EDA of Spanish electricity generation 2015-2018

Group 16: Time Series - Assignment 3 - EDA Report

Daniel Petterson (300467629), Rob Tomkies (300588353), and Daniel Wrench (300398222)

Submitted 31st August 2021

This report is accompanied by an online dashboard, developed in R Shiny. The app allows interactive visualisation and comparison of each time series, and is discussed in greater detail in the section Time series visualisation.

Background

We chose to study a time series dataset about electricity generation in Spain over a four-year period. This data originally comes from the "ENTSO-E Transparency Platform" dashboard [1]. ENTSO stands for *European Network of Transmission System Operators for Electricity*, and this website is the central collection and visualisation repository of electricity generation, transportation and consumption data across Europe. European Member States are required to submit this data to this platform. The data for Spain from the start of 2015 to the end of 2018 was collected from the dashboard and published on the data repository website Kaggle [2].

The raw dataset from Kaggle contained variables for electricity consumption, generation, and pricing, recorded every hour of the four years, as well as day-ahead forecasts. For this exploration of the data, we limited our investigation to electricity generation, which is divided into several different generation sources/methods.

The question we formulated to guide our EDA is as follows:

How does the amount of electricity generated by different sources change over time, individually and with respect to other sources?

This question is worth investigating due to the pressing global challenge of climate change, and the consequent need for countries to transition to a sustainable energy economy. In order for this to be possible, we need to understand how electricity is currently being generated and what the long-term trends are, to see whether a nation's energy policy should continue or diverge from the current trajectory. Furthermore, governments need to know how electricity generation sources vary seasonally and what interdependencies there are.

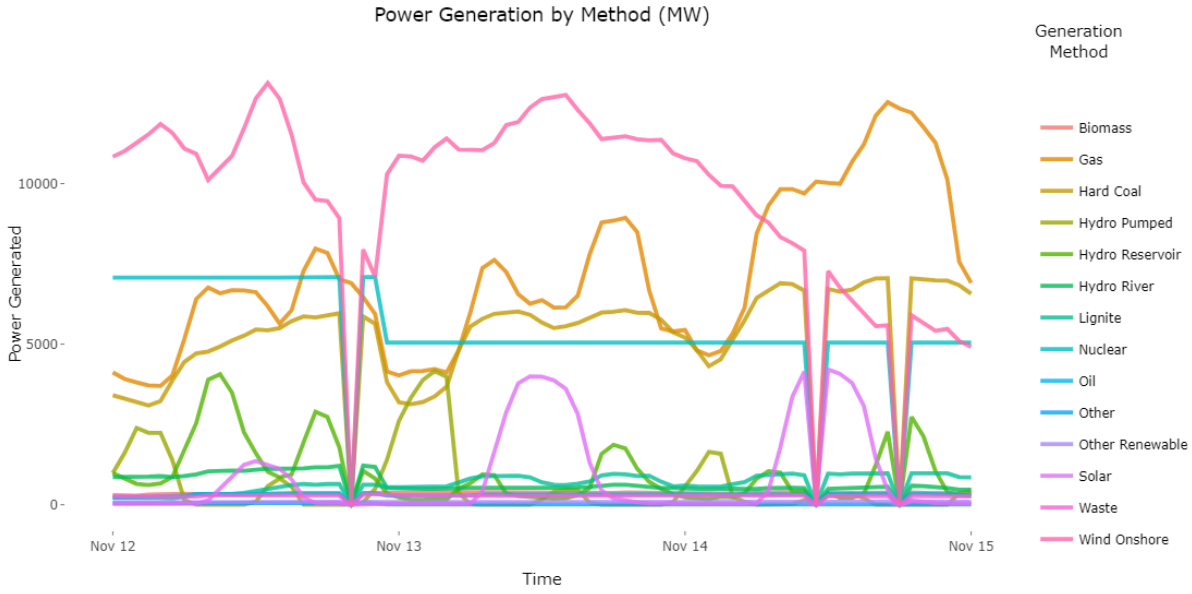
To learn more about this topic we conducted a literature review, and found that in 2009 Spain became subject to EU Directive 2009/28/EC which mandates an EU-wide commitment to expanding the use of renewable energy [3]. The target was for 20% of all energy consumed in the bloc to be generated by renewable sources by 2020. The Spanish government subsequently published the Plan de Energias Renovables 2011-2020 outlining how it planned to meet its commitments [4]. The amount of energy generated each year that our data covers does not significantly change, but the composition of its generation sources is changing in favor of renewable sources [5, 6].

Data

After removing the variables that were not relevant to our analysis, and those that contained no observations, the dataset comprised one date-time index variable and 14 numeric variables, corresponding to 14 different electricity generation sources (solar, nuclear, etc.). Each variable contained hourly recordings of the amount

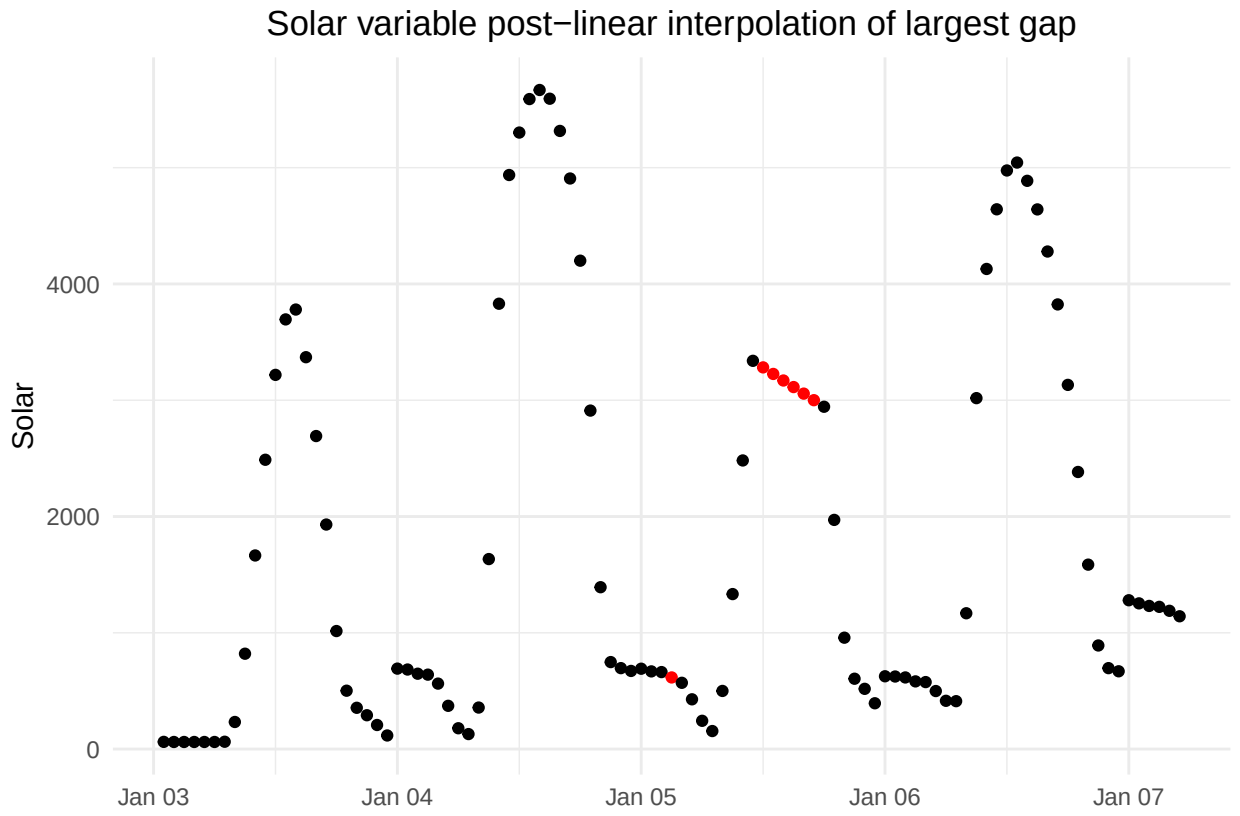
of power generated by that source, in megawatts (MW), in Spain, between January 1st 2015 and December 31st 2019. This equates to about 35,000 observations for each source.

