Statistical analysis of the production of electricity in Romania

Abstract

We study the influence of seasonality in different electric energy production methods by a statistical analysis of the electric energy production. This methodology enables us to extract trends and seasonality in the eight most used energy production methods in Romania. Our findings are that there are strong positive and negative correlations between the energy productions methods considered.

Keywords: statistical analysis; energy production, seasonality.

Introduction

As electricity cannot be conveniently stored without massive investments in technology and infrastructure and export of surplus electric energy cannot always be exported, there always must be a balance between the production and consumption of electrical energy. As consumption of energy cannot be adjusted because it depends on the consumers, the production side needs to be easily controllable. In some cases, such as coal and gas & oil powerplants the energy output is easily controllable but in others such as hydroelectric, solar and wind power plans the energy output is affected by seasonality

The purpose of this paper is to perform a statistical analysis on the production of electric energy of Romania to better understand the seasonality and the correlation between eight methods used to generate electricity.

Parameters for statistical Analysis

To gain a better understanding of the trends in our energy production data we use the following statistical measures.

*Mean*

The central tendency of our data is measured using the mean (or the expected value)

The mean value is calculated using the formula:

*Standard deviation (Std)*

The amount of dispersion of our data is measured using the standard deviation. A low value of the standard deviation shows that the data values tend to be close to the expected value of the set a higher value indicates that the data values are spread out on a bigger interval.

The standard deviation is calculated using the formula:

*Coefficient of skewness (Skew)*

In a paper on time series analysis Grigoletto (2009) argued that for short- and medium-term predictions the expected value is less likely to be achieved and in general, the more skewed the data, either positive or negative, the less accurate the data analysis is. To take this into account we also want to analyse the extremes of the data set by calculating the skew. The coefficient of skewness can be positive, negative, or zero. We consider that a distribution is fairy symmetrical if the coefficient of skewness is between -0.2 and 0.2, moderately skewed if the coefficient of skewness is between -1 and -0.2 or between 0.2 and 1, and highly skewed if the coefficient of skewness is less than -1 or greater than 1.

The coefficient of skewness is calculated using the formula:

*Coefficient of kurtosis (Kurt)*

Another important parameter is the coefficient of kurtosis, which has been successfully used by Loperfido (2020) for outlier detection in time series. The general explication is that a positive kurtosis means that our data distribution has a lighter tail than a normal curve with the same mean and standard deviation and thus a higher peak and fewer outliers, a negative kurtosis translates into a heavier than normal tailed distribution and thus a lower peak and more outliers.

As kurtosis is measured against the normal distribution, if the kurtosis is between -0.5 and 0.5 the data distribution can be assumed to be normal and is called a mesokurtic distribution, ff the kurtosis Is less than -0.5 the distribution is called platykurtic, and if the kurtosis is greater than 0.5 the distribution is called leptokurtic.

The coefficient of kurtosis is calculated using the formula:

*Coefficient of variation (Variation)*

It is well known from a paper by Sears (1964) that big values of standard deviation in relation to the average value may lead to limitations in the quality of statistical analysis. So, we use the coefficient of variation to rank the reliability of our analysis. The coefficient of variation is used in precise manufacturing for variability assessment Reed(2002) and in finance Weber(2004) to better understand the risk-return trade-off, there is no direct interpretation for it in the statistical analysis of energy production, so we are going to assume that a coefficient of variation over 0.5 is associated with energy production methods that are highly dependent on day to day condition and are highly influenced by phenomena that are unpredictable in the long term such as the weather in general, cloud positioning, and wind speeds.

The coefficient of variation is calculated using the formula:

Plots

The following section consists of plots of electrical energy production by method of production on the interval 7th March 2019 – 9th March 2022, together with a visual analysis of the plots.

