EDA of Spanish electricity generation 2015-2018

Group 17: Time Series - Assignment 3 - EDA Report

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

Due 5pm Tuesday 31st August

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 and Data

1-3 pages, 6 marks

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 (https://transparency.entsoe.eu/dashboard/show). 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. This is where the data for this report was downloaded from (https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather). *Chuck the links in this paragraph in footnotes?*

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 hourly electricity generation, which is divided into several different generation sources/methods.

In 2009 Spain became subject to EU Directive 2009/28/EC which mandates an EU-wide commitment to expanding the use of renewable energy (https://eur-lex.europa.eu/eli/dir/2009/28/oj). 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.(https://www.idae.es/sites/default/files/documentos/publicaciones_idae/documentos_11227_per_2011-2020_def_93c624ab.pdf)

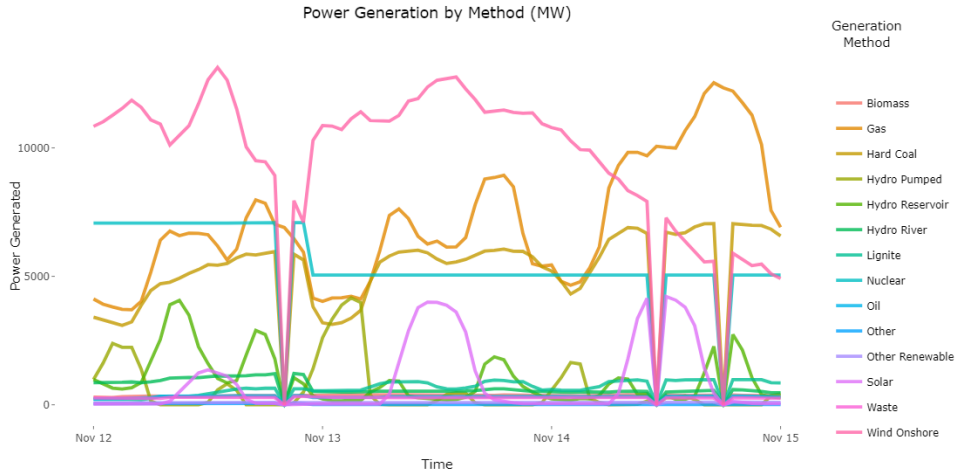
The question we formulated to guide our EDA is as follows:

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.