As well as some missing values (discussed further below), we discovered some outliers in the data. Interestingly, these occurred across variables at the exact same timestamps: we identified 3 rows in the raw dataset in which all but one generation source recorded a value of 0. This is shown in the plot below of three days in November 2017, where three negative spikes to zero for all sources except Gas clearly stand out.

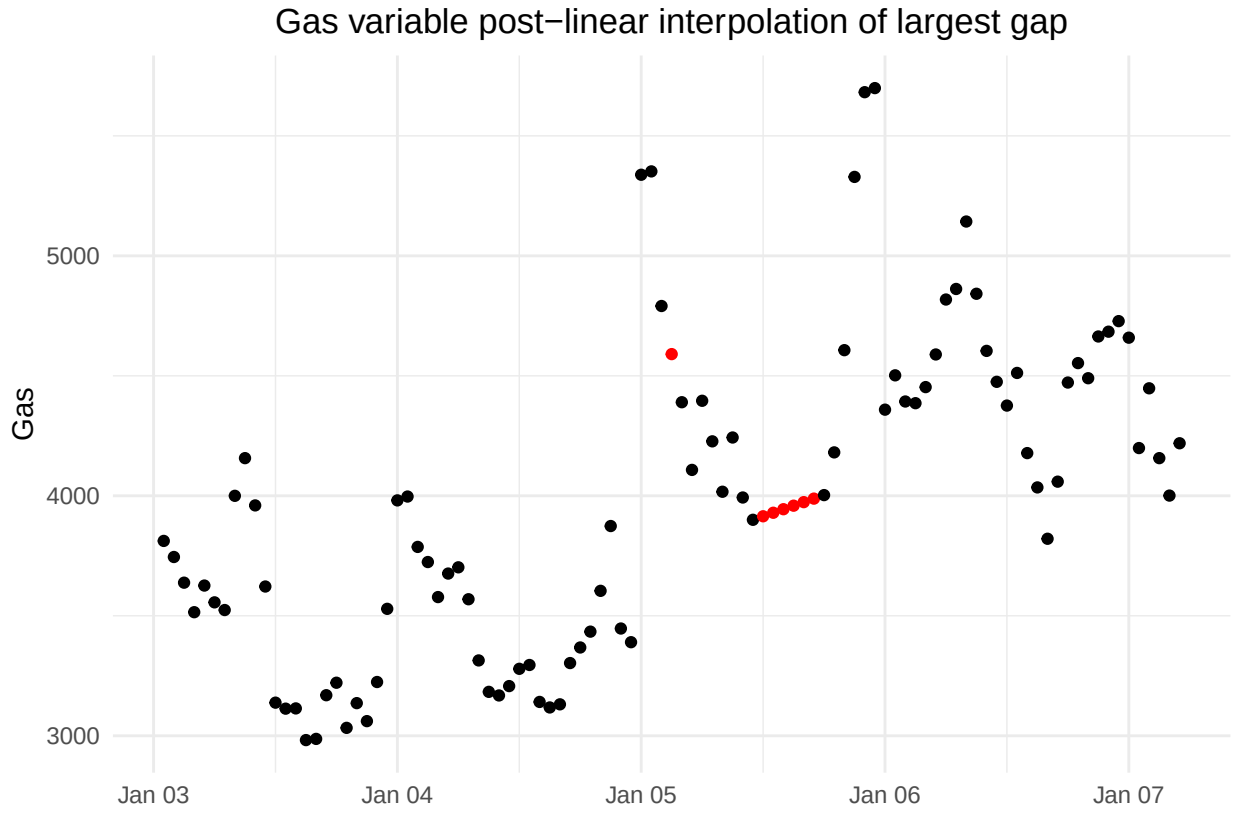


Because they seem very unlikely to be true values, these outliers were assumed to be incorrectly coded as non-missing and were removed. We also identified two other rows where values were a significant departure from the typical range of that source, but were not equal to zero. Because these values are less obviously mistakes they have not been removed.

After removing the outliers, there was a small amount of missing data (256 observations in total). Each variable had between 18 and 22 missing values, which occurred across a total of 26 rows/hours. Just like the outliers described above, most of these missing values occurred across variables: there were 21 rows where data for all or all but one variables were missing, and five rows where only one variable was missing. The missing data did not appear to have any structure, but it was more common towards the start of the time series. Because most of the missing values were isolated, we were confident in imputing them using linear interpolation. However, there was one area of slight concern: of the 26 rows with missing data, six of them were consecutive. The plots below show how this interpolation affected two of the variables.



The plot above, with interpolated values in red, clearly shows that linear interpolation over the adjacent missing values did not follow the strong seasonal pattern observed for this variable.



The linear interpolation over the adjacent missing values in the Gas variable looks as if it may also result in some distortion of the fluctuation in this variable, but it is less obvious than for Solar.

Overall, despite linear interpolation not following the trend of the data for some variables, this gap represents a very small proportion of each time series (less than 0.02% of the series), so we were content to make this small distortion of the data and apply the interpolation method consistently across each variable.

Ethics, Privacy, and Security

Ethics

While there is no intention to utilise this study for any purpose beyond academic, the ethical concerns relating to this project entirely come from possible uses and interpretations of the outputs of this study in the real world. Primarily, these would relate to political decisions of national energy policy and investment. Spain has an established nuclear sector which has had opposition since the 1950s when the government initially announced its intentions to develop the capability [7]. The arguments against nuclear largely focus on the hazardous waste, potential consequences of disaster, and association with weapons of mass destruction. If this paper were utilized to argue for further development of Spain's nuclear power industry, these concerns must be considered. Similarly, as mentioned in Background, there are growing concerns around hydrocarbon-based fuel sources and their impact on climate and air quality. This study specifically makes no recommendations, but if people decide to develop hydrocarbon-based energy sources, they should consider the ethics of their impact on the environment. Therefore, global warming is both a political and, in some ways, an ethical issue, and so this would also need to be considered if information from this study were used to support any decisions made about this global challenge.

