformation on outcome variable.

homoscedasticity: in this case, hard to give a definite answer. On the left-hand side of the plot, so when the fitted values are particularly small, the residuals tend to be greater. To investigate further, one idea is to run the **Breusch-Pagan** test, which fits a linear regression model to the previous residuals and rejects the homoscedasticity hypothesis if a sizable portion of the variance is captured by the additional explanatory variables. *Table 3.6* provides a summary.

The test detects heteroscedasticity. This means that the least squares estimator is still linear and unbiased, but it may no longer be the best. That is, there is another estimator with smaller variance. In order to correct for heteroscedasticity, a **Box Cox** transformation is applied on the dependent variable, electricity price. A transformation of this type makes a non-normal dependent variable into a normal shape [4]. Code for applying Box-Cox transformation on the dependent variable is left to the appendix: in any case, as displayed in *Table 3.7*, the overall situation improves but the *BP test* still detects heteroscedasticity.

3.1.6 Weighted Least Squares

Another option to better deal with non-constant variance is to deviate from ordinary least squares to use weighted least squares model. Weighted least squares corrects homoscedasticity by weighting each observation by the reciprocal of its estimated variance. Observations with small estimated variances are weighted with a higher score than observations with large estimated variances. The precise steps are reported in the appendix, but the idea is to store the residuals from the *EU model* and then run a regression on the residuals themselves. The weights for the final model are then obtained by taking the reciprocal of the square root of the exponential of the fitted values from this intermediate model:

$$W = \frac{1}{\sqrt{e^{\hat{y}}}}$$

Weights are inversely proportional to the error variance. An observation with small error variance has a large weight since it contains relatively more information than

Coefficients Comparison		
coef	OLS	WLS
intercept	-3.81	-3.71
log(GAS)	0.21	0.24
DEP	0	0
log(CON)	-0.1	-0.11
REN	0	0
log(GDP)	0.18	0.16
log(OIL)	0.32	0.3
MAR	0	0
EA	0.23	0.23
EFTA	0.24	0.27
EU	0.2	0.2

Table 3.8: Comparison of Coefficients obtained through OLS and WLS.

an observation with large error variance. *Table 3.8* compares the resulting coefficients from the *EU model*, with Ordinary Least Squares, and the linear model with weighted regression.

There are no major differences between the two models but the one built through Weighted Least Squares is supposed to be more robust. The interpretation of the resulting coefficients and standard errors remains the same as before. The fact that some heteroscedasticity is detected while not using weights is important to be noted because the estimators may not be optimal but it should not be too worrisome: most real world data is in fact heteroscedastic. If the sample size is large enough, however, the variance of the least squares estimator is supposed to be sufficiently small to get precise estimates. That is why from the following section, the use of the weights inside models is going to be carried out only sporadically.

3.1.7 Subset Selection

To have an idea of the goodness of models there are several options: one could look at the *Adjusted* R^2 values, as previously done in the chapter, or at some alternative accuracy metrics, as *AIC*, *BIC* or *CP*. R^2 is the proportion of variation in the outcome that is explained by the predictor variables: sadly, any variable which is added to the model is going to increase this accuracy metric, which is hence usually adjusted by the number of variables n. Akaike's Information Criteria (shortened AIC), Bayesian Information Criteria (shortened BIC) and Cp's basic idea is instead to penalize the inclusion of additional variables to the model. The lower the levels for these metrics, the more accurate the model.

The take-home message is that having a very large model including variables which

Subset Selection				
Excl Var	N of Vars	Adj R^2	BIC	Cp
	10	0.62	-1183	11
DEP	9	0.62	-1186	13
MAR	8	0.62	-1180	25
EFTA	7	0.61	-1158	52
EA	6	0.6	-1123	94
EU	5	0.59	-1092	133
log(CON)	4	0.57	-1058	177
log(OIL)	3	0.54	-966	292
REN	2	0.5	-858	438
log(GAS)	1	0.4	-647	753

Table 3.9: Comparison of accuracy metrics while removing one variable at a time.

are not really significant can decrease the likelihood of the final predictions produced being reliable. The next step is to assess whether the model with all the variables from the *EU model* is supposed to be the best or it is better to focus on a subset of those predictors. *Table 3.9* displays the performance of the different models while one variable at a time is excluded from the model.

The first variables to be excluded are Dependence on Imports and Market Concentration, followed by the geographical dummy variables. Contrarily, log(GDP) does not appear in the table which means that it is the last variable to be left. To decide which the best model is one has to look at the accuracy metrics: adjusted R^2 and Cp would suggest to opt for the model including all the variables, while BIC suggests to exclude DEP.

The problem with the accuracy metrics described above is that they are computed on the same data set that is used for training the model. When there is enough information availability an option is to use a a part of the data to actually measure performance only. For instance, to understand which one of two models is the most reliable it is only necessary to compare the performance of the two models on the real-world data that is left for validation. In the case of the electricity data set at disposal, the idea is to split information on a temporal basis: the training set includes data prior to 2015, validation set includes the most recent information.

At this point, simple and weighted linear regressions are recomputed on the smaller size training set. Performance of these two models on the validation set is reported in *Table 3.10*. The accuracy metric used is the *Mean Absolute Error*, which is computed as the arithmetic average of the absolute errors.

Ordinary Least Squares seems to outperform Weighted Least Squares when testing the models on the validation set. Adjusted R^2, BIC and Cp metrics suggested to keep all the variables in the models, or in the worst case drop DEP. Now it could

Mean Absolute Error			
OLS	1.81	WLS	1.89

Table 3.10: Comparison of performance between OLS and WLS, measured in Euro cents.

Mean Absolute Error, OLS									
log(GAS)	REN	log(GDP)	log(OIL)	MAR	EA	EF	EU	NVars	MAE
x	x	x	x	x	x	x	x	8	2.02
x	x	x	x		x	x	x	7	2.05
x	x	x	x		x	x		6	1.97
x	x	x	x			x		5	1.96
x	x	x	x					4	2.01
	x	x	x					3	2.22
x		x						2	2.03

Table 3.11: Models obtained through OLS and corresponding error.

be interesting to check whether also Mean Absolute Error result proves to go in this direction. *Table 3.11* summarizes information regarding the best model for N=1,...,10 where N is the number of variables in the model. Resulting Mean Absolute Error, measured in Euro cents, is reported in the final column. Models are obtained through Ordinary Least Squares.

As expected, the best performing model is the one built including all the variables of the data set. As discussed earlier, however, Ordinary Least Squares may not be the only option. *Table 3.12* summarizes what happens when the data points effect on the model is weighted by W.

Interestingly enough, the model which occurs to be the best performing on the validation set is not the one including all the variables but the one which includes only

Mean Absolute Error, WLS									
log(GAS)	REN	log(GDP)	log(OIL)	MAR	EA	EF	EU	NVars	MAE
x	x	x	x	x	x	x	x	8	2.07
x	x	x	x		x	x	x	7	2.09
x	x	x	x	x	x			6	2.08
x	x	x	x	x				5	2.07
x	x	x	x					4	1.95
	x	x	x					3	2.23
		x	x					2	1.75

Table 3.12: Models obtained through WLS and corresponding error.

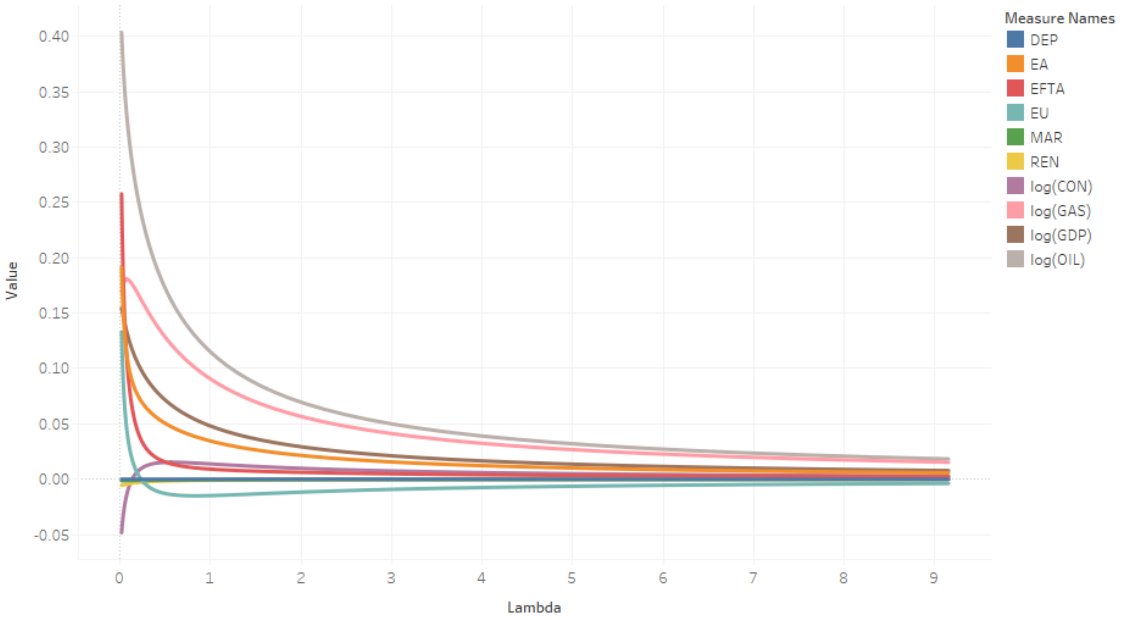

Figure 3.6: Coefficient Estimates for ridge regression.

two, Gas price and Real GDP per person. This outcome seems to oppose the idea that using all the variables is the best available option. *WLS* with two variables is also able to outperform the best *OLS* model by more than half a cent. All the code for measuring performance for Ordinary Least Squares, Weighted Least Squares and Exhaustive Subset Selection is left for the appendix.

3.1.8 Ridge Regression

Rather than completely excluding some predictor variables it could be beneficial to use smaller coefficients with respect to the ones of the linear regression model. The linear model could be wrong in assigning high coefficients to variables that are not as important, rather than being wrong in including them in the model. To shrink the coefficients one idea is to use Ridge regression: the resulting model is similar to least squares, except the coefficients are determined by minimizing a slightly different quantity, which includes a *shrinkage* penalty λ. The use of the penalty has the effect of shrinking the estimates towards zero. As λ approaches infinity, the impact of the shrinkage penalty grows, making the coefficient estimates converge to zero. [14] This property of ridge regression is showed in *Figure 3.6*, in which a new model for the electricity price prediction is built.