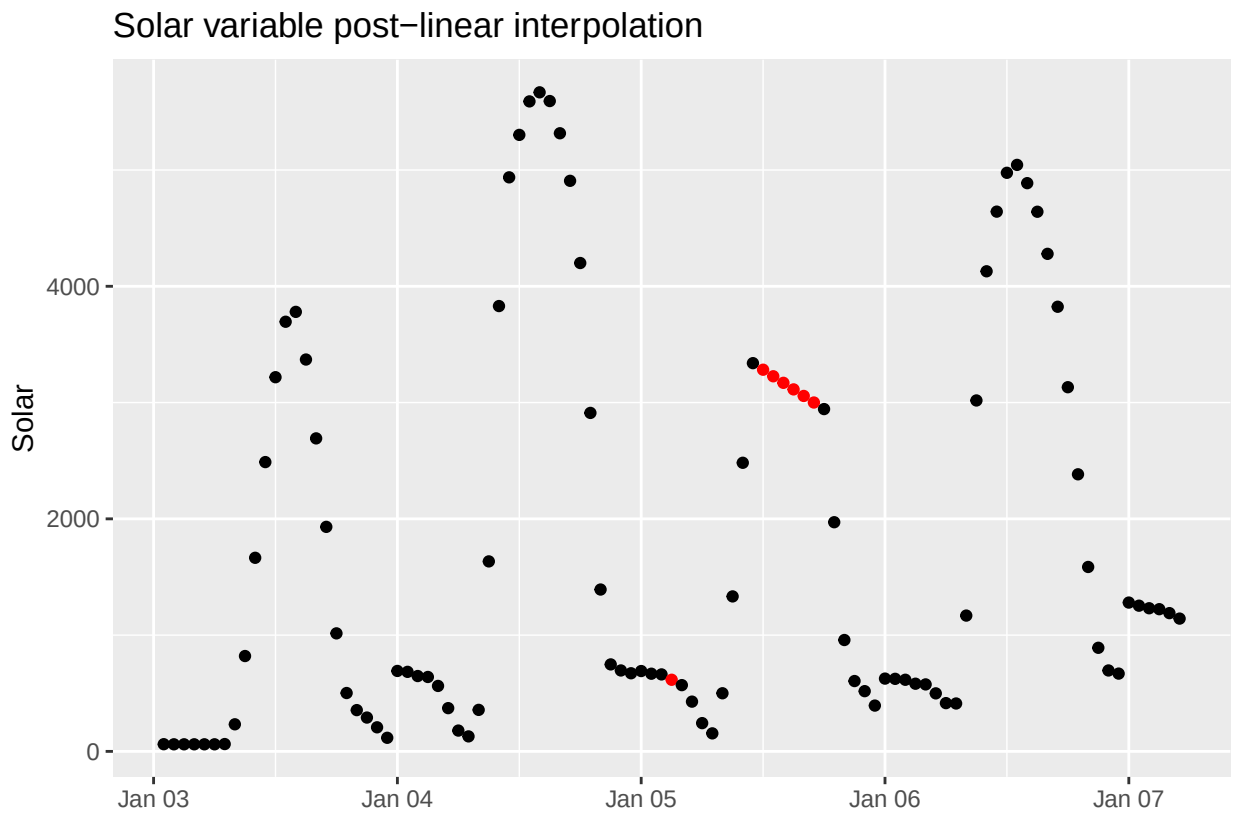
After removing the variables that were not relevant to our analysis, and those that contained no observations, the dataset contained hourly recordings of the amount of power generated from 14 different sources in Spain between 2015 and 2019. The definitions of each of these sources are reasonably self-explanatory, and we have expanded in further detail during our analysis when it is helpful.

As well as some missing values (discussed further below), we also discovered some outliers. Interestingly, these occurred across variables at the exact same timestamps. Specifically, we identified 3 rows in the raw dataset in which all but one generation source recorded a value of 0. This is shown by the three negative spikes in the plot below:

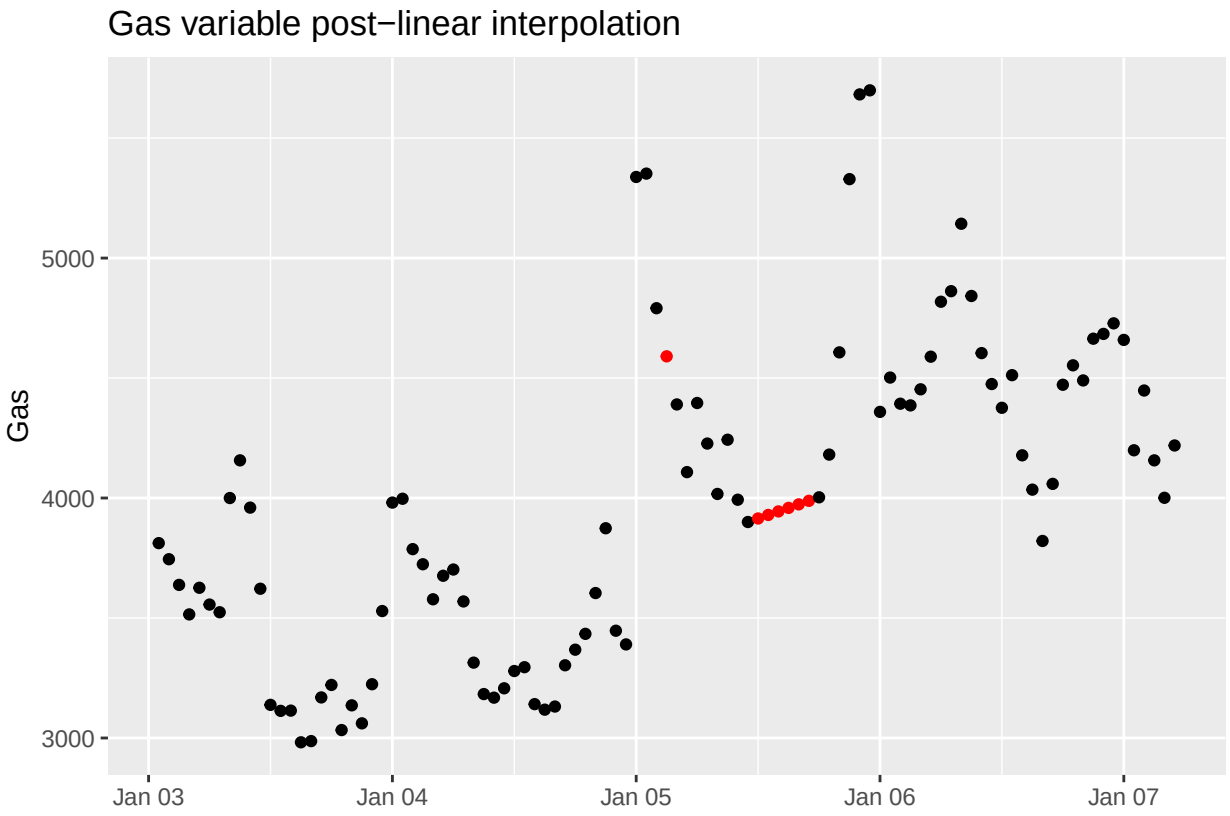


Because they seem very unlikely to be true values, these outliers were removed. *We also identified two other rows where values were a significant departure from the typical range of that source, but were not equal to zero. These have not been removed because they could be plausible values (**confirm whether we are OK with this, otherwise I (Dan W) am happy to remove these too**)*

After removing the outliers, each variable had between 18 and 22 missing values. Most of these missing values occurred across variables: there were 21 rows for which data for all or all but one variables were missing, and five rows in where only one variable was missing. In most cases, we were confident in imputing the missing values using linear interpolation. However, there was one area of slight concern: of the 26 rows with missing data, six of them are consecutive. The plots below show how this imputation affected two of the variables.



The linear interpolation over the adjacent missing values in the Solar variable 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 utilize 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 (https://en.wikipedia.org/wiki/Anti-nuclear_movement_in_Spain). 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 and Data, 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 completed from this data is at a national level and so

assumes a uniform complete accessibility to energy from all sources, this is inherently not the case. If the government decided to shift to a non-diversified energy model or banned local diesel generators there is the small possibility that areas, most likely rural, would no longer have access to electricity and have an adverse effect on these people lives. More likely is where, due to political pressure, the government chooses to invest in a more expensive source of energy, for example due to a drive to achieve a higher proportion of green energy (https://www.theguardian.com/environment/2019/jun/14/power-to-the-people-how-spanish-cities-took-control-of-energy). Pushing power prices up would disproportionately affect people from a lower socioeconomic background and fuel a growing divide between the classes and creating a growing poverty in the country. This could be further exacerbated if with the investment in new technology automation is increased, leading to greater efficiency at the cost of lower skilled jobs, again disproportionately affecting the lower socioeconomic class.

There is a small risk of the creation of 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 (https://minorityrights.org/country/spain/) 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 (https://ec.europa.eu/newsroom/just/redirection/document/45791). Furthermore, this data is recorded and stored at a national level from generation source rather than a conglomerate of personal data and as such 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 the environment 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 would only be a concern if the data was comprised of the sum of individuals energy use added up and then anonymised. A larger concern relates to if the data was of finer resolution, such as region or street where theoretically if the sample size is small, even with anonymised data, individuals could be identified with local knowledge, again this is not the case as information is provided at a national level and measures taken at source. This means that data anonymisation techniques such as differential privacy are not only not necessary, but detrimental to the accuracy.