There are ethical issues to consider beyond technology and climate change as well, in terms of the effect of energy sector policy on people. Any modelling using this data is at a national level, and one must remember this does not mean there is equal access to energy from each source across the nation. For example, if the government decided to shift to a non-diversified energy model, or banned local diesel generators, it is possible that certain areas, most likely rural, would have less access to electricity, having an adverse effect on the people who live in these areas. A more likely scenario is one in which, due to political pressure, the government chooses to invest in a more expensive source of energy in order to achieve a higher proportion of green energy [8]. Pushing power prices up would disproportionately affect people from a lower socioeconomic background, which could lead to greater poverty and a greater class divide in the country. This problem could be exacerbated if such investment comes with more automation, leading to greater efficiency at the cost of lower skilled jobs and again disproportionately affecting the lower socioeconomic class.

There is a small risk of creating feedback cycles relating to the above concerns. This would only occur if decisions made based on the results of this study created a negative affect as discussed above, which was then integrated into another model which the further exacerbated the issue in turn. As this study is a non-reactive one-off and there is no intention to base decision-making upon the models of this exercise, the concern of feedback loops has been considered outside the remit of the ethical concerns of this study.

Spain has many formally recognized ethnicities and minorities [9], however, the Spanish Personal Data Protection Act No. 15/1999 does not extend any explicit protection for or guidance on the use and storage of data from these groups. This means that handling approaches and considerations like the protective measures and regulations offered to people such as the Māori in New Zealand are not present or relevant in Spanish law [10]. Furthermore, because this data is recorded and stored at a national level from generation source, there is little about personal ethics and the use of personal data that is relevant or of concern to this study. There are no religious concerns with energy modelling beyond those discussed to do with environmental and social impacts.

Privacy

Privacy relates to your right not to be asked invasive questions or to needlessly provide personal information, as well as the assurance that any publications made from the data cannot be used to identify you. The data

from this study was collected at source of generation and then generalized to a national resolution. This means that there has never been any personal information stored within the data, and this would only be a concern if the data was compiled as the sum of individuals' energy use. A possible concern could arise if the data was collected from relatively small areas, such as suburb- or street-level. In this case it is possible that even with anonymised data, individuals could be identified using local knowledge. Again, this is not the case as information is provided at a national level. This means that data anonymisation techniques such as differential privacy are not only unnecessary, but detrimental to the accuracy of the data.

Considering now corporate privacy, because this data is available publicly from Spanish TSO Red Electric España, there is no concern of a corporate privacy breach. Further to this, the body holding the data is the national government body, which will have ensured that it complies with both Spanish and European data privacy regulations.

Security

Confidentiality, integrity, and availability are the three main factors that should be considered when ensuring the security of a data project. Confidentiality is ensured through the resolution of the data being at a national scale and being collected at the source, thus ensuring no identifiable information is available in the first place. This data is publicly available and so there is no need to ensure that this data is kept fully secure while handling it. If there were identifiable aspects to the original data, the responsibility to ensure this information is secure lies with the body holding the information and to ensure they are compliant with any national or European data laws. The analysis, organization and modelling were completed and stored using Microsoft Teams and a Git repository, which all require, at minimum, password protection or multi-factor authentication. While the information within each storage service does not need to be protected, each service has a responsibility to protect the personal information of our team.

Integrity was maintained for this project through the upkeep of a Git repository, ensuring that we had a complete record of changes to the code and report. This way the code was kept up-to-date while allowing us to keep records of previous versions in case we needed to roll back edits to the code. We made sure we keep an unedited version of the original dataset. Access to edit files was only allowed through team members accounts which were all password protected. Even if someone obtained access to the files and made unregulated changes, these would be recorded through Git, ensuring it would be simple to stay on top of our work.

Availability of data and files was ensured through utilizing cloud-based platforms for all data, codes, and reports. This ensured that all members of the team could access all files in their most up-to-date format, provided an internet connection was present. Each update to the formats would be associated with the respective team member account and so would be authorized within the platform. All platforms are mainstream and therefore utilize top-level protection. Should the servers go down due to something like a DDOS attack, each Git repository is stored locally on devices as well, so one of us would still have the most up-to-date file.

Exploratory Data Analysis

In the first section, Summary statistics, we look at statistics for each source over the entire time series, and calculate correlations between sources.

In the second section, Time series visualisation, we investigate some of these correlations, using charts to compare how different sources vary over time with respect to one another. Here we also examine interesting long-term trends and seasonality.

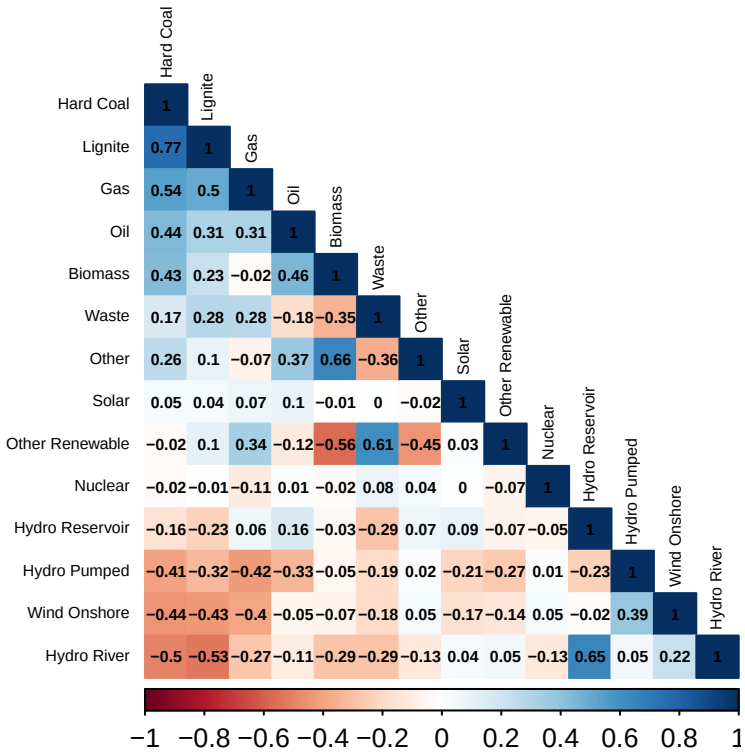
Summary statistics

	Mean	SD	CV	Min	Median	Max	Histogram
Nuclear	6264	838	13	998	6564	7117	
Gas	5623	2202	39	0	4970	20034	

	Mean	SD	CV	Min	Median	Max	Histogram
Wind Onshore	5466	3213	59	234	4850	17436	
Hard Coal	4257	1962	46	576	4475	8359	
Hydro Reservoir	2606	1835	70	134	2165	9728	
Solar	1433	1680	117	2	616	5792	
Hydro River	972	401	41	283	906	2000	
Lignite	448	355	79	0	509	999	
Hydro Pumped	476	792	167	0	68	4523	
Biomass	384	85	22	0	367	592	
Oil	298	52	18	44	300	449	
Waste	269	50	19	39	279	357	
Other Renewable	86	14	16	4	88	119	
Other	60	20	34	0	57	106	

The table above shows summary statistics (to 0dp) for every generation source in the dataset, ordered from largest to smallest mean hourly value. This allows a quick comparison of the magnitude and variability of each source. Of the top five sources, two are fossil-fuel based (Gas and Hard Coal). Nuclear is a unique source in that it has the largest average hourly generation but a relatively low standard deviation, compared to other major sources. This combination of high mean and low variability is best summarised by the coefficient of variation (CV) statistic, calculated as the standard deviation divided by the mean and then multiplied by 100. Nuclear has the lowest CV out of all sources. Wind Onshore has the largest variability in hourly output, followed by Gas. Solar, however, is a source with very strong seasonality, and this is shown by its higher CV than both Wind Onshore and Gas.