By looking at Fig. 1, the daily plot of electrical energy produced with coal (with blue) and the seven-day running average (with orange) we notice that the seven-day running average is approximately constant on the interval March 2019 - December 2020. This period is then followed by a 50% drop on the interval January 2021- August 2021 and a sudden 350% increase in the month of September. This sudden decrease and increase in the electrical energy produced with coal seem to correlate perfectly with the 2020 Covid 19 market crash and recovery. The chart has not reached its previous heights up to date and is currently sitting at a mean value on the last 9 months of approximately 1250.

In Fig. 2 we have the daily plot of electrical energy produced with hydropower (with blue) and the seven-day running average (with orange). There seems to be a periodicity in the lows and highs of the data. The highs correlate with the rainy season for our country (August, September, October, and November) and the lows correlate with the summer and winter seasons. The spring season is a sort of a transitioning period in which the energy production increases slightly because of the melting snow but not enough to reach even a quarter of the highs during the rainy season.

Chart

Description automatically generatedChart

Description automatically generated

Fig. 1 Coal Fig. 2 Hydropower

By looking at Fig. 3, the daily plot of electrical energy produced with Oil & Gas (with blue) and the seven-day running average (with orange) we notice an inverse correlation with the highs and lows of the chart in Fig. 2. The highs in the Oil & Gas take place exactly when the lows of Hydropower do and the same is true for the lows. This is, of course normal, as the production of electrical energy with Hydropower is lower in the summer and winter months there needs to be an increase in the electrical energy produced with Oil & Gas so that the needs of the consumers are met. Then, when the production of electrical energy with Hydropower increases again the energy produced with Oil & gas needs to be lower so that the power grid does not have excess energy.

When looking at Fig. 4, the daily plot of electrical energy produced with nuclear power (with blue) and the seven-day running average (with orange), the first thing we notice is that our only nuclear powerplant in Romania functions either at 700MW or 1400MW. For most of the year the electrical output of the nuclear power plant is 1400MW, but on the upmost region of the peaks in the chart in Fig. 1 the nuclear powerplant output is decreased at 700MW this is done to keep the power grid from having excess energy. Overall, the electrical energy output of the nuclear power plant seems to be the most constant out of all the methods considered.

Chart

Description automatically generatedChart, histogram

Description automatically generated

Fig. 3 Oil & Gas Fig. 4 Nuclear

In Fig. 5 we have the daily plot of electrical energy produced with wind power (with blue) and the seven-day running average (with orange). This chart seems to be the noisiest one, where even the seven-day running average has large variations in relatively short periods of time. This behaviour is characteristic of a method of power production that is heavily reliant on unpredictable factors, in this case, the local wind speed. But nonetheless, there seems to be seasonality in the chart, there seems to be a higher output in the autumn, winter, and spring seasons when the winds are generally stronger than in the summer season.

In Fig. 6 we have the daily plot of electrical energy produced with Solar power (with blue) and the seven-day running average (with orange). It is self-evident that there will be seasonality in the data as the energy output of solar power will be greater in the summer than in the winter, with spring and autumn seasons being seen as transition periods where the power output increases respectively decrease gradually.

Chart, line chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Fig. 5 Wind Fig. 6 Solar

By looking at Fig. 6, the daily plot of electrical energy produced with biomass (with blue) and the seven-day running average (with orange) we notice the same behaviour as we did for the chart for electrical energy produced with coal. There seems to be a correlation between the behaviour of the biomass energy production in the period January 2021- November 2021where we have a 60% drop and a sudden 280% increase in the month of December. Again, this sudden decrease and increase in the electrical energy produced with coal seem to correlate perfectly with the 2020 Covid 19 market crash and recovery.

For Fig. 8, the daily plot of electrical energy that is imported or exported (with blue) and the seven-day running average (with orange), we notice a high daily variability where it goes from importing to exporting high amounts of electrical power almost every day. This constant alternation between importing and exporting electrical power makes it hard for a trend to form but nonetheless, there is a small correlation between the highs of electric energy production using hydroelectric power and the negative lows (exporting) of this chart.