The next step is now to decide where to draw the vertical line on the graph to settle for a specific value of the parameter λ. The best way to do so is, once again, using part of the data to measure performance, in this case through cross-validation. The λ parameter which results as the best performing is 0.018, so extremely low: this is an

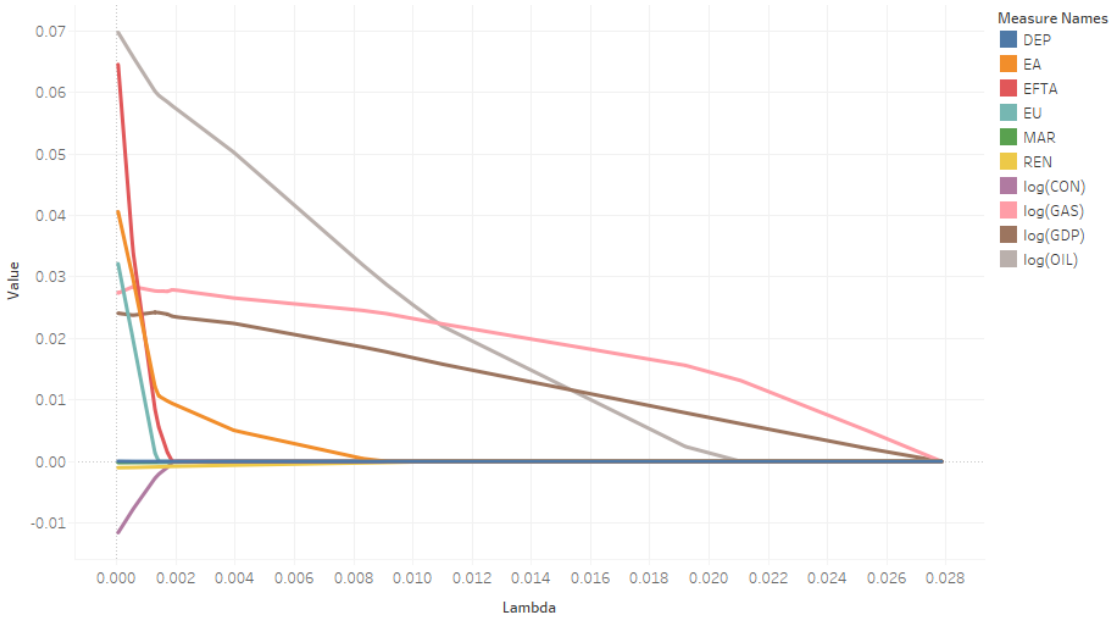

Figure 3.7: Coefficient Estimates for lasso.

indication that regularizing the parameters may not be that beneficial, in this context. In fact, the average value error of the prediction with respect to the actual electricity price is of about 1.8 cents, not a real improvement with respect to previous models.

3.1.9 Lasso

One drawback of ridge regression models is that even variables that are not significant do have a non-zero estimate. Ridge regression includes all predictors in the final model, since increasing the value of λ will tend to reduce the coefficient estimates, but will result in no variable exclusion. If however the shrinkage penalty formulation is slightly changed, some of the coefficients shall be forced to be exactly equal to zero when the tuning parameter λ is sufficiently large. This type of regularization can be considered as a sort of blend between ridge regularization, because the coefficient estimates are shrunk towards zero, and subset selection, because some of the coefficient estimates are effectively turned to zero. As it was the case for ridge regression, the choice of the λ value is critical: cross validation on the training set, once again, is going to be used to settle for a specific value of this parameter. *Figure 3.7* displays how the coefficient estimates shrink to zero while the value of λ is increased.

The way the coefficients shrink to zero is less regular with respect to what happens for ridge regression. *GAS*, *OIL* and *GDP* coefficients are the last ones to get to zero, a further hint that they are the most significant variables at disposal. Once again, the λ value that results as the best performing through cross validation is low, $5 * 10^{-5}$, so this lasso regression is quite close to linear model regression without regularization.

Models with Interactions						
GDP	GAS:OIL	CON:REN	GAS:REN	GAS:GDP	NVars	MAE
x	x	x			3	2.1
			x	x	2	1.69
				x	1	1.78

Table 3.13: Variables and Errors of the models with interactions.

Also the mean absolute error on the training set does not decrease: 1.81 cents. Lasso and Ridge regularization techniques respectively use so-called L_1 and L_2 penalties. One can go from one system to the other simply trying the model and looking at the one that performs best, but another possibility is to linearly combine the two systems, with the so-called *Elastic Net* regression. In this case cross validation is not only used to choose λ but also to settle for the best α, a parameter that allows to determine the weight of L_1 with respect to the L_2 penalty to be applied to the model. In this section Elastic Net is not going to be used simply because the results with these regularization techniques is rather unsatisfactory.

3.1.10 Interactions and Polynomials

Until now one assumption was made: that the predictor variables have a linear interaction with the outcome variable. This idea is partially supported by regression plots, but some meaningful interactions may have been excluded from the model. For instance, it may be that gas price and share of renewable energy are not that significant when considered alone but the interaction, so the first multiplier by the latter, is. A situation of high gas price and simultaneous high share of renewable of energy produced may hide some meaning that only the interaction term would bring to the surface. In the case of the electricity price regression, this is exactly what happens, and the coefficient of this GAS:REN interaction term is negative, meaning that the higher its value, the lower the final electricity price. To determine which the most significant interaction terms are Subset Selection is again used, with the accuracy metric being the Mean Absolute Error on the validation set. *Table 3.13* displays the coefficients and corresponding errors for the models with the lowest number of variables (which are also the best performing ones). Please note that some of the variables are still taken in their logged form.

Allowing the models to include interactions results in a wide use of those terms, that are often preferred to the single predictor values that were used in the *EU model*. In particular, when only two variables are included as predictors, and these two variables are the interaction terms *log(GAS):REN* and *log(GAS):log(GDP)*, the result in performance is remarkable.

Models with Polynomials							
GDP	OIL	REN	DEP^2	GAS	MAR	NVars	MAE
x	x	x	x	x	x	6	2.34
x	x	x	x	x		5	2.22
x	x	x	x			4	2.41
x	x	x				3	2.18
x	x					2	2.17
x						1	1.96

Table 3.14: Variables and Errors of the models with polynomials.

Looking at interaction terms is one option, but another assumption that it could be interesting to relax is that the predictor variables enter the regression model only linearly. There may some variables, as it is the case for Dependence on Imports, which are meaningful not in their linear term, but rather in their second order polynomial. *Table 3.14* illustrates which are the variables used in the models which perform best on the validation set. Only one polynomial term, DEP^2, is included, but this does not seem to improve performance since the model with the lowest mean absolute error is a simple linear regression with GDP as predictor.

3.2 Dynamic European Models

The results discussed in the previous section are based on static models; however, the electricity price is likely to exhibit a certain degree of path-dependency, implying that the dynamic nature of this variable should be accounted for in the model. Thus, it would be interesting to re-estimate the model in a **dynamic**-panel framework. The options at this point are two: either a lagged predictor variable, which accounts for the previous record for the same country, is created, as it is done by [12], or a different outcome variable is used, changing the scope of the regression from predicting the final price of electricity to predicting the temporal difference in price from one semester to the following. To allow for this latter idea each of the predictor variables are to be changed obtaining each observation of the data set as the difference between that variable value in that semester and the variable value in the previous semester. To put it in mathematical terms:

$$y_{t_j - t_{j-1}} = \beta_0 + \beta_1 x_{1_{t_j - t_{j-1}}} + \beta_2 x_{2_{t_j - t_{j-1}}} + ...$$

The process to achieve this in R is left for the appendix. In any case most of the variables and time range used are the same as for the Static models, with the exception of course that the data set is slightly shorter (it starts from the second semester of 2000 and not from the first one). The only variable from the static models which is not going to be used is *OIL*, because not enough data is available to have good approximations of reality in a dynamic sense. Differently with respect to the previous section,

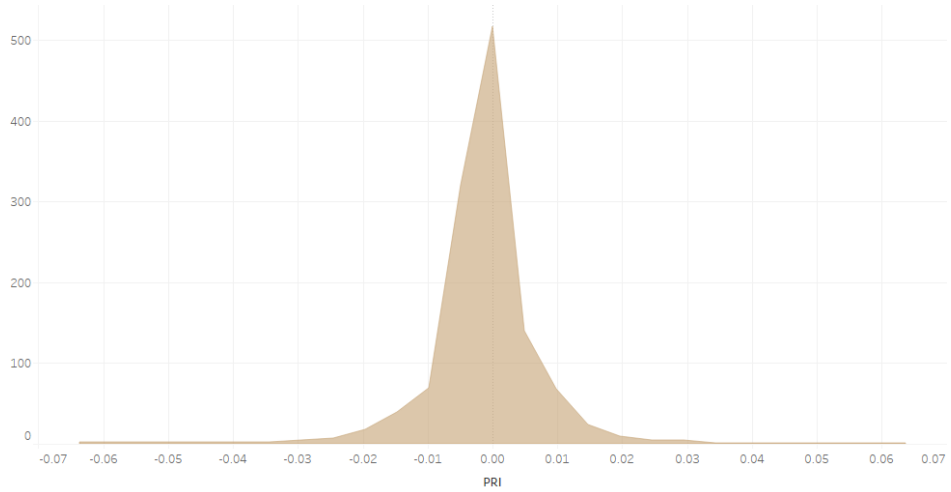

Figure 3.8: Distribution of the outcome variable, PRI.

EMI is instead going to appear in all the dynamic models. The problem with using the emission intensity in the static setting is in the fact that it is does not provide any meaningful information: for instance an *EMI* observation equal to 101 indicates a country that is getting worse in the context of its emissions per capita. If however no information is used to characterize that country with respect to others, i.e. in this context, the absence of a country identifier, makes that information quite useless when used in its absolute term. In this dynamic context, instead, if the difference from one term to the next one is -1, in the case of one percent decrease in emission intensity, or 1, one percent increase in emission intensity, that information can be used to infer the direction and magnitude of the change in price.

Of course changing the nature of the outcome variable in this way is going to significantly decrease the Mean Absolute Error of the regression: on average, the difference from one period to the other is going to be a positive number very close to zero. So the aim shifts to creating a model that is able to predict the direction and the magnitude of the change as accurately as possible. *Figure 3.8* provides an histogram of the outcome variable distribution.

The first thing to be noticed is that the distribution shows long tails: most of the observations fall within a very narrow range around zero but there is also a minority number of records falling outside this span. One may wonder where and when the price results in a 7 cent oscillation.