As discussed in the Background section, understanding interdependencies between electricity generation sources is important to maintaining energy availability in the short- and medium-term, but also to manage energy transition in the long-term. We calculated Pearson correlation coefficients as a way of investigating linear associations in the dataset. (NB: Due to the autocorrelation properties of time series data, these should be interpreted with caution, but can still give insights into associations.)



Nuclear, the largest overall contributor to electricity generation, does not have any significant correlations (here defined as correlations with absolute value greater than 0.3). Therefore, not only is Nuclear an incredibly steady generation source, it is also one that does not obviously depend on or affect the amount of electricity produced by other sources. Solar, on the other hand, is a much more variable source, but also has no significant correlations, showing it is also a very independent generation method.

Three generation sources that are very *dependent*, or closely related to other sources, are Hard Coal, Hydro River, and Other Renewable, shown by their large number of significant correlations on the correlation plot. The strongest linear association in the dataset is a positive correlation of 0.77 between Hard Coal and Lignite. Lignite is another type of coal, so it makes sense that these two sources would change in tandem. Other correlations are investigated further in the discussion of the time series plots in the next section.

Time series visualisation

The R Shiny dashboard (https://danielpetterson.shinyapps.io/scripts/) accompanies this section of the EDA. The dashboard allows interactive exploration of each electricity generation source over time, with customisation of which time period to look at, how to average the data, and how to compare the variables. Here we focus on the most interesting patterns.

The plot below shows the five sources with the greatest long-term change over the four-year period. (These five sources also happen to have the lowest mean hourly values of the 14 studied in this dataset.)

Biomass, Oil, and Other generation sources all show a rapid decrease from a high in 2015 to a low in early 2016, followed by a relatively constant long-term trend for the rest of the series. On the other hand, Waste and Other renewable sources show a sustained increase from 2015 to early 2017, and again followed by a relatively constant long-term trend for the remainder of the series. The similar or opposite nature of the trends between these generation sources are reflected in strong positive or negative correlations. For example, Biomass and Other sources have the second strongest positive linear association of any two sources in the

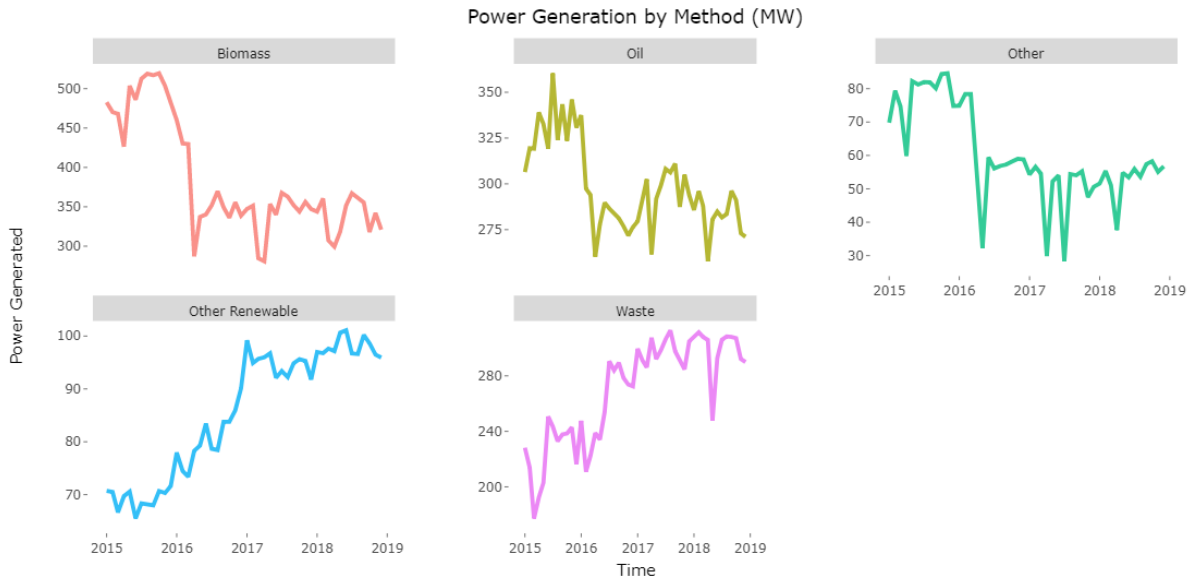
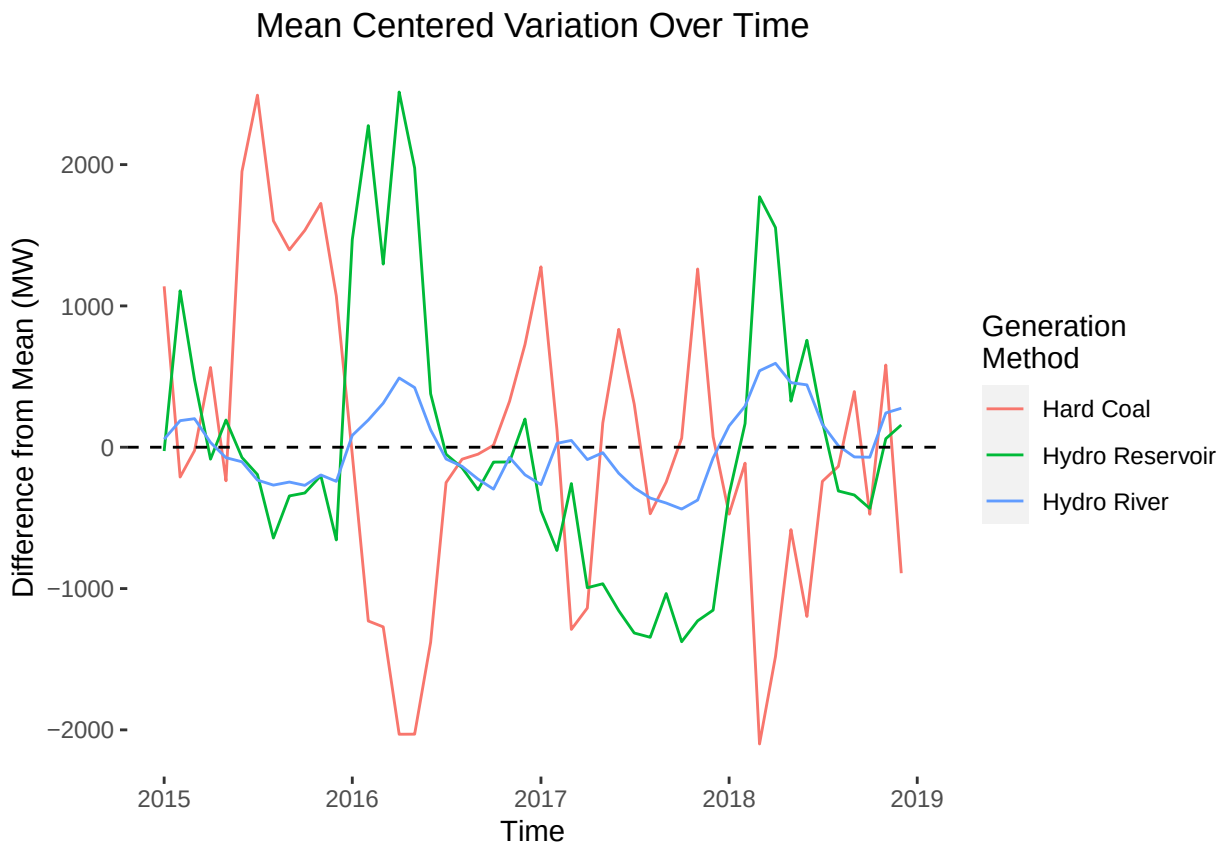