Chart

Description automatically generated with medium confidenceA picture containing pencil

Description automatically generated

Fig. 7 Biomass Fig. 8 Import/Export

Results of Statistical Analysis and Histogram analysis

The selected results of the daily distribution of energy production grouped by the method of production are presented in Table 1.

The highest three means of electric energy production are hydroelectric, nuclear and coal, these the production methods amount for approximately 65% of the total energy production of Romania.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Mean | Std | Skew | Kurtosis | Variation |
| Coal | 1270 | 315.2 | (-)0.231 | 0.103 | 0.248 |
| Hydroelectric | 1853 | 724.3 | 0.661 | 0.101 | 0.390 |
| Oil & Gas | 1150 | 431.5 | (-)0.02 | (-)0.877 | 0.375 |
| Nuclear | 1299 | 236.54 | (-)2.12 | 2.68 | 0.182 |
| Wind | 766 | 672.96 | 0.972 | 0.013 | 0.879 |
| Solar | 154 | 225.67 | 1.304 | 0.366 | 1.468 |
| Biomass | 59 | 14.02 | (-)0.345 | (-)0.676 | 0.240 |
| Import/Export | 248 | 700.09 | (-)0.248 | (-)0.458 | 2.818 |

Table 1: Results of the statistical analysis by method of production

Pt perioada de test

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Mean | Std | Skew | Kurtosis | Variation |
| Coal | 1230 | 188 | -0.007 | 0.28 | 0.15 |
| Hydroelectric | 1825 | 747 | 0.71 | -0.29 | 0.40 |
| Oil & Gas | 1226 | 378 | -0.30 | -0.42 | 0.31 |
| Nuclear | 1289 | 245 | -1.92 | 1.83 | 0.19 |
| Wind | 756 | 641 | 0.99 | 0.21 | 0.84 |
| Solar | 133 | 206 | 1.47 | 0.91 | 1.55 |
| Biomass | 65 | 8.44 | -0.39 | 0.11 | 0.12 |
| Import/Export | 228 | 760 | -0.17 | -0.49 | 3.33 |
| Productie | 6501 | 934 | 0.14 | -0.12 | 0.14 |
| Consum | 6726 | 987 | 0.28 | -0.69 | 0.14 |

Ce prezic

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Mean | Std | Skew | Kurtosis | Variation |
| Coal | 1212 | 143 | 0.15 | 0.27 | 0.11 |
| Hydroelectric | 0544 | 489 | 1.14 | 0.26 | 0.31 |
| Oil & Gas | 1225 | 306 | -0.06 | -0.64 | 0.25 |
| Nuclear | 1289 | 228 | -1.94 | 1.89 | 0.17 |
| Wind | 752 | 340 | 1.01 | 0.59 | 0.45 |
| Solar | 133 | 204 | 1.53 | 1.11 | 1.53 |
| Biomass | 54 | 0.69 | -0.41 | -0.36 | 0.10 |
| Import/Export | 154 | 373 | -0.49 | -0.54 | 2.41 |
| Productie | 6432 | 601 | 0.12 | -0.5 | 0.09 |
| Consum | 6715 | 500 | 0.18 | -0.09 | 0.07 |

By looking at the histogram of the energy produced by coal in Fig. 9 we observe that the highest frequency of occurrence is at 1250, a value close to the mean of 1270 calculated in Table 1. This means that the data distribution is skewed. This fact is confirmed by the coefficient of skewness calculated as -0.231, which is between -1 and -0.2 and is associated with a data distribution moderately skewed to the right (higher values). The coefficient of kurtosis was calculated as 0.103, which is between the interval of -0.5 and 0.5, so our distribution fairly resembles a normal one and is called a mesokurtic distribution. The standard deviation in relation to the mean, calculated through the coefficient of variation is 0.248 which is smaller than 0.5 and corresponds with data that is distributed near the mean and is not highly influenced by long-term unpredictable phenomena.