The anomaly in the data is found in the fact that from the second semester of 2002 to the first one of 2003 the price of electricity in Norway falls by 7 cents which are then regained in the second semester of 2003. This weird behavior seems more of a recording mistake rather than a real price fluctuation experienced by the households, given that

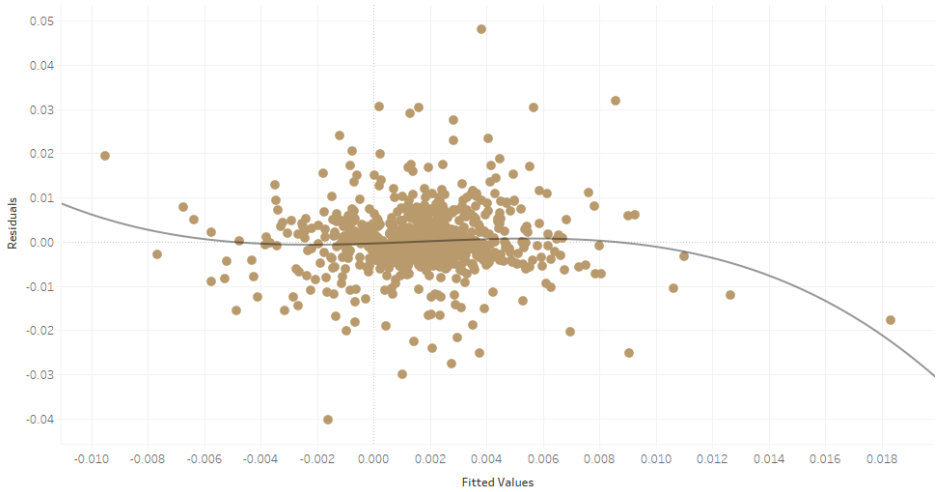

Figure 3.9: Residuals vs Fitted plot, dynamic model.

there is no explanation in the data looking at the other predictors and no sign in the literature that such an important price oscillation actually took place. Thus, the best way out of this simply results in excluding those variables from the data set.

3.2.1 Linear Model

As it was done in the static setting, the first model to be tested is linear, with multiple predictor variables. However, while in the previous section all the values showed a positive sign, in this dynamic model it will happen quite often to incur in negative observations, both for the outcome and the predictor variables. This is not a problem in the context of the linear regression, but it does not allow to apply the simple logarithmic transformation on any of these variables. In some cases, as for the outcome PRI, some slightly more complicated transformations could be applied as for instance $log(PRI + 1)$. But it would not really be worth it to go in this direction, given that there would be some significant loss in complexity and interpretability. As the linear regression name recalls, the relationship between the predictors and the outcome variable is assumed to be linear. To check whether this assumption is met, the fitted values vs residuals plot of *Figure 3.9* comes in handy.

One note: this time the model is built directly on the training set, which is composed by observations prior to year 2015, leaving subsequent years observations to validate the results.

Apart for some observations on the margins, no distinct pattern is discernible in the most data-dense part of the plot, hinting that the relationship between the predictors and the outcome variable can be safely assumed as linear. Also, it is to be noticed how the largest residual absolute value turns out to be lower than 2 cents: keeping the 7

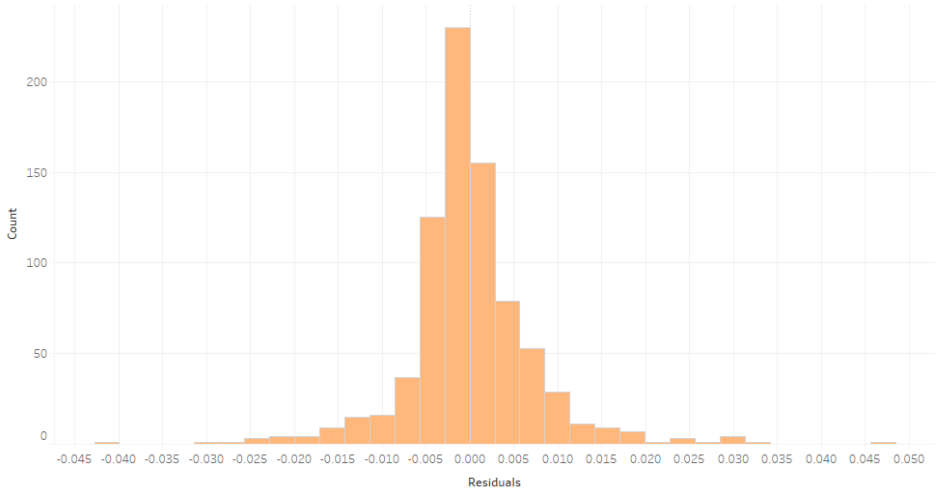

Figure 3.10: Residuals vs Fitted plot, dynamic model.

Variance Inflation Factor			
	GVIF	Df	$GVIF^{\frac{1}{2*Df}}$
GAS	1.1	1	1
DEP	1.1	1	1.1
CON	1	1	1
EMI	1.2	1	1.1
MAR	1	1	1
REN	1.1	1	1.1
GDP	1.1	1	1.1
EU	1	2	1

Table 3.15: VIF test.

cents change in price would have likely messed up the regression. *Figure 3.10* displays how the residuals of the regression are distributed.

Most of the residuals fall very close to zero: 708 observations out of 800 result in less than one cent of error. The distribution shows some long tails both on the left and right side. However, given the high number of observations, this should not constitute a problem. The *Variance Inflation Factor* can instead be used to check whether there is collinearity among the regression variables, as summarized by *Table 3.15*.

The GVIF and adjusted GVIF values are both very low, indicating that there is no hidden source of collinearity: the variables work independently from each other.

Figure 3.11 plots the Squared Root of the Absolute values of the residuals against the fitted values: the objective is to detect heteroscedasticity.

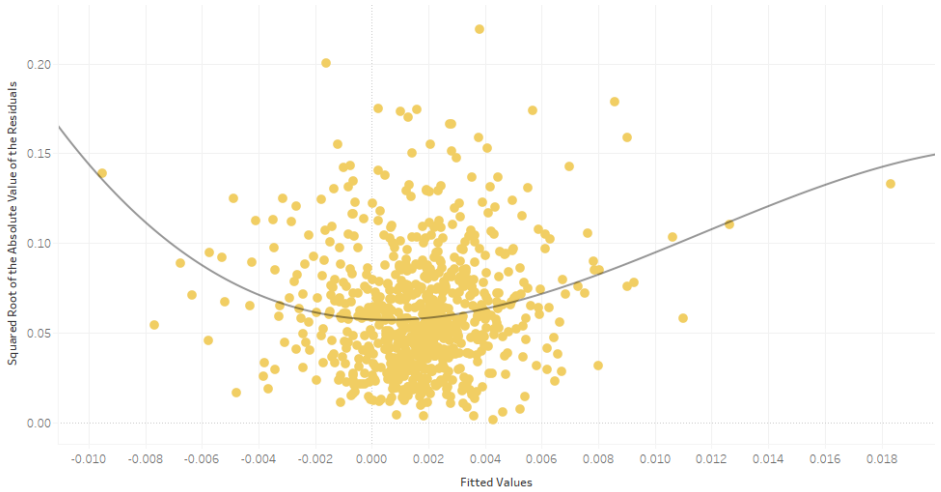

Figure 3.11: Scale-Location plot.

Dynamic Linear Model		
	Coef Sign	Significance
GAS	+	***
DEP	+	
CON	-	
EMI	+	
MAR	+	
REN	-	
GDP	+	**
EFTA	-	
EU	+	

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3.16: Coefficients and Significance.

When the change in price is significant, both in the positive and in the negative direction, the model is not really good in calling the right prediction. When instead the change in price is smaller, the residuals are on average small as well. This means that the model is likely to be quite conservative and that the heteroscedasticity hypothesis is not to be rejected. This is an indication of the fact that the linear model with ordinary least squares may not be the best choice in the dynamic setting. *Table 3.16* provides a summary of the linear regression in order to understand the impact of the predictor variables on the outcome. If the coefficient is positive, an increase in the predictor variable is going to result in an increase in the price; if instead the coefficient is negative, an increase in the predictor variable is going to result in a decrease in the price. The significance column is instead used to discriminate between variables whose coefficient is different from zero at the 95% significance level and variables whose coefficient does not.

Mean Absolute Error, in Euro Cents			
Linear Model	0.4867	With SubSel	0.5123

Table 3.17: Comparison of performance between Linear Model with and without Subset Selection.

Differently to the static case, only two variables are significant in the dynamic setting: GDP and GAS. The change in the electricity price from one term to the other is unlikely to be explained by changes in the market concentration, use of renewable sources or consumption patterns. It seems that the number of variables included in the model is too large, so the focus should go on a smaller subset. However, when tested on the validation set, the models with less variables turn out performing a little bit worse with respect to models with less variables, as showed in *Table 3.17*.

The model is wrong by less than half a cent when the prediction power of the model is tested on new data: this may seem as a good result, especially if compared to the accuracy of the static models. Yet, the reader should keep in mind that the outcome variable has changed with respect to the previous setting. When it is the whole electricity price to be predicted there is relatively high variability (the standard deviation of the outcome is of about 3.3 cents), while when it is only the change from one term to the other to be predicted there is low variability (the standard deviation of the outcome is of about 0.8 cents). Thus, in proportion, in the dynamic setting, a lower level of variability is captured.

3.2.2 Regression Trees

Since the linear regression models seem not to be really well-performing on this data, another option is to use regression trees. The idea behind this technique is to divide the predictor space into distinct regions $R_1, R_2, ..., R_J$. These regions have the shape of high-dimensional boxes. For each observation that falls into the region R_j, the same prediction, i.e. the average of the response values for the training observations in R_j, is output. In the case of the electricity price regression, the predictor space is created on the basis of the splits displayed of *Figure 3.12*. The information used to build this tree is not the complete set: only data from years prior to 2013 are used, while the rest is left to perform some parameter tuning.

The most discriminant splits have to do with the price of natural gas: if it goes down, then it is likely that the price of electricity will go down as well and conversely, if the former increases, the latter will also increase. This tree excludes the possibility that, if there's been a decrease in the gas price greater than 1.2 cents, the direction of the change in electricity price is going to be positive. To more precisely forecast the magnitudes of price changes, one other predictor comes in play in particular: the real Gross Domestic Product per capita. Especially in cases of significant economy

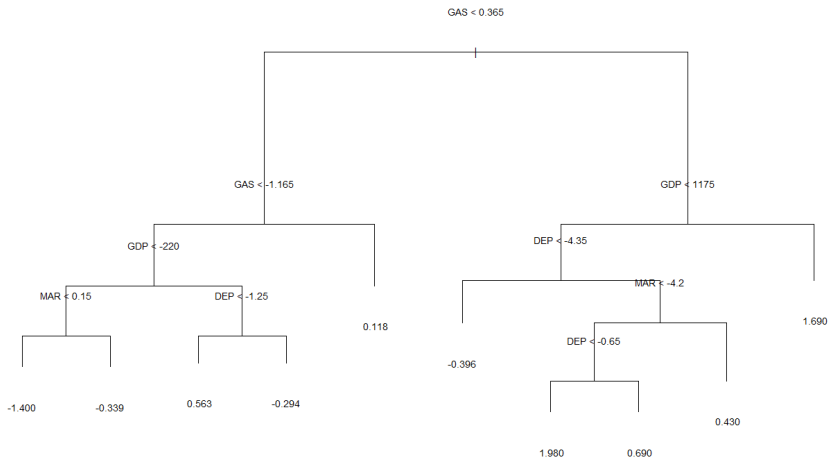

Figure 3.12: Regression Tree.