Figure 1: **Monthly averages of Biomass, Oil, Other, Other Renewable, and Waste**

data, with a correlation of 0.66. Meanwhile, the opposing trends mean that Biomass and Other Renewable sources have the strongest negative linear association of any two sources, with a correlation of -0.56.

Unlike with Hard Coal and Lignite, or Hydro River and Hydro Reservoir, it is not immediately obvious why these sources follow similar or opposite trends. Also, due to their low contribution to overall electricity generation, these marked long-term trends do not have a large effect on total energy output.

Hard Coal and Hydro River, as shown in the correlation plot, are two sources with several large correlations with other sources. The plot below shows some examples of this interdependence, by plotting the mean-centered variation over time of Hard Coal, Hydro River, and Hydro Reservoir.



Due to the way the peaks and troughs of each source tend to either be in phase or 180° out of phase, we clearly observe a close positive correlation between the two Hydro sources (correlation = 0.65), and a negative correlation between Hard Coal and both these sources (correlation with Hydro River = -0.5, with Hydro Reservoir = -0.16). The reason behind these relationships is likely to be that a more reliable energy source such as coal is used to generate electricity when renewable (and therefore more desired) sources such as hydro power have lower output than is needed.

The most obvious seasonality we detected was for Solar generation. In fact, Solar displayed seasonality on two different scales. In the following plot below we can see that this source is seasonal on a daily scale, peaking at around 4000-5000MW/hour between 12pm and 2pm and reaching a low of 40-600MW/hour between 11pm and 6pm. This is an intuitive periodicity: solar plants produce most electricity when they are receiving the most sunlight around noon, and much less during the night when it is dark.

In the next plot we can see that Solar is also seasonal on a monthly scale. Electricity generation by solar peaks at around 2000MW in June or July each year, and hits a low of between 800MW and 1000MW in November or December each year. Once again, this periodicity makes intuitive sense, with higher generation in the summer months and lower generation in the winter months.

From background research we learnt that although Spain was an early pioneer in renewable energy, changes in government policy, including the reduction of financial incentives, have stymied the growth of renewable energy capacity in the country. This is especially true for solar power plants, which have remained at a similar generation capacity between 2011 and 2017 [11].