We now look at Fig. 10, the histogram of the energy produced through hydroelectric means, and we observe that the highest frequency of occurrence is at 1750 MW, a little bit off from the expected value of 1853 calculated in Table 1, which shows skewness in the data distribution. Again, we look at the calculated coefficient of skewness of 0.661, which is greater than 0. and less than 1, so the data distribution for hydroelectric power generation is moderately skewed to the left (lower values). The coefficient of kurtosis is 0.101, which is between -0.5 and 0.5, so our distribution fairy resembles a normal one and is called a mesokurtic distribution. The coefficient of variation calculated as 0.309 is below 0.5, so we can say that our data is mainly distributed near the mean value and is not highly affected by long-term unpredictable phenomena.

**Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated**

Fig. 9 Coal Fig. 10 Hydroelectric

In Fig. 11, the histogram of energy production through Oil & Gas, we remark that the highest frequency of occurrence is at 1150, exactly as the value that was calculated in Table 1. So, our distribution is almost symmetrical, a fact that is confirmed by the coefficient of skewness of -0.02, this means that our data distribution is not skewed at all. The coefficient of kurtosis was calculated as -0.877 which is smaller than -0.5 so our distribution is heavier tailed than normal distribution with the same standard deviation and thus has a lower peak and is called a platykurtic distribution. The coefficient of variation is 0. 375, which is smaller than 0.5, so we say that our data is mainly distributed near the mean value and is not highly affected by long-term unpredictable phenomena.

For Fig. 12, the histogram of energy production through nuclear means, we see that the highest frequency of occurrence is at 1400, quite different from the expected value of 1299 calculated in Table 1, so we expect our data distribution to be skewed. Our expectations are confirmed by the coefficient of skewness of -2.12 which is well below the threshold of -1, this means that our data distribution is highly skewed to the right (higher values). The coefficient of kurtosis was calculated as 2.68, which is greater than 0.5, so our data distribution is lighter tailed than a normal distribution with the same standard thus having a higher peak and is called a leptokurtic distribution. The coefficient of variation that was obtained is 0.182, which is below 0.5 and the lowest of all the ones calculated in Table 1, this shows that most of the data is packed near the expected value and that it is not highly affected by long term unpredictable phenomena, but it also shows that nuclear energy has the most constant output from all the other means of energy production.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Fig. 11 Oil & Gas Fig. 12 Nuclear

By looking at the histogram of energy production through wind power in Fig. 13, we observe that the highest frequency of occurrence is at 0 while the mean value calculated in Table 1 is 766. This means that our distribution is skewed, an effect which can clearly be seen visually, but as confirmation we have the coefficient of skewness calculated as 0.972, which is between 0.2 and 1 so our data distribution is moderately skewed to the left (lower values). The coefficient of kurtosis was calculated as 0.013 and is between -0.5 and 0.5, so our distribution fairly resembles a normal one and is called a mesokurtic distribution. The standard deviation reported to the mean, calculated through the coefficient of variation is 0.879, which is greater than 0.5 and corresponds to a data distribution that is not packed near the mean and is highly influenced by long-term unpredictable phenomena, in this case, the local wind speed.

We now switch to Fig. 14, the histogram of energy production through solar, we remark that the highest frequency of occurrence is at 0 and the expected value calculated in Table 1 is 154, so our distribution is skewed, an effect which can be seen visually. This is confirmed by the coefficient of skewness of 1.304, which is greater than 1 and is associated with a distribution heavily skewed to the left (lower values). The coefficient of kurtosis was calculated as 0.366, which is between -0.5 and 0.5 so our distribution fairly resembles a normal one and is called a mesokurtic distribution. The coefficient of variation is 1.468, greater than 0.5 which corresponds to data that is not mainly distributed near the expected value and is highly influenced by long-term unpredictable phenomena, in this case, the general weather and the positioning of the clouds.