Beyond personal privacy, as this data is available publicly from Spanish TSO Red Electric España there is no concern of corporate property breach. Further to this, the body holding the data is the national government body, they 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 ensure the security of the project. Confidentiality is ensured through the resolution of the data being at a national scale and being collected at 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 through Microsoft Teams, a git repository and Overleaf which all require a minimum password protection or multifactor authentication. While

the information within each storage means does not need to be protected, each company has a responsibility to protect the personal information of our team.

Integrity was maintained through the upkeep of a git repository ensuring that we had a complete record of any code and any changes made to the code throughout. This ensured the code present was the most up to date and relevant while also allowing us to keep records of versions throughout the project in case of needing to roll back edits to the code. The data was linked directly to the Kaggle dataset to ensure no local edits would affect it; this does mean that we relied on the integrity of the Kaggle dataset. Access to edit all 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 also be recorded through git and recorded, ensuring that it would be simple to a point we were sure the code was correct and our work.

Availability of data and files was ensured through utilizing cloud-based platforms for all data, codes, and reports. This ensures 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 so utilize top level protection however should the servers go down due to something such as a DDOS attack, each git repository is stored locally on devices as well so one of us would have the most up to date file.

Exploratory Data Analysis

3-8 pages, 15 marks

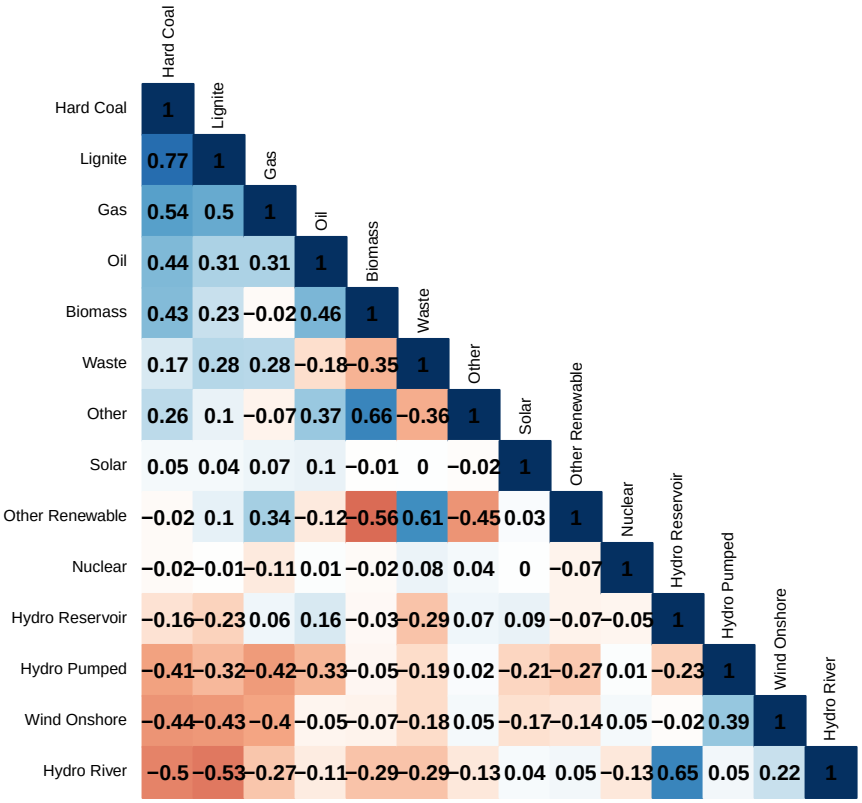
Overall statistics

	Mean	SD	CV	Min	Median	Max	Histogram
Nuclear	6263.97	838.33	0.13	998.00	6564.50	7117.00	<U+2583><U+2581><U+2583><U+2587>
Gas	5622.70	2201.51	0.39	0.00	4969.50	20034.00	<U+2583><U+2587><U+2582><U+2581><U+2581>
Wind Onshore	5465.60	3213.25	0.59	234.00	4850.00	17436.00	*<U+2587><U+2586><U+2585>*<U+2582><U+2581><U+258
Hard Coal	4257.09	1961.71	0.46	576.00	4475.00	8359.00	*<U+2586><U+2585><U+2585><U+2585><U+2587><U+258*
Hydro Reservoir	2605.66	1835.07	0.70	134.00	2165.00	9728.00	<U+2587><U+2586><U+2585>_<U+2583><U+2582><U+25
Solar	1432.95	1680.00	1.17	2.00	616.00	5792.00	<U+2587><U+2582><U+2581><U+2581><U+2581><U+2581