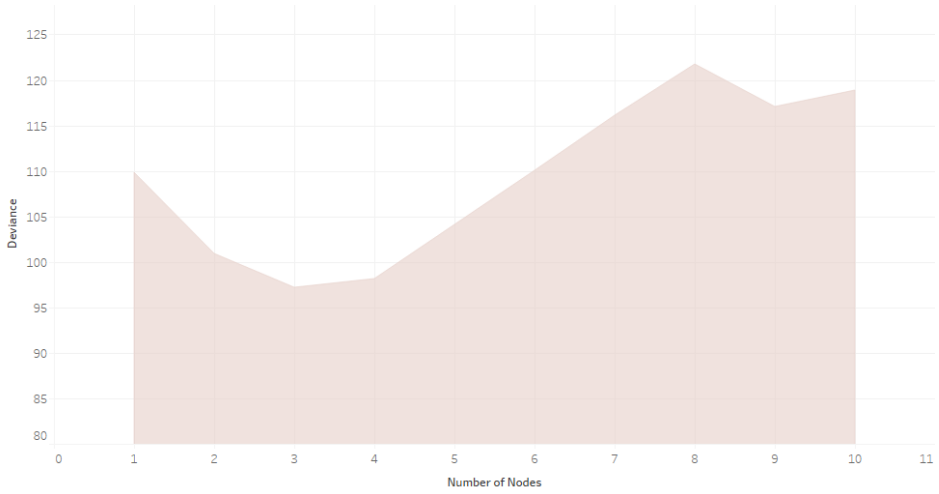

Figure 3.13: Tree Size and Relative Performance.

expansions (*GDP* increase greater than 1,175 euros), the electricity price increase is going to be particularly sizable. The two final variables that come in play are *MAR* and *DEP*, while there is no indication that the effect of changes in the share of renewable energy produced or in the consumption patterns do have any effect on the regression of the outcome variable.

The one displayed in *Figure 3.12* is a complete tree, but the risk with trees is to overfit. In order to avoid this, information not used to build the tree (data from 2012 to second half of 2014) is employed to compute the error and decide whether to reduce the number of splits in the tree or not. The total deviance of each tree in the cost-complexity pruning sequence is used as the metric to choose how many nodes to exclude. *Figure 3.13* displays performance on the data left for parameter tuning. The lower the deviance, the more effective the model.

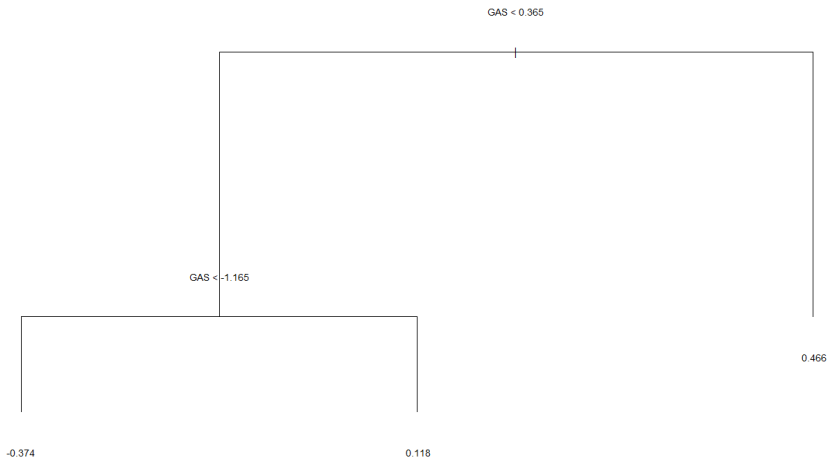

Figure 3.14: Tree after validation.

Mean Absolute Error, in Euro Cents			
Tree	0.564	Pruned Tree	0.511

Table 3.18: Comparison of performance between Trees before and after Pruning.

It seems that the complete 10-nodes model is not the best performing of the batch at all, as the deviance is lower when the number of splits is three. *Figure 3.14* shows the nodes and the shape of the tree that has been pruned.

The only relevant variable that has been kept in the model is the natural gas price: when the price for this good increase, the price of electricity goes up as well. There are instead two options when the price of natural gas has decreased since the last term: if the fall has been sizable (higher than 1 cent), then the outcome value is going to be negative almost by half a cent , otherwise the electricity price increase will correspond to 0.1 cents. *DEP*, *MAR* and surprisingly also *GDP* are excluded because their presence in the model is the result of overfitting.

It is now clear which the best performing tree is, but no indication yet regarding the performance on the validation set, which includes data from later years. *Table 3.18* summarizes performance of the regression tree in terms of Mean Absolute Error, as previously done for the other models.

The error is lower after pruning is performed with respect to the case the large tree is left intact. However, overall performance does not correspond to a real step forward with respect to linear models.

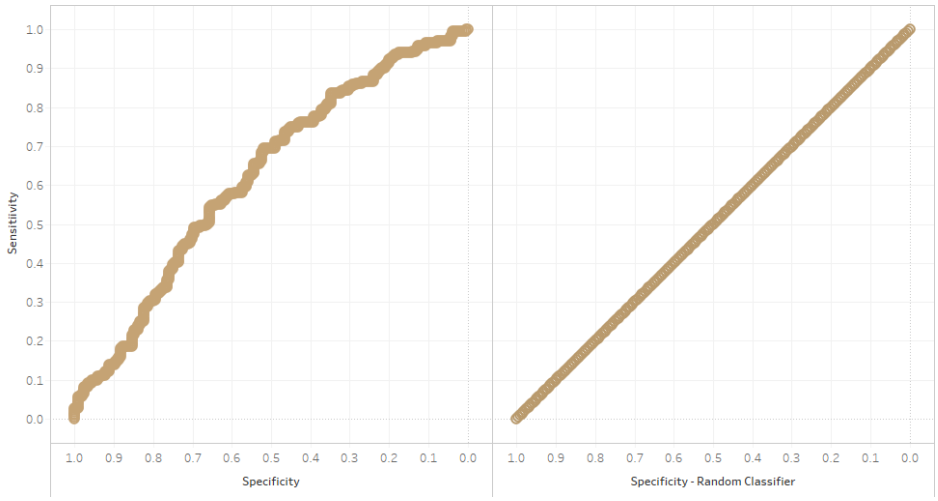

Figure 3.15: ROC curves.

3.2.3 Direction of Change

The magnitude of change from one term to the other seems to be more the result of chance, rather than something that can easily be predicted. If however the problem is simplified to the prediction of only the direction of the change, the hope is that the result may be more promising. The outcome variable this time is going to be the tally "up"/"down", in relation to the change in price. The algorithm run is a generalized linear model where the outcome is binomial.

Also in this case it is found that the most important predictor to explain the behavior of the outcome is *GAS*, while the other variables seem to not have a significant effect. The model shows performance superior to a random classifier, having the former an accuracy equivalent to 0.601 with respect to 0.5 of the latter. However, this does not come as a big improvement: *Figure 3.15* on the left displays the Receiver Operative Characteristic curve for the problem at hand, while the random classifier performance (on the specificity-sensitivity basis) instead is showed on the right.

The fact that the difference between the two plots is not so striking indicates that the classifier's ability to discriminate is not strong. The take-home message from the analysis of dynamic models is that most of the variability is likely to be explained by non-systematic factors, although the natural gas price fluctuations resemble the movements in the electricity price.

3.3 Global Models

For the time being, the idea has been to focus on data from European countries, and this was the case for various reasons. The EU provides accurate and reliable information with good frequency. Plus, for various aspects it is a common market so there are many similarities in the structure of their electricity systems. The reader may however wonder whether the findings that were discussed so far are only valid for European countries: excluding the rest of the world may be a source of bias if the interest is in looking at global interactions between variables. The role of this section is exactly the one of providing a benchmark to European models to check whether previous findings are globally valid.

3.3.1 Variables

Data for the global analysis is taken from *World Bank* instead of *Eurostat*. The type of regressor variables that are used aim to resemble as much as possible the ones from the previous regression problems. However, differences are unavoidable so variables will hereby be presented:

- The outcome variable is, of course, the electricity price (**PRI**), which this time is measured in U.S. cents per kWh (so not in Euros as before). For the records, a monthly electricity consumption is assumed, for which a bill for the equivalent of a warehouse is computed in the largest business city for the month of March.
- The net of energy imports as percentage of energy use (**DEP**).
- The Emission Intensity (**EMI**) variable is measured quite in a different way with respect to before. It is now the kg equivalent of CO2 emitted per 2010 US dollar of GDP.
- The electric power consumption (**CON**) is measured as kWh per capita.
- **GDP** per capita is again assessed with Purchasing Power Parity, which is a measurement of prices in different countries that uses the baskets of good as benchmark for evaluating purchasing power. The principle is the same of the real GDP used by *Eurostat*.
- Pump price for gasoline, used a proxy for oil price (**OILp**), same variable as for the European discussion.
- Electricity production from renewable sources (**REN**), excluding hydroelectric (as percentage of total).
- Electricity production from oil sources as percentage of total, (**OILs**), is a variable which was not available from Eurostat, which only disclosed information regarding share of energy produced from renewable sources.

Data from Solomon Islands					
Year	PRI	DEP	GDP	Region	Income
2015	98.7	-17.4	2236	East Asia & Pacific	Lower middle income
2016	99.9	-17.4	2272	East Asia & Pacific	Lower middle income
2017	96	-17.4	2338	East Asia & Pacific	Lower middle income
2018	67.9	-17.4	2421	East Asia & Pacific	Lower middle income
2019	67.4	-17.4	2466	East Asia & Pacific	Lower middle income

Table 3.19: Data from Solomon Islands.

- Another information data not available before that can be useful: electricity production from coal sources as percentage of total **(COA**).

The handling of information and creation of the data set is left to the appendix. One other important difference with respect to the European setting is that information for the price of electricity is only available from 2015 on, so this data set is going to be wider (a very large number of countries is included), shorter (time range only 5 years), and complete (since all countries' *PRI* information is disclosed for all the years). Since included information is also sourced from very small and remote places, there may be some inconsistencies: Mauritius is an island nation with a very low wealth level which is also an energy exporter (negative dependency value), but in spite of this electricity is extremely expensive, as displayed in *Table 3.19*.