Chart, histogram

Description automatically generatedA picture containing histogram

Description automatically generated

Fig. 13 Wind Fig. 14 Solar

For Fig. 15, the histogram of energy production through biomass fuel, we see that the highest frequency of occurrence is at 65, while the expected value calculated in Table 1 is 59, so our data distribution is skewed, a fact that is confirmed by the coefficient of skewness of -0.345 which is between -1 and -0.2 and is associated to a data distribution moderately skewed to the right (higher values). The coefficient of kurtosis was calculated as -0.676 which is less than -0.5, so our data distribution is heavier tailed than a normal distribution with the same standard thus having a lower peak and is called a platykurtic distribution. The coefficient of variation is 0.24, which is less than 0.5, this shows that most of the data is distributed near the mean and is not highly affected by long-term unpredictable phenomena.

We now look at Fig. 16, the histogram Import/Export of electrical energy, in which negative values represent export of electrical energy and positive values represent import of electrical energy. We see that the highest frequency of occurrence is at 500 while the mean value calculated in Table 1 is 248, this means that our data is skewed. The skewness is confirmed by the coefficient of skewness of -0.248, which is between -1 and -0.2 so our data distribution is moderately skewed to the right (higher values). The coefficient of kurtosis was calculated as -0.458 which is between -0.5 and 0.5 but still extremely close to the limit value of -0.5, so our data distribution is at the limit between a mesokurtic and a platykurtic distribution. The coefficient of variation obtained is 2.818, which is well above the threshold of 0.5, this shows that the data is not mainly distributed near the mean and that it is highly affected by long term unpredictable phenomena, in this case, other all the other parameters that influence the rest of the energy production affect the Import/Export of electrical energy. This is normal as energy is imported or exported depending on the energy production in that day, which tends to follow a seasonal trend but is also highly influenced by long-term unpredictable phenomena.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Fig. 15 Biomass Fig. 16 Import/Export

Conclusions

During the statistical analysis of the energy production of Romania on the 7th of March 2019 – 9th of March 2022 period we have found that most other means of production of electrical energy are highly negatively or positively correlated with the energy production of hydroelectric power. The energy production through coal and biomass also seems to be positively correlated with the Covid 19 market crash and subsequent recovery.

We have observed seasonality in energy production through hydro, solar, oil & gas, and nuclear power. Again, this seasonality seems to be mainly the result of the rainy season, when electric energy generated through hydropower increases and so the oil & gas and nuclear power production of energy needs to be decreased so that the power grid doesn’t get overloaded, and summer season in the case of solar power.

Solar, wind, and import/export have the highest coefficients of variation, well above 0.5, representative of data distributions that are affected by long terms unpredictable and highly unpredictable phenomena such as cloud positioning, local wind speed, general weather. At the opposite, end nuclear power has the smallest coefficient of variation, this is a result of the constant 140MW output of the only nuclear power of Romania, only in special cases such as power grid overloading dost the output gets lowered to 700 MW.

References

Grigoletto, M. and Lisi, F., 2009. Looking for skewness in financial time series. *The Econometrics Journal*, *12*(2), pp.310-323.

Loperfido, N., 2020. Kurtosis-based projection pursuit for outlier detection in financial time series. *The European Journal of Finance*, *26*(2-3), pp.142-164.

Reed, G.F., Lynn, F. and Meade, B.D., 2002. Use of coefficient of variation in assessing variability of quantitative assays. *Clinical and Vaccine Immunology*, *9*(6), pp.1235-1239.

Searls, D.T., 1964. The utilization of a known coefficient of variation in the estimation procedure. *Journal of the American Statistical Association*, *59*(308), pp.1225-1226.

Weber, E.U., Shafir, S. and Blais, A.R., 2004. Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. *Psychological review*, *111*(2), p.430.