Hydro River	972.27	400.63	0.41	283.00	906.00	2000.00	<U+2582><U+2587><U+2587><U+2585><U+2585><U+2585
Lignite	448.17	354.62	0.79	0.00	509.00	999.00	<U+2587><U+2581><U+2581><U+2582><U+2582><U+2583><U+2581
Hydro Pumped	475.59	792.31	1.67	0.00	68.00	4523.00	<U+2587><U+2581><U+2581>
Biomass	383.56	85.27	0.22	0.00	367.00	592.00	<U+2581><U+2582><U+2586><U+2587><U+2581><U+2583
Oil	298.37	52.45	0.18	44.00	300.00	449.00	<U+2581><U+2583><U+2585><U+2587>_<U+2581><U+25
Waste	269.44	50.16	0.19	39.00	279.00	357.00	<U+2581><U+2581><U+2582>_<U+2586><U+2587><U+25
Other Renewable	85.64	14.06	0.16	4.00	88.00	119.00	<U+2581><U+2585><U+2585><U+2587><U+2586><U+2581
Other	60.23	20.23	0.34	0.00	57.00	106.00	<U+2581><U+2583><U+2587><U+2581><U+2582><U+2583>

The table above shows summary statistics 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. We can see that Nuclear has the largest average hourly generation, but a relatively low standard deviation. This combination of high production and low variability, an important measure for understanding

an electricity system, is best summarised by the coefficient of variation (CV) statistic. Nuclear has the lowest CV out of all sources. This statistic helps us understand the difference between intermittent sources (high CV, such as Solar and Wind Onshore), and continuous sources (low CV, such as Nuclear, Oil, 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 correlations as a way of investigating relationships in the dataset.



Nuclear, the largest overall contributor to electricity generation, does not have any significant correlations. Therefore, not only is Nuclear an incredibly steady generation source, it is also one that does not depend on or affect the amount of electricity produced by other sources. Solar, on the other hand, is a much more variable source, but this correlation plot shows it is also a very independent generation method.

The three least independent generation sources - those most related to other sources - are Hard Coal, Hydro River, and Other Renewable. The largest correlation 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 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 much to smooth the data, and how to compare the variables. Here, we focus on the most interesting patterns.

The plot below shows the five sources in the data with the most significant long-term trends over the four-year period. The similar or opposite nature of these trends between these generation sources are reflected in strong

positive or negative correlations.