From 2015 to 2017 one kWh was costing almost 1$, which is an extremely high price: the effect of Mauritius observations would be very high on the model. Since the purpose of this section is to build a model that works well for most countries in the world, and it is not really of interest to accurately fit the information from a 1 million people nation, Solomon Islands observations are simply excluded from the model.

3.3.2 Random Effects

The aim is to build a model for the prediction of the electricity price that is as global as possible, i.e. behaving well not for single countries or continents, but with observations of any provenance. From here the idea of using random effects, as if geographical information was not really of interest in this case. *Table 3.20* provides a summary of the random effect model. Variables enter the model linearly, without logarithmic transformations.

Interestingly all three *REN*, *COA*, and *OILs* have a positive coefficient. Although they describe opposite ways of producing energy, ranging from the most sustainable, *REN*, to the most polluting, *COA*, the greater the value for each variable, the greater the electricity is expected to be. Since the shares of energy produced obviously sum up to 100, this may indicate that using ways to produce electricity alternative to these

Coefficients			
	Estimate	Std. Error	p
(Intercept)	1.39	1.49	***
EMI	-5.17	1	***
DEP	0.00	0.00	
CON	-0.00	0.00	
COA	0.07	0.01	***
REN	0.13	0.03	***
GDP	0.00	0.00	***
OILp	4.89	1.12	***
OILs	0.11	0.02	***

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Global Performance			
R-Squared	0.2431	Adj R-Squared	0.2362

Table 3.20: Fixed Effects Model summary.

(e.g. generation through gas or nuclear energy) is likely to have a negative effect on the electricity price. Or, simply, that the random effects model is not the best option for performing this type of analysis.

3.3.3 Linear Model

As the case was for the previous models, using random effects may result in a model showing little variance but too much bias. Instead, including all the country identifiers would give the opposite result. A good compromise is to use the region, which is already present in the data set and can provide some information about which parts of the world tend to have higher electricity prices with respect to others. One decision that has to be taken when building the linear model with geographical identifier is whether to use the non-percentage variables in their logarithmic form or not. If from one side a small portion of model interpretability is lost, the gain is in the reduced influence of outliers or high-leverage points, and in the fact of the outcome price being output as always positive. *Figure 3.16* shows the difference in the residual distribution between models without (the one up) and with (the one below) logarithmic transformations. It is possible to evaluate how the second is more symmetrical, even though still a little bit skewed.

When the logarithmic transformation is applied, also the other assumptions beside residuals' normality, as for instance homoscedasticity seem to be respected, so the model of this type is preferred. The presence of world region dummy variables allows the model to have different intercepts depending on the country in consideration. *Ta-*

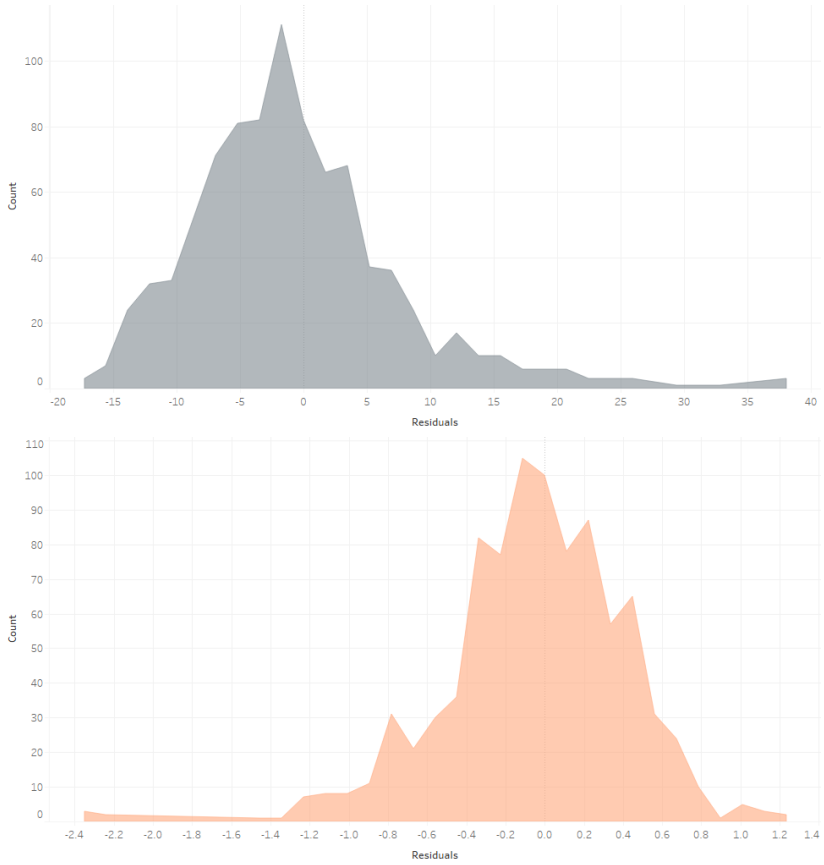

Figure 3.16: Residual distributions for models without and with logarithmic transformation applied on non-percentage variables.

ble 3.21 displays a summary of the model with region identifier. The model is trained not on all the available observations but only on years prior to 2019, whereas last year information is left for validating the results.

All region dummy variables' coefficients are negative, meaning that the base case (East Asia and Pacific) is the area in which this good is the most expensive. Conversely, the price is the lowest, on average, in Middle East and North Africa (shortened **MENA**), in Europe and Central Asia (**ECA**). There is not enough information to call that the intercept for North America (**NA**) price, is significantly different from the one of the base case. For Sub-Saharan Africa (**SSA**), South Asia (**SA**) and Latin America and Caribbean (**LAC**), the intercept is lower with respect to the base case.

The total variance captured by the model is about 30%, more than with the random effects model but still quite low. However, the wide heterogeneity of the data at hand has to be considered. Interestingly, the variables *OILs*, *COA* and *REN* still have a positive coefficient. The error on the validation set is once again measured in Mean Absolute Error (with the outcome variable re-converted with an exponential transformation) and is slightly lower than 5 dollar cents. The mean value for *PRI* is 17.44 cents.

Coefficients			
	Estimate	Std. Error	p
(Intercept)	17.1	1.65	***
EMI	-3.7	1.04	***
DEP	0.01	0.00	*
CON	-0.00	0.00	
COA	0.04	0.01	**
REN	0.12	0.03	***
GDP	0.00	0.00	***
OILp	6.8	1.12	***
OILs	0.1	0.02	***
ECA	-9.45	0.97	***
LAC	-3.47	1.11	***
MENA	-8.02	1.26	***
NA	-2.94	2.75	
SA	-5.59	1.51	***
SSA	-6.5	1.07	***

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Global Performance			
MAE	4.959	Adj R-Squared	0.3286

Table 3.21: Fixed Effects Model summary.

3.3.4 Regularization

The risk of leaning on linear regression is that the model fits the training data too closely, so not only the systematic component of the variance, but also the unsystematic one is fit. To avoid this behavior, called overfitting, instead of minimizing simple ordinary least squares, penalties are included in order to shrink the coefficients. In the static setting, the choice was between two alternative methods of regularization, Ridge regularization and Lasso, where the difference is in the form of the penalty term. In Ridge regression, the quantity to be minimized is:

$$L_{ridge}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j=1}^{m} \hat{\beta}_j^{\,2},$$

where the λ term is the most important parameter that decides for the amount of shrinkage of the coefficients. Inside the penalty term of the Lasso regression, instead of a square operation, the absolute value of the β_j is considered.

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j=1}^{m} |\hat{\beta}_j|$$

To understand whether it is better to use ridge or lasso penalties, the best solution is to look at performance on validation set. However, it is also possible that an hybrid

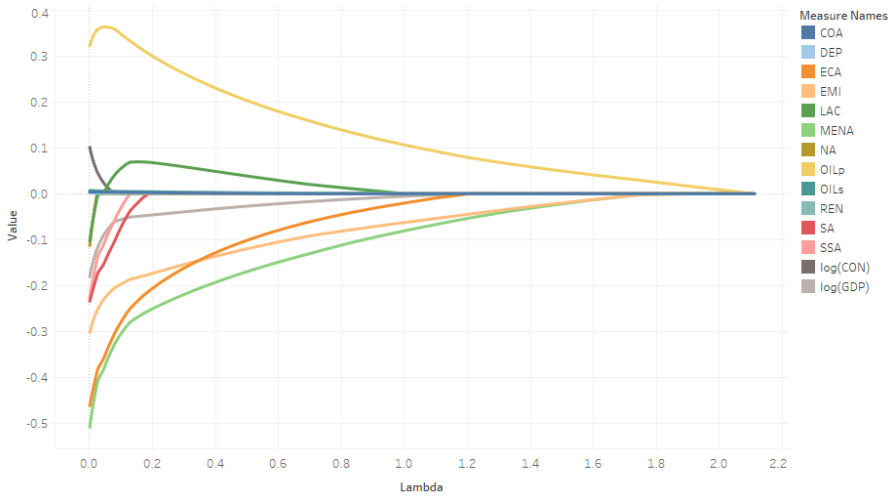

Figure 3.17: Coefficient Estimates for the Elastic Net regression.

approach between the two results in best accuracy. This is the idea of elastic net: introducing a parameter α that controls for how much of each regularization to use.

$$L_{elas}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j{}^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|)$$

The higher the α, the closer to lasso penalization, the lower the α, the closer to ridge regression. Best possible value has to be found not only for λ but also for α. To tune those parameters, cross validation is used. In the case of the electricity price regression, the value for the α parameter that results as best performing is 0.1, very close to straight ridge regression. *Figure 3.17* displays how the coefficients shrink in the regularization process with α equal to 0.1.

Despite the fact that this type of regularization is closer to L_1, so Lasso, than L_2, so Ridge, for an high enough λ value, all coefficients shrink to zero. The corresponding λ value, in the case of the electricity price regression is close to zero (==0.00278), and as a matter of performance there is not a great improvement with respect to the linear regression with Ordinary Least Squares (Mean Absolute Error is 4.962 cents).

3.3.5 Interactions

Similarly to what happens for the static European setting, increasing the number of predictors is more fruitful rather than decreasing it through regularization. Both adding polynomials and adding interactions strategies are tested and the latter action is associated to a better performance on the validation set. The number of parameters increases to 30 plus the intercept, 21 of which are interactions. The corresponding Mean Absolute Error is 4.522 cents. The fact that the relationship between the predictors and the outcome is best described through such a complex model may be a sign that

the validation set is not adequately large and there may be some overfitting. Since electricity price information is only collected from 2015, in the future it will be possible to have a clearer picture of the effect of each of the variables, and of the significance of the contributions.