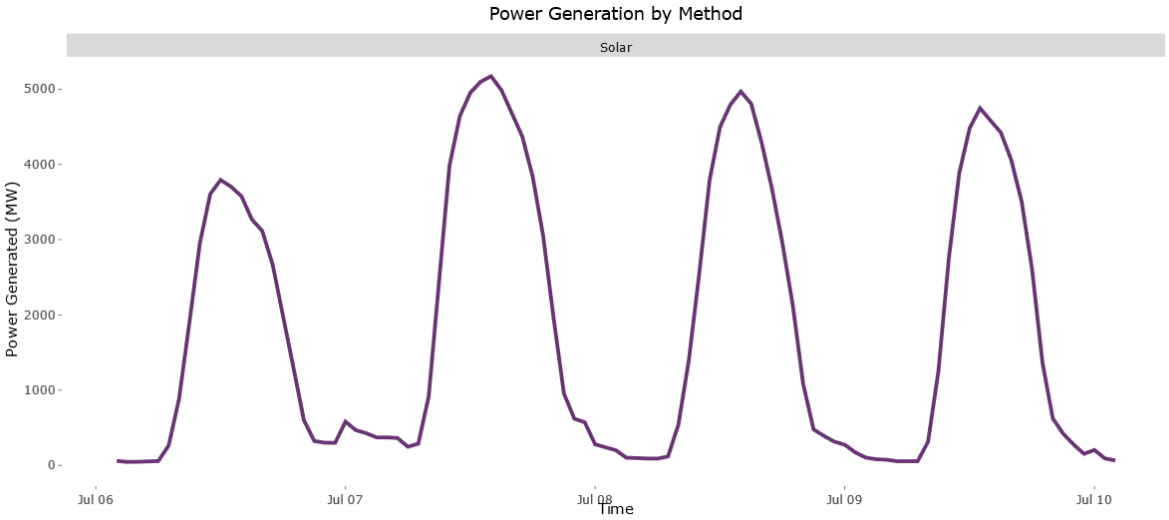


Figure 2: **Hourly averages of Solar (subset from July 2018)**

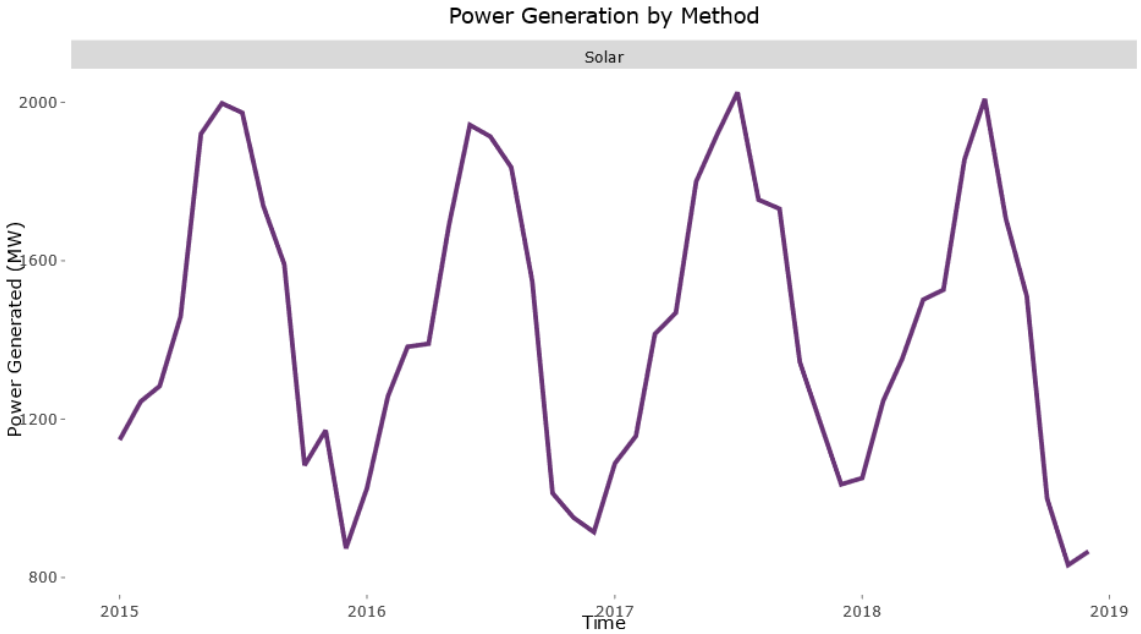
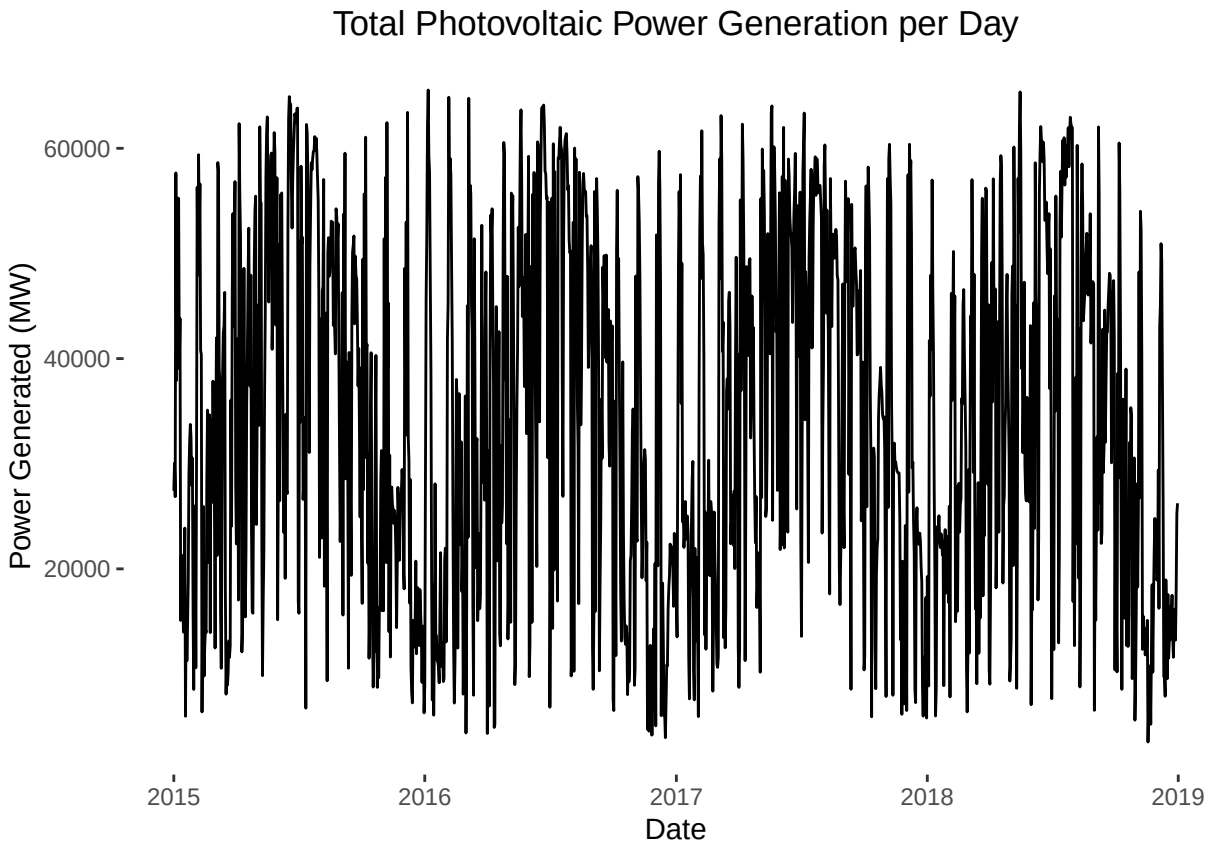


Figure 3: **Monthly averages of Solar**



From the above plot we can see that total daily power generated from photovoltaic solar systems shows little change across the years we investigated. The similarity of high peaks in winter and summer is likely due to almost all of the solar generation capacity in Spain being photovoltaic rather than thermal: while there are less sunlight hours in winter, the amount and intensity of sun would be similar in the middle of the day. In fact, solar panels are less efficient in the high temperatures of Spanish summers, but no clear evidence for this was found in the data.

While other seasonal effects are less clear, by deconstructing the signal we could find overarching trends, variance, and some degree of seasonality in most of the data. The signal decomposition for our natural hydro sources is shown below. A clear, yearly seasonal cycle is shown, likely demonstrating the existence of a wet and dry season in Spain. On top of this a repeated monthly spike and drop is seen, likely due to some form of maintenance cycle. By overlaying the smoothed seasonal variation with Solar (plot title "Seasonal Contribution"), we can see that they are roughly anti-phased. While this will help to provide a constant year-round supply, the magnitude difference suggests that there will be some residual seasonal signal present in the other generation sources, given the assumption of constant year-round generation of electricity. This is seen in further signal decomposition, and would help to explain the strong negative correlation between hydro sources and many hydrocarbon energy sources.

Proportion of fossil fuel usage over time Overall we see a number of trends, further to the seasonality deficit, that underlie the fossil fuel usage over time. As shown below, once seasonality and variance are removed, all display a trend of decreasing to mid 2016, then increasing to mid 2017 before decreasing again While the inflection points are not identical, they are extremely similar, which can likely be partially explained by common political factors. The first inflection point centres around June 2016, the time of a Spanish general election with issues centering around green energy and environmental friendliness. Articles talk about the drive at the time to support renewable energy and the uncertainty in the fossil fuel sector, which lead to a diversification away from them [12]. Following this Spain had an exceptionally dry year [13], leading to the much lower seasonal peak seen in hydro, and we see a steady increase in fossil fuel sources to compensate. In early 2017, the Spanish government implemented the energy efficiency action plan [14],