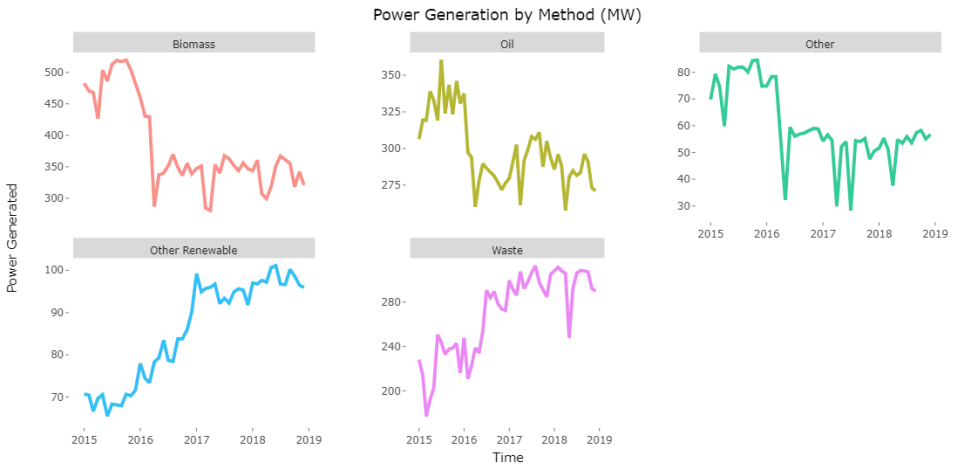
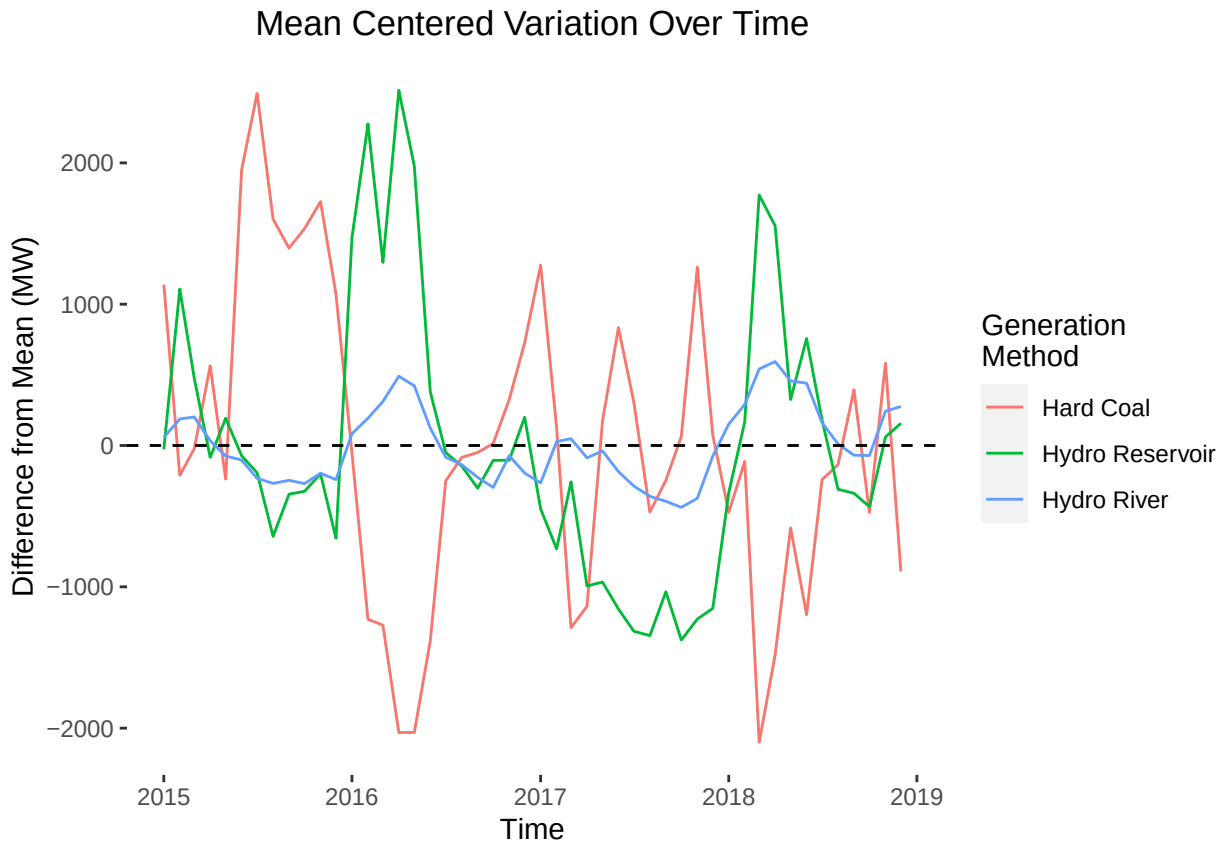


Figure 1: **Biomass, Oil, Other Renewable, Other, and Waste time series with monthly smoothing**

Hard Coal and Hydro River, as shown in the correlation plot, are the two least independent generation sources - in other words, those most related to the values of 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 are either 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 both these sources and Hard Coal (correlation with Hydro Reservoir = -0.16, with Hydro River = -0.5). The reason behind these relationships is likely to be that a more reliable energy source such as a coal is used to generate electricity when renewable (and therefore more desired) sources such as hydro power have lower output than is needed.