Chapter 4

Conclusions

The purpose of this study is to find a good formula for predicting magnitude and direction of fluctuations in the price of electricity for household use. While traditional approaches focus on determining the effect of specific variables, most notably of market liberalization and investment on renewable sources for its generation, the focus here shifts to prediction accuracy. For this purpose, the econometric approach which is typical in this field is replaced with a Machine Learning attitude. The novelty with respect to previous approaches is also in the quantity of countries which are included in the study and in the time range, factors which try to make the results as much comprehensive as possible.

Different families of models are used in order to predict the electricity price and its change over time. The analysis starts with a focus on Europe: algorithms for trying to infer the amount paid by households in their bills given national data are developed. The initial approach is *static*: information regarding past electricity prices' information does not enter the model. The result is that, among the information at disposal, the Gross Domestic Product and the natural gas price data are the most meaningful, and both these variables are positively related with the outcome. The best performing model is a multiple linear regression in which the predictor variables are interactions between *GAS*, *GDP* and *REN*. On one hand, the predictor space is rather limited, since only two predictors are included, on the other those variables are not taken in their raw form but are first processed through some simple feature engineering. On average, the predictions of this model are wrong for about 1.7 euro cents, whereas the mean and standard deviation of the distribution are 10.7 and 3.4 cents, respectively.

The focus then shifts to foreseeing the magnitude of price change from one term to the other, given the fluctuations in the predictor space. European households experience changes in the bill amounts and a common question is what those price changes are due to. This is a more difficult problem with respect to predicting the electricity price from scratch, although the mean absolute error is of course smaller. Once again, the most meaningful information is related to *GAS* and *GDP*: different types of models

Model	Mean Absolute Error
Static Models	
Linear	1.8
Regularized	1.8
Interactions	1.7
Dynamic Models	
Linear	0.48
Tree	0.51
Global Models	
Linear	5
Regularized	5
Interactions	4.5

Table 4.1: Main models and corresponding performance on validation set.

are tested: multiple regression models and trees primarily. The accuracy of either of the two is not really satisfactory, but the former outperforms the latter by a small margin. The Mean Absolute Error of the multiple regression model is 0.48 cents, with the distribution of the outcome variable in this case having mean 0.11 and standard deviation 0.79.

Finally, to benchmark the findings, the last section is left to the discussion of a more general problem, the one of predicting the electricity price not only in Europe but in any world country. In this case, the natural gas price information is not available and this results in a decrease in the prediction precision. Linear models, both regularized and processed through subset selection are tested on this data set. The best performing model is a multiple regression model including several interaction terms, and this comes as a stunning difference with respect to previous settings, in which the best accuracy was obtained by using few predictors. Complexity wins over rigidity in this case. The Mean Absolute Error is 4.5, while the outcome distribution has a mean of 17.4 and a standard deviation of 9.8 cents. *Table 4.1* provides a summary of the models described in this study and its corresponding performance.

4.1 From Here

While *Eurostat* has been disclosing electricity price data for about 50 years now, *World Bank* has been doing the same for only 5 years. The hope is that in a decade or so it will be possible to have a much clearer picture of the effect of wealth, renewable investment, also on a global scale.

One of the shortcomings of this study is the focus on the demand side, rather than the supply: especially for countries which are net importers the price paid by the consumer

depends on decisions from the provider country. An idea for future work in this field would be to engage on this aspect as well.

Humankind is at the moment experiencing uncertain times, and it may seem presumptuous to build a compass for the future. However, data from the recent past can help us understand where we are heading to. Processes that are taking action right now, as the slow breaking up of electricity monopolies and the investment on renewable sources for electricity generation, seem to have marginal reverberations on the final price paid by consumers. The slowing down of the global economy as a consequence of the health crisis is likely to have a negative effect on the electricity bills, shrinking the amounts due by households.

Appendices

Data Cleaning

The library *Eurostat* is used to get the tables from the **R** interface, in particular the *get_eurostat* function. For instance, to obtain the price of electricity for years after 2007 the code is:

```
X <- get_eurostat("nrg_pc_204", time_format = "num")
X <- X[X$consom == "4161903" & X$tax == "X_TAX" & X$currency == "EUR",
c(6:8)]
X <- X[!(X$geo == "EA" | X$geo == "EU27_2020" | X$geo == "EU28"),]
```

To widen the data set, a series of merge functions is prompted. For instance, to add information regarding dependency from imports:

```
I <- get_eurostat("nrg_ind_id", time_format = "num")
I <- I[grep("TOTAL", I$siec), ]
I <- I[,c(3:5)]
names(I)[3] <- "DEP"
X <- merge(X, I, by=c("time", "geo"), all.x=TRUE)
```

As mentioned when introducing the renewable energy share variable, the major problem is in the transition from *Eurostat* to *World Bank* data for information prior to 2004. In this case the library *wbstats* is used, and in particular the function *wb*. There are small differences in the names assigned to the countries: in *Eurostat* Greece is **EL** while in the *World Bank* it is **GR**, United Kingdom is **UK** while in *World Bank* it is **GB**. Data integration between the two data sources is carried out using the following code:

```
J <- wb(indicator = "EG.FEC.RNEW.ZS", startdate = 2000,
  enddate = 2004, country=c(levels(X$geo), "GB", "GR"))
J$iso2c <- gsub("GB", "UK", J$iso2c)
J$iso2c <- gsub("GR", "EL", J$iso2c)
J <- J[,c(2,3,6)]
names(J) <- c("time", "WBren", "geo")
```

```
I <- merge(I, J, by=c("time", "geo"))
I$FAC <- I$REN / I$WBren
I <- I[,c(2,5)]
J <- merge(J, I, by="geo")
J$REN <- round(J$WBren * J$FAC, 3)
J <- J[J$time < 2004, c(1,2,5)]
U <- rbind(U, J)
X <- merge(X, U, by=c("time", "geo"), all.x=TRUE)
```

For a discrete number of countries, information regarding *market concentration* is sometimes missing. In the case of Bulgaria, for instance, all the values prior to 2013 are not available; those NAs are substituted with the first available year (2013).

```
X[X$geo == "BG" & X$time %in% c(2000:2012), "MAR"]
<- X[X$geo == "BG" & X$time == 2013, "MAR"]
```

While for Bulgaria it was use possible to simply copy and paste the information from following years, the Netherlands never disclosed any information regarding market concentration. In this case data is substituted with averages from countries with similar market structure, as Belgium, Germany, Luxembourg and Denmark.

```
for (i in 1:nrow(X)) {
  year <- X$time[i]
  if (X$geo[i] == "NL" & year%%1==0){
    DEvalue <- X[X$geo == "DE" & X$time == year, "MAR"]
    BEvalue <- X[X$geo == "BE" & X$time == year, "MAR"]
    LUvalue <- X[X$geo == "LU" & X$time == year, "MAR"]
    DKvalue <- X[X$geo == "DK" & X$time == year, "MAR"]
    NLvalue <- mean(c(DEvalue, BEvalue, LUvalue, DKvalue), na.rm=T)
    X[X$geo == "NL" & X$time == year, "MAR"] <- NLvalue
  }
}
```

Eurostat, at the time of the writing, has not yet published data regarding emission intensity for years 2019 and 2020. Since these values are likely to have changed in the last years, the assumption is that the direction of the change follows the trend of the difference from 2017 to 2018, with some inertia (equal to 0.33). The achieve this on **R**:

```
for (country in levels(X$geo)){
  if (country != "GE") {
  diff <- (X[X$time == 2018 & X$geo == country, "EMI"] -
```

```
X[X$time == 2017 & X$geo == country, "EMI"])
  X[X$time == 2019 & X$geo == country, "EMI"]
<- round(X[X$time == 2018 & X$geo == country, "EMI"] + diff / 3, 1)
  X[X$time == 2020 & X$geo == country, "EMI"]
<- round(X[X$time == 2018 & X$geo == country, "EMI"] + 2 * diff / 3, 1)
  }
}
```

Price of natural gas tends to increase over time. For this reason, when substituting Not Availables for the Gas variable when it is missing, both the history of the price for that country and the average of countries in that year is taken into account. For countries that never disclosed information regarding the natural gas price, averages from neighboring or similar countries is used. In the case of the missing value for Malta, for instance, an average of the information from Italy and Greece is used.

```
for (i in 1:nrow(X)) {
  year <- X$time[i]
  country <- X$geo[i]
  if (country %in% c("BE", "CZ", "DE", "EE", "EL", "LI", "LU", "MK",
  "PL", "PT")
& is.na(X[i, "GAS"])) {
      X[i, "GAS"] <- weighted.mean(c(mean(X[X$geo==country,"GAS"],
    na.rm=T),
mean(X[X$time==year,"GAS"],na.rm=T)),c(0.25,0.75))
  }
  else if (country == "MT") {
    value1 <- X[X$geo == "IT" & X$time == year, "GAS"]
    value2 <- X[X$geo == "EL" & X$time == year, "GAS"]
    value <- mean(c(value1, value2), na.rm=T)
    X[i,"GAS"] <- value
  }
}
```

For all other Not Availables referring to non mid-term records, for variables such as Dependence, Consumption, Inflation and Emission Intensity, both the average from the year and the average from the country is taken into consideration, using a weighted mean. This schema is implemented so that if information regarding previous years is missing, the NA can still be replaced with a reasonable value. To bring an example, the code for substituting Dependence missing values is reported below.

```
for (i in 1:nrow(X)) {
  year <- X$time[i]
  country <- X$geo[i]
```

```
  if ((is.na(X[i, "DEP"]) & year%%1==0)) {
      X[i, "DEP"] <- weighted.mean(c(mean(X[X$geo==country,"DEP"],
      na.rm=T),
mean(X[X$time==year,"DEP"],na.rm=T)),c(0.0005,0.9995),na.rm=T)
  }
```

For Gross Domestic Product missing values an additional intake of care is needed, since geographically close regions may present completely different wealth levels. In the case no historic value is found, countries that are expected to have a similar GDP are used as a benchmark. The problem with the above-mentioned technique is in the fact that using the average from historic values is likely to under-estimate the real GDP value, since for most years the economy tends to grow rather than shrink. For this reason the mean is positively adjusted by multiplying it by 1.02 (the European economy in recent years has been growing by about 2 percentage point per year on average).

```
historic1 <- c()
for (i in 1:nrow(X)) {
  year <- X$time[i]
  country <- X$geo[i]
  if (country == "XK" | country == "AL" | country == "ME" |
      country == "MK" | country == "BA") {
    value1 <- X[X$geo == "RO" & X$time == year, "GDP"]
    value2 <- X[X$geo == "HR" & X$time == year, "GDP"]
    value3 <- X[X$geo == "RS" & X$time == year, "GDP"]
    value4 <- X[X$geo == "BG" & X$time == year, "GDP"]
    outside <- mean(c(value1, value2, value3, value4),na.rm=T)
    value <- weighted.mean(c(outside,1.02*mean(historic1)),
    w=c(0.0005,0.9995),na.rm=T)
    historic1 <- c(historic1, value)
    X[i,"GDP"] <- value
  }
```