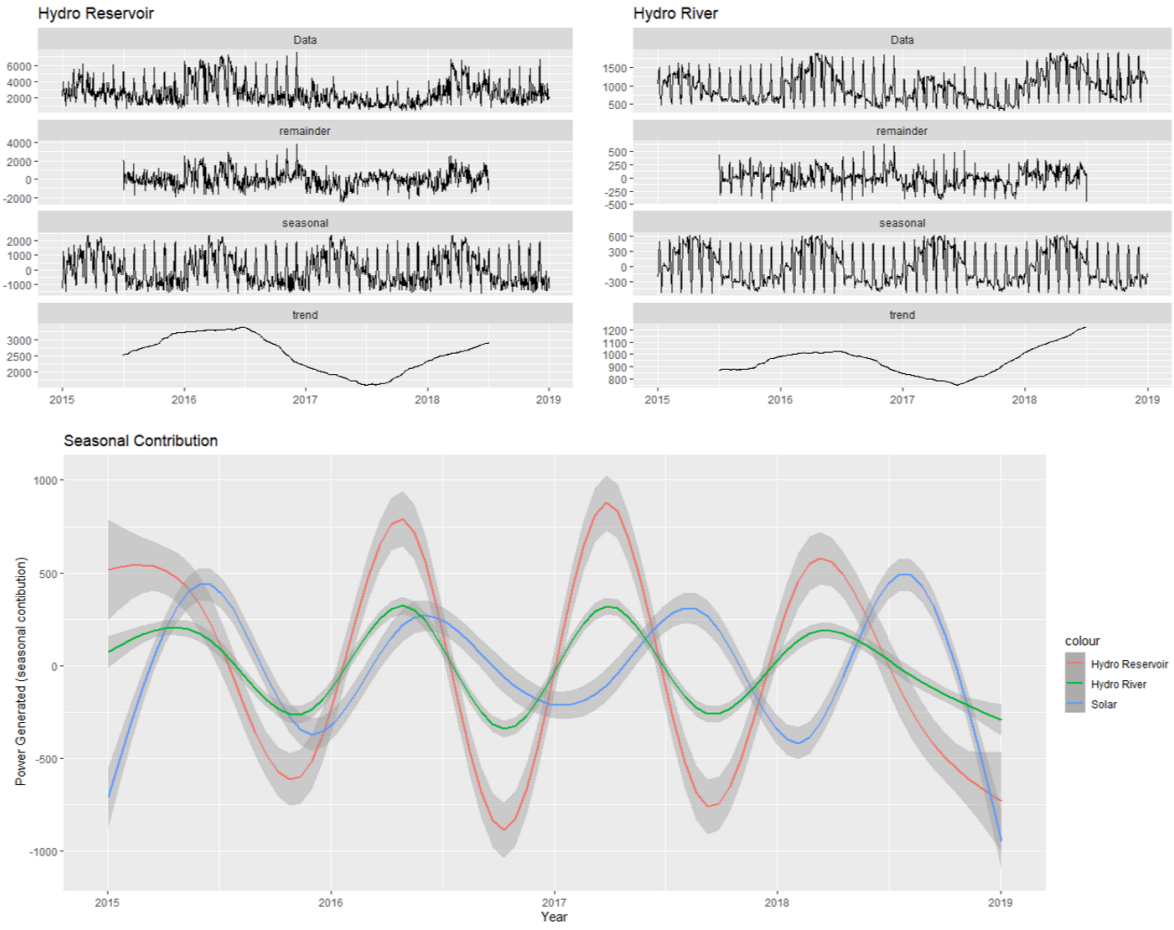


Figure 4: **Signal Decomposition of Hydro Sources**

which reaffirms and incentivises the use of non-fossil fuel based energy. Combined with increased rainfall, this diversification away from fossil fuels lead to a decrease in the electricity generated by them.

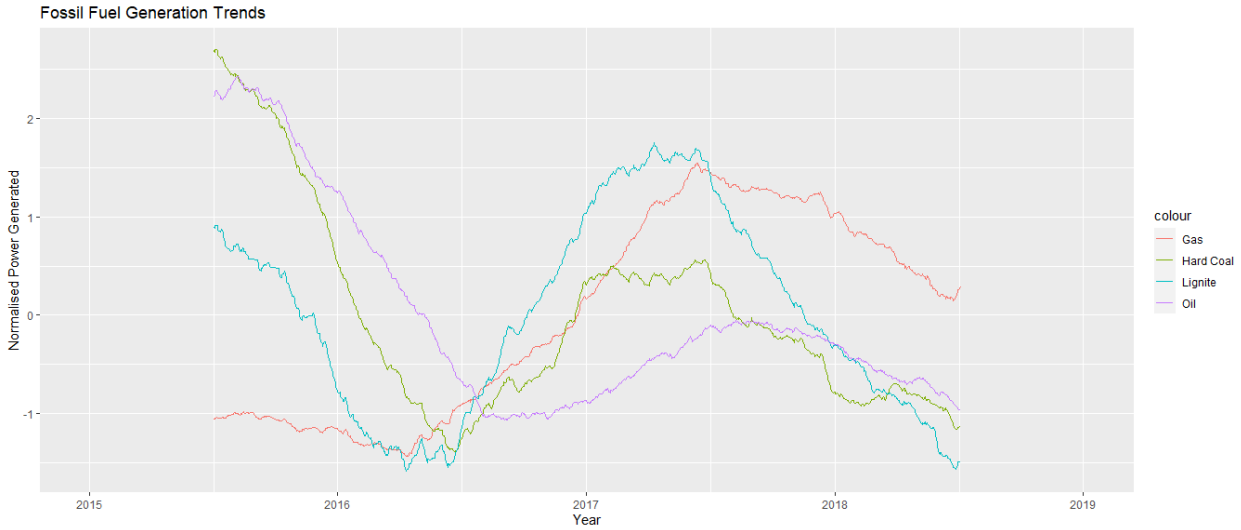
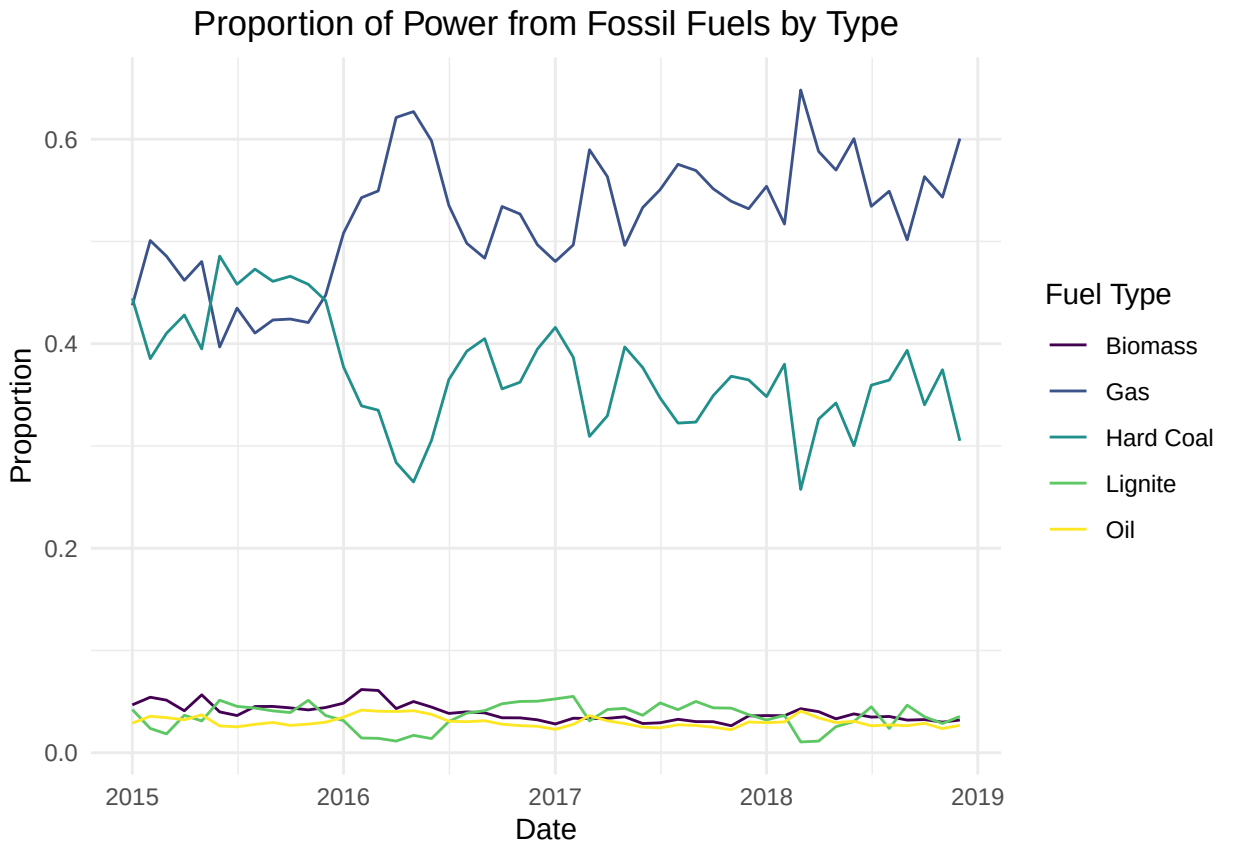