The only 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 variable is seasonal on a daily scale, peaking between 11am and 2pm and hitting a low at 8am and the few hours prior. This is an intuitive periodicity: a location receives the most sunlight around noon, and no sunlight during the night.

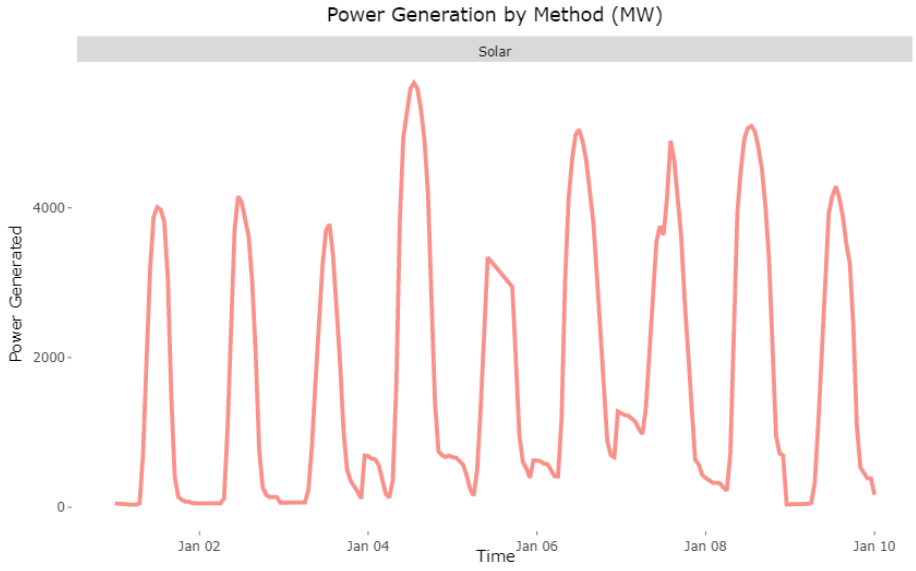


Figure 2: **Sample of Solar time series with hourly smoothing**

In the next plot we can see that Solar is 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.

Although an early pioneer in renewable energy, changes in government policy including the reduction of financial incentives 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 (https://www.sciencedirect.com/science/article/pii/S0301421519304598#sec5).