The remaining missing values all refer to mid-year records. When both the following semester value and the previous semester value are present, a simple average is computed.

```
for (i in 1:nrow(X)) {
  year <- X$time[i]
  country <- X$geo[i]
  if (year %% 1 != 0)
    for (col in c("DEP", "CON", "EMI", "MAR", "REN", "OIL")) {
      prev <- X[X$time==(year-0.5) & X$geo==country, col]
      after <- X[X$time==(year+0.5) & X$geo==country, col]
```

```
      X[i, col] <- mean(c(prev,after),na.rm=T)
    }
}
```

Membership to EU

To understand whether the membership to the European Union corresponds to a higher electricity price, left constant all other model parameters, it is necessary to first introduce some membership dummy variables. The **EA** parameter becomes one when the country belongs to the EU Economic Area, where Euro is used as a currency; for all other member countries the identifier is simply **EU**. Among the non-member countries, the ones that agreed to join the European Free Trade Association are identified as **EFTA**.

```
for (row in 1:nrow(DP)) {
  if (as.character(DP[row, "geo"]) %in% c("BG", "RO", "CZ", "HU",
  "HR","SE", "DK", "PL", "UK")) {
    DP$EU[row] <- "EU"
}
  else if (as.character(DP[row, "geo"]) %in% c("NO", "IS", "LI")) {
    DP$EU[row] <- "EFTA"
  }
  else if (as.character(DP[row, "geo"]) %in% c("BE", "EL", "LT", "PT",
  "ES", "LU", "FR","SI", "MT", "SK", "DE", "IT", "NL",
  "FI", "EE", "CY", "AT", "IE", "LV")){
    DP$EU[row] <- "EA"
    }
}
```

Going Dynamic

To make the new data set dynamic the idea is to subtract from one observation the value from the same country in the previous semester. To do so a dictionary is built through the function *hash*, a copy of the original data set is stored and then a for loop is triggered. Through this for loop, values inside the newly created data set are updated with the correct difference term by using the dictionary.

```
d <- hash()
toRemove <- c()
U <- X
```

```
X$geo <- as.character(X$geo)
for (i in 1:nrow(X)) {
  country <- X[i, "geo"]
  if (X[i, "geo"] %in% c(keys(d))) {
    U[i,c(-1,-2)] <- X[i,c(-1,-2)] - d[[country]]
  } else {toRemove <- c(toRemove, i)}
  d[[country]] <- c(X[i,c(-1,-2)])
}
X <- U[-toRemove,]
```

Global Data

Previous data was mostly obtained through *Eurostat*, which however provides information only for European countries. To build a more global data set, the *World Bank* works as the source. The R library to retrieve information is *WDI*. Cities and regions are removed from the data set, which includes only national information.

```
W <- WDI(indicator = c("IC.ELC.PRI.KH.DB1619", "EG.IMP.CONS.ZS",
"EN.ATM.CO2E.KD.GD", "EG.USE.ELEC.KH.PC", "NY.GDP.PCAP.PP.CD",
"EP.PMP.SGAS.CD", "EG.ELC.PETR.ZS","EG.ELC.RNWX.ZS", "EG.ELC.COAL.ZS"),
start = 2014, end = 2019, extra = TRUE)
names(W) <- c("iso2c", "country", "year", "PRI", "DEP", "EMI","CON",
"GDP", "OILp","OILs", "REN", "COA", "iso3c", "region", "capital",
"longitude", "latitude", "income", "lending")
W <- W[W$income != "Aggregates",]
W <- W[!is.na(W$income),]
```

The novelty with this type of data retriever is that it includes information regarding the region and the income level. This can be very useful when filling Not Availables. If a NA is found it is substituted with the latest value from the same country or if that is not available with the average from countries of the region in consideration. The same process is carried out to fill NAs for *OIL*, *REN* and *COA* columns.

```
for (row in 1:nrow(W)) {
  area = W$region[row]
  if (W$year[row] == 2014 & is.na(W$DEP[row])) {
    W$DEP[row] <- mean(W[W$region == area,"DEP"],na.rm=T)
  }
}

for (row in 1:nrow(W)) {
```

```
  country = W$country[row]
  if (is.na(W$DEP[row])) {
    if (is.na(W[W$country == country & W$year == 2015,"DEP"])) {
      W$DEP[row] <- W[W$country == country & W$year == 2014,"DEP"]
    }
    else {
      W$DEP[row] <- W[W$country == country & W$year == 2015,"DEP"]
    }
  }
}
```

A similar process with respect to the one described above is carried out for filling consumption and oil NAs. The only difference is that countries with the same wealth level are used instead of countries from the same region.

```
for (row in 1:nrow(W)) {
  area = W$income[row]
  if (W$year[row] == 2014 & is.na(W$CON[row])) {
    W$CON[row] <- mean(W[W$income == area,"CON"],na.rm=T)
  }
}

for (row in 1:nrow(W)) {
  country = W$country[row]
  if (is.na(W$CON[row])) {
    if (is.na(W[W$country == country & W$year == 2015,"CON"])) {
    W$CON[row] <- W[W$country == country & W$year == 2014,"CON"]
    }
    else {
      W$CON[row] <- W[W$country == country & W$year == 2015,"CON"]
    }
  }
}
```

Rows in which either *PRI* or *GDP* information is missing, are simply excluded from the data set.

```
W <- W[!is.na(W$PRI),]
W <- W[!is.na(W$GDP),]
```

Models

Static

To create a simple linear model, the command *lm* needs to be run in R. For instance, *EU Model* was computed by running the following.

```
LMeP <- lm(log(PRI) ~ log(GAS) + DEP + log(CON) + REN + log(GDP)
    + log(OIL) + MAR + EU, DP)
```

To compute analysis-of-variance tables for model objects produced by **lm**, the function *anova* is used. The objective is to compare models in order to evaluate differences in significance.

```
LM <- lm(log(PRI) ~ log(GAS) + DEP + log(CON) + REN + log(GDP)
     + log(OIL) + MAR, DP)
anova(LM, LMeP)
```

The library *caret* is instead used to carry out the **Box-Cox** transformation on the outcome variable. A new model is then built upon the new outcome and the previous predictor values.

```
bpt <- caret::BoxCoxTrans(DP$PRI)
DP <- cbind(DP, nP=predict(bpt, DP$PRI))
LMeB <- lm(nP ~ log(GAS) + DEP + log(CON) + REN + log(GDP)
+ log(OIL) + MAR + EU, DP)
```

Since applying the Box-Cox transformation to the outcome variable is not always sufficient to solve problems of non-constant variance, another option is to use **Weighted Least Regression** instead of Ordinary Least Regression. Steps are displayed below.

```
DP$resi <- LMeP$residuals
LMeR <- lm(log(resi^2) ~
 log(GAS) + DEP + log(CON) + REN + log(GDP) + log(OIL) + MAR + EU, DP)
DP$varFunc <- exp(LMeR$fitted.values)
LMwls <- lm(log(PRI) ~ log(GAS) + DEP + log(CON) + REN + log(GDP)
+ log(OIL) + MAR + EU, weights = 1/sqrt(varFunc), DP)
```

To perform subset selection, the library *leaps* is used. It is necessary to specify the option *nvmax*, which represents the maximum number of predictors to incorporate in the model. In the case of this research $nvmax = 5$, so the function will return up all the variables model, that is, it returns the best 1-variable model, the best 2-variables model, ..., the best 12-variables models. Backward and forward step-wise regression is instead faster and less computationally expensive because it works as a sequential system of inclusion (or exclusion) of variables, without having to compute the best subset for each N from 1 to $nvmax$. Code for performing exhaustive subset selection on R is reported below.

```
leaps::regsubsets(log(PRI) ~ log(GAS) + DEP + log(CON) + REN
+ log(GDP) + log(OIL) + MAR + EU, DP, nvmax = 12)
```

To split the data set in training and validation sets on a temporal basis (before and after 2015), it is sufficient to prompt the following:

```
train <- (DP$time < 2015)
test <- (!train)
```

Now the idea is to use the training set to build a model which can then be tested on the most recent validation data. To correctly determine the mean difference in Euro cents between the real information and the model prediction it is necessary to compute the exponential transformation on the model prediction, since in the model this variable is used in its logged form.

```
LMeP <- lm(log(PRI) ~ log(GAS) + DEP + log(CON) + REN + log(GDP)
+ log(OIL) + MAR + EU, DP, subset=train)
pred <- predict(LMeP, DP[test,])
mean(abs(DP$PRI[test] - exp(pred)))
```

If instead weighted least squares was used instead of OLS for building the model:

```
DP$resi[train] <- LMeP$residuals
```

```
LMeR <- lm(log(resi^2) ~ log(GAS) + DEP + log(CON) + REN + log(GDP)
+ log(OIL) + MAR + EU, DP, subset=train)

DP$varFunc[train] <- exp(LMeR$fitted.values)
LMgls <- lm(log(PRI) ~ log(GAS) + DEP + log(CON) + REN + log(GDP)
+ log(OIL) + MAR + EU, weights = 1/sqrt(varFunc), DP, subset=train)
pred <- predict(LMgls, DP[test,])
mean(abs(DP$PRI[test] - exp(pred)))
```

In the previous two models all variables except *EMI* were used, but it may be that reducing the number of variables actually boosts the performance on the validation set. To evaluate the mean absolute error between the predictions and the actual data at disposal, a vector with the errors is created in order to store the performance information. For every step N in the cycle 1-10 the best subset of variables including N of them is determined and then mean absolute error is computed. In this example the weights are used in the runs but the code is very similar if OLS is preferred to WLS.

```
W <- 1/sqrt(DP$varFunc[train])
tLM <- regsubsets(log(PRI) ~ log(GAS) + DEP + log(CON) + REN + log(GDP)
+ log(OIL) + MAR + EU, DP[train,], weights=W, nvmax = 11)
test.mat <- model.matrix(log(PRI) ~ log(GAS) + DEP + log(CON) + REN
+ log(GDP) + log(OIL) + MAR + EU, data=DP[test,])
val.errors <- rep(NA, 10)
for (i in 1:10){
  coefi <- coef(tLM, id=i)
  pred <- test.mat[,names(coefi)] \%*\% coefi
  val.errors[i] <- mean(abs(DP$PRI[test] - exp(pred)))
}
which.min(val.errors)
val.errors[which.min(val.errors)]
coef(tLM, 4)
val.errors
```