Figure 5: **Signal Decomposition of Hydro Sources**

While the total amount of power generated from fossil fuels varies over time in response to the shortfall between renewables/nuclear and consumer demand, there is a very clear inverse relationship between the proportion of power produced from Liquid Natural Gas (LNG) and the proportion produced from Hard Coal. The rise in LNG usage is likely driven by dropping prices and the availability of infrastructure to import it en masse, given that Spain is home to one third of Europe's LNG import capacity [15].



Individual contributions

1 page, 2 marks

As discussed in Assignment 2, each member independently researched topics for the project, which were then discussed and ranked through preferences of topicality, relevance, data quality, and interest, before conducting an ideas session to develop and refine an initial investigation question. The structure and responsibilities of our team were then assigned as shown below. Following this the investigation proper was conducted, consisting of data analysis and a literature review. Each member participated in weekly team meetings which were in person at first, before the Covid lockdown, at which point communications were shifted to an online platform. Outside of these meetings collaboration continued through discussion on Teams Chat as well as commenting throughout code and report.

Role	Team Member
Repository Manager	Daniel Petterson
Document/Bibliography Manager	Daniel Wrench
Data/Project Manager	Robert Tomkies

Who	What	Outcome
Repo Manager	Set up Repo	Github Repo
Everyone	Explore Data	Understanding of data
Everyone	Refine Question	Improved question
Everyone	Literature Review	An understanding of relevant literature

In addition to these, each member of the team carried out a variety of other functions as the need arose.

Daniel P

Daniel P served as the repository manager which involved setting up the Github repository and ensuring that it was well organised and any conflicts were dealt with. Daniel P also built the Shiny dashboard to allow in-demand analysis of the data by the various stakeholders.

The primary focus of his section of literature review was the change in national energy generation capacity in the years leading up to the time our data was collected and the impact of environmental or financial variables in that time frame. In addition to this he analysed changes in fossil fuel generation methods and how they reacted to changes in renewable energy generation.

Rob

Rob took on the role of timeline/project manager. This initially involved the set up of the online workspace in Microsoft Teams, including the Kanban board and project timeline in the form of Gantt charts. These were maintained throughout the project, updating with new tasks and marking as in-progress or complete so that progress could be quantified, ensuring the project remained on track. The workspace also featured polls for when to have meetings and easy access to the Git repository.

Rob also conducted both wider literature review and data analysis. The literature review focused primarily on the political environment present in Spain at the time to better understand some of the drivers for the trends seen. Further to this he examined different energy time series analysis studies and the techniques utilised within them to help inform the most appropriate approached to apply to this study. His analysis focused on the seasonal decomposition and trend isolation within the data, as well as a distribution analysis to examine the behavior of the sources independently within the time series.

Beyond this, Rob took responsibility for the consideration and write up of ethics, privacy and security concerns relating to this project, as well as contributing towards the data analysis write-up, and reviewing some of the Shiny app code.

Daniel W

Daniel W was the document and bibliography manager. As part of this he was in charge of structuring, proof-reading, and editing the report. During the data preparation stage, he investigated missing data, outliers, and the effect of linear interpolation. During the analysis stage, he investigated correlations, long-term trends, and seasonality. He then produced plots and wrote about these in the final report. He also contributed to the literature review and helped to identify issues during development of the Shiny dashboard.

Conclusion

An exploratory data analysis of electricity generation in Spain between 2015 and 2018 was conducted to investigate how generation by different sources has changed over time. This dataset required cleaning of a small number of missing values and outliers; these were filled in using linear interpolation. We found that Nuclear was the most important generation source over the period, not only due to its high mean output but also its low variability. A number of low-output sources showed significant long-term change in their average output. Seasonality was observed in solar generation (on two scales), and to a lesser extent in hydroelectric sources. The overall generation amount remained roughly constant for Spain and as such this seasonality induced echo seasonality in some of the fossil fuel sources. This analysis could help inform government policy around clean energy investment, and there were some concerns around ethical implications of how this research could affect future policy around climate change and affordable energy.

References

1. ENTSO. Central collection and publication of electricity generation, transportation and consumption data and information for the pan-european market. https://transparency.entsoe.eu/dashboard/show

2. Jhana. N Electrical demand, generation by type, prices and weather in Spain. https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather

3. Directive 2009/28/EC of the European Parliament and of the Council of 23 April 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC. https://eur-lex.europa.eu/eli/dir/2009/28/oj

4. Plan de Energias Renovables 2011-2020. In: IDAE. https://www.idae.es/sites/default/files/documentos/publicaciones_idae/documentos_11227_per_2011-2020_def_93c624ab.pdf

5. (2016) El Sistema Electrico Espanol 2016. https://www.ree.es/sites/default/files/11_PUBLICACIONES/Documentos/InformesSistemaElectrico/2016/inf_sis_elec_ree_2016.pdf

6. (2018) El Sistema Electrico Espanol 2018. https://www.ree.es/sites/default/files/11_PUBLICACIONES/Documentos/InformesSistemaElectrico/2018/inf_sis_elec_ree_2018.pdf

7. Anti-nuclear movement in Spain. In: Wikipedia. https://en.wikipedia.org/wiki/Anti-nuclear_movement_in_Spain

8. Power to the people: how Spanish cities took control of energy. In: The Guardian. https://www.theguardian.com/environment/2019/jun/14/power-to-the-people-how-spanish-cities-took-control-of-energy

9. World Directory of Minorities and Indigenous Peoples: Spain. https://minorityrights.org/country/spain/

10. Analysis and comparative review of equality data collection practices in the European Union. https://ec.europa.eu/newsroom/just/redirection/document/45791

11. Gürtler K, Postpischil R, Quitzow R (2019) The dismantling of renewable energy policies: The cases of Spain and the Czech Republic. Energy Policy 133:110881. https://doi.org/https://doi.org/10.1016/j.enpol.2019.110881

12. (2016) Limited Risk of Energy Reforms Reversal in Spain; Political Risk Remains. https://www.fitchratings.com/research/corporate-finance/limited-risk-of-energy-reforms-reversal-in-spain-political-risk-remains-24-02-2016

13. Fernández L (2021) Spain: Average rainfall 2013-2020. Statista

14. Government of Spain T Ministry of Energy, Digital (2017) 2017-2020 national energy efficiency action plan

15. (2018) General Overview of Spanish LNG Sector. https://ec.europa.eu/energy/sites/ener/files/documents/prieto_-_lng_experience_spain.pdf