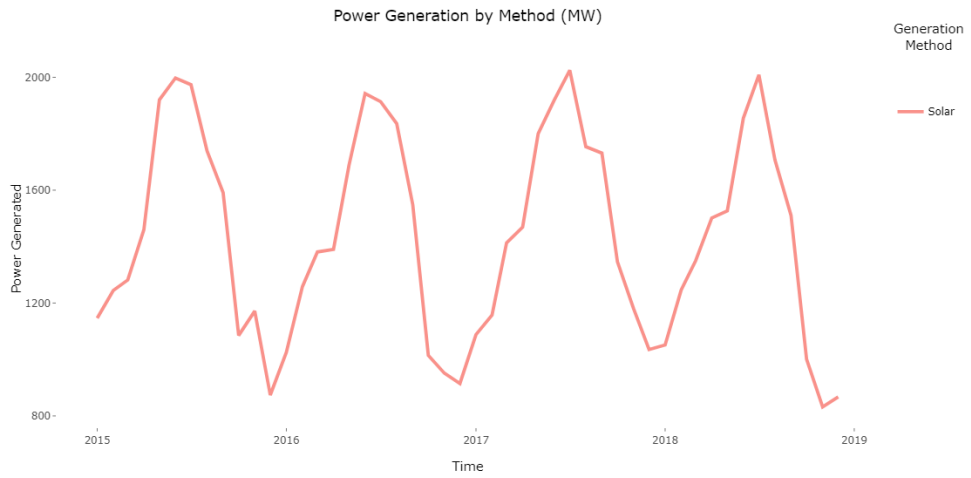
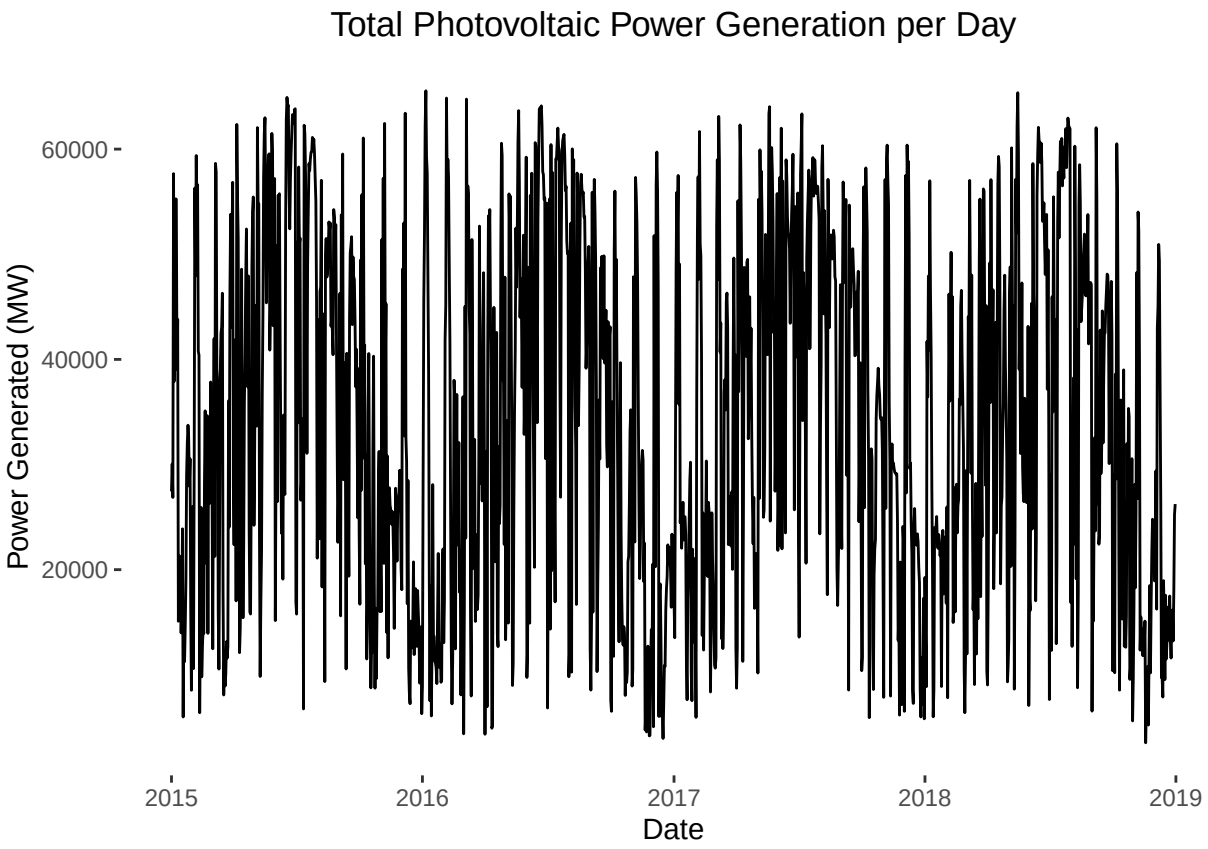


Figure 3: **Solar time series with monthly smoothing**



From the above plot we can see that total daily power generated from photovoltaic solar systems shows little change across the years we investigated.

Individual contributions

1 page, 2 marks

Quick bullet points before writing up

Daniel P:

Rob:

Daniel W:

- Investigated and wrote about outliers and effect of linear interpolation
- Investigated and wrote about correlations, long-term trends and seasonality
- Proof-read and edited Ethics, Privacy and Security section.

Conclusion