The library *glmnet* can be used in order to create ridge, lasso and elastic net models. In particular, the function *cv.glmnet* enables to choose the parameter λ through cross validation. A seed is set simply to make the results reproducible.

```
x <- model.matrix(log(PRI) ~ log(GAS) + DEP + log(CON) + REN + log(GDP)
  + log(OIL) + MAR + EU, DP)[,-1]
y <- DP$PRI
set.seed(35)
ridge <- cv.glmnet(x[train,], log(y[train]), alpha=0, standardize=T)
bestlam <- min(ridge$lambda)
```

Now that information regarding the λ value which enables to achieve best performance (on the training set), Mean Absolute Error is computed on the validation set.

```
ridge.pred <- predict(ridge, s=bestlam, newx=x[test,])
mean(abs(exp(ridge.pred) - y[test]))
```

For the visualization of *Figure 3.6* the output of the following code is used:

```
mat <- as.matrix(ridge$glmnet.fit$beta)
forplot <- data.frame(t(rbind(mat, ridge$glmnet.fit$lambda)))
```

If L_1 is preferred to L_2 penalty the functions to be used do not really change: the only major difference is in the α value to be set inside the *cv.glmnet* function.

```
lasso.mod <- glmnet(x[train,], log(y[train]), alpha=1)
set.seed(35)
cv.out <- cv.glmnet(x[train,], sqrt(y[train]), alpha=1)
bestlam <- cv.out$lambda.min
pred <- predict(lasso.mod, s=bestlam, newx=x[test,])
C <- as.vector(coef(cv.out, bestlam))
length(C[C != 0])
mat <- as.matrix(cv.out$glmnet.fit$beta)
forplot <- data.frame(t(rbind(mat, cv.out$glmnet.fit$lambda)))
```

To include interaction terms it is only necessary to create a *regsubsets* object that builds models not only looking at the distinct variables but also takes in the multiplied terms.

```
tLM <- regsubsets(log(PRI) ~ (log(GAS) + DEP + log(CON) + REN + log(GDP)
 + log(OIL) + MAR)^2 + EU, DP[train,], nvmax = 30, really.big = T)
test.mat <- model.matrix(log(PRI) ~ (log(GAS) + DEP + log(CON) + REN
+ log(GDP) + log(OIL) + MAR)^2 + EU, data=DP[test,])
val.errors <- rep(NA, 30)
for (i in 1:30){
  coefi <- coef(tLM, id=i)
  pred <- test.mat[,names(coefi)] %*% coefi
  val.errors[i] <- mean(abs(DP$PRI[test] - exp(pred)))
}
```

Until now predictor variables have only entered the model linearly but it may actually be that also including polynomials does improve performance. To do so and then compare their performance through subset selection the function *regsubsets* is again used.

```
tLM <- regsubsets(log(PRI) ~ poly(log(GAS),2) + poly(DEP,2) +
  poly(log(CON),2) + poly(REN,2) + poly(log(GDP),2) + poly(log(OIL),2)
  + poly(MAR,2) + EU, DP[train,], nvmax = 18)
test.mat <- model.matrix(log(PRI) ~ poly(log(GAS),2) + poly(DEP,2)
 + poly(log(CON),2) + poly(REN,2) + poly(log(GDP),2) + poly(log(OIL),2)
 + poly(MAR,2) + EU, data=DP[test,])
val.errors <- rep(NA, 17)
for (i in 1:17){
  coefi <- coef(tLM, id=i)
  pred <- test.mat[,names(coefi)] %*% coefi
  val.errors[i] <- mean(abs(DP$PRI[test] - exp(pred)))
}
```

Dynamic

The steps for building linear models and for applying subset selection on the variables are similar to the ones already described in the past section, for the static setting. What is new is the inclusion of regression trees and logistic regression. The library *tree* is used for this purpose. The training set is further divided into to smaller sets, one of which is used to tune the node quantity parameter, which controls for overfitting.

```
treeset <- C[train,]
val <- (treeset$time >= 2012)
traini <- !(val)
tre <- tree(PRI ~ GAS + DEP + CON + EMI + MAR + REN + GDP + EU,
  treeset, subset=traini)
plot(tre)
text(tre,cex=0.85,digits=3)
```

Some information data is left to decide whether to prune the tree in order to improve performance. The function *prune.tree* does the job.

```
valset <- C[val,]
my.tree <- prune.tree(tre,newdata=valset)
```

To choose the final tree model among the many, deviance is used as a metric.

```
opt.trees <- which(my.tree$dev == min(my.tree$dev))
best.leaves <- min(my.tree$size[opt.trees])
my.tree.pruned <- prune.tree(tre,best=best.leaves)
```

Function *glm* is used to build generalized linear models of the binomial family. If the probability computed by the model is greater than 0.5 the model is going to output a positive prediction (which in this case is going to correspond to a "up" prediction. The confusion matrix of the classification problem can be obtained using the *table* command.

```
C$increase <- ifelse(C$PRI > 0, "Up", "Down")
C$increase <- as.factor(C$increase)
lR <- glm(increase ~ GAS + DEP + CON + GDP + REN + EMI + MAR + EU, C,
 family=binomial, subset=train)
pred <- predict(lR, C, type="response")
glm.pred=rep("Down",434)
glm.pred[pred[test] >.5]="Up"
table(prediction=glm.pred , groundtruth=C$increase[test])
mean(glm.pred == C$increase[test], na.rm=TRUE)
```

The Receiver Operating Characteristic curve is used to compare performances when the 0.5 probability threshold is moved. To compute the *ROC* curve the library *pROC* is used.

```
C$prob <- pred
library(pROC)
ROC <- roc(increase ~ prob, C[test,])
plot(ROC)
```

Global

The most significant difference of the global models with respect to European ones is the use of cross-validation in order to find the best values of both the parameters α and λ. For this purpose, the library *caret*, and in particular the function *Train* is used.

```
Elastic <- train(
  log(PRI) ~ EMI + DEP + log(CON) + COA + REN + log(GDP) + OILp + OILs
  + region, data = D[train,], method = "glmnet",
  trControl = trainControl("cv", number = 10),
  tuneLength = 10
)
pred <- predict(Elastic$finalModel, s=Elastic$bestTune$lambda,
  newx=x[test,])
```

Bibliography

[1] Gdp per capita. `https://data.worldbank.org/indicator/NY.GDP.PCAP.CD`.

[2] Demand Forecasting for Electricity. *Body of Knowledge on Infrastructure Regulation*, 2017.

[3] Laurence Ball et al. What causes inflation? *Business Review*, (Mar):3–12, 1993.

[4] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

[5] Václav Brož and Evžen Kočenda. Dynamics and factors of inflation convergence in the European union. *Journal of International Money and Finance*, 86:93–111, 2018.

[6] Zsuzsanna Csereklyei, M d Mar Rubio-Varas, and David I Stern. Energy and economic growth: the stylized facts. *The Energy Journal*, 37(2), 2016.

[7] Belén del Río, Ana Fernández-Sainz, and Itziar Martinez de Alegria. Industrial electricity prices in the European Union following restructuring: A comparative panel-data analysis. *Utilities policy*, 60:100956, 2019.

[8] Fabio Domanico. Concentration in the European electricity industry: The internal market as solution? *Energy Policy*, 35(10):5064–5076, 2007.

[9] Carlo V Fiorio, Massimo Florio, et al. The reform of network industries, privatization and consumers' welfare: Evidence from the EU15. *UNIMI-Research Papers in Economics, Business, and Statistics*, 1088, 2009.

[10] Georg Fuchs, Benedikt Lunz, Matthias Leuthold, and Dirk Uwe Sauer. Technology overview on electricity storage. *ISEA, Aachen, Juni*, page 26, 2012.

[11] Jean-Michel Glachant and Sophia Ruester. The EU internal electricity market: Done forever? *Utilities Policy*, 31:221–228, 2014.

[12] Marie Hyland. Restructuring European electricity markets–a panel data analysis. *Utilities Policy*, 38:33–42, 2016.

[13] Tooraj Jamasb and Michael Pollitt. Electricity market reform in the European Union: review of progress toward liberalization & integration. *The Energy Journal*, 26(Special Issue), 2005.

[14] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[15] Faisal Jamil and Eatzaz Ahmad. The relationship between electricity consumption, electricity prices and GDP in Pakistan. *Energy policy*, 38(10):6016–6025, 2010.

[16] D. S. Kirschen, G. Strbac, P. Cumperayot, and D. de Paiva Mendes. Factoring the elasticity of demand in electricity prices. *IEEE Transactions on Power Systems*, 15(2):612–617, 2000.

[17] Hooi Hooi Lean and Russell Smyth. Multivariate Granger causality between electricity generation, exports, prices and GDP in Malaysia. *Energy*, 35(9):3640–3648, 2010.

[18] Anil Markandya and Paul Wilkinson. Electricity generation and health. *The lancet*, 370(9591):979–990, 2007.

[19] Gustavo A Marrero. Greenhouse gases emissions, growth and the energy mix in Europe. *Energy Economics*, 32(6):1356–1363, 2010.

[20] Blanca Moreno, Ana J López, and María Teresa García-Álvarez. The electricity prices in the European Union. the role of renewable energies and regulatory electric market reforms. *Energy*, 48(1):307–313, 2012.

[21] Machiel Mulder and Bert Willems. The Dutch retail electricity market. *Energy policy*, 127:228–239, 2019.

[22] Hiroaki Nagayama. Electric power sector reform liberalization models and electric power prices in developing countries: An empirical analysis using international panel data. *Energy Economics*, 31(3):463–472, 2009.

[23] Marc Ringel. Liberalising European electricity markets: opportunities and risks for a sustainable power sector. *Renewable and Sustainable Energy Reviews*, 7(6):485–499, 2003.

[24] Neil J Salkind. *Encyclopedia of research design*, volume 1. Sage, 2010.

[25] International Energy Agency Staff. *Energy Balances of Non-oecd Countries 2009/Bilans Energetiques Des Pays Non-membres De Locde 2009*. Organization for Economic, 2009.

[26] Simone Tagliapietra et al. Beyond coal: facilitating the transition in Europe. Technical report, 2017.

[27] W Kip Viscusi, Joseph E Harrington Jr, and David EM Sappington. *Economics of regulation and antitrust*. MIT press, 2018.

[28] Rafał Weron. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International journal of forecasting*, 30(4):1030–1081, 2014.