Research (technical) Article

Multiple Data-Driven Missing Imputation for Wind Energy Datasets

Sergii Kavun [1,2*], Alina Zamula [3]

[1] Interregional Academy of Personnel Management, Kyiv, Ukraine; kavserg@gmail.com
[2] Luxena Ltd, Lead of Data Science Team, Kyiv, Ukraine; sergii.kavun@luxena.com
[3] Global Logic, Kharkiv, Ukraine; zamula.alina@gmail.com

* Correspondence: kavserg@gmail.com; Tel.: +38-0677-09-5577
Ukraine, c. Uzhorod, Kavkazska str., 33, 88017

Abstract: Missing data introduces ambiguity in data analysis, and addressing this issue is crucial for accurate insights. In this paper, we focus on the task of multiple data imputation in wind energy datasets with time-series data. We propose a novel statistical approach to handle multiple missing values, including single, double, triple, quadruple, and pentuple gaps in time-series data. Our approach involves the development of an advanced imputation algorithm specifically tailored for wind energy datasets. We evaluate the performance of our approach on real-world wind energy datasets, demonstrating its effectiveness in generating reliable and complete data. The imputed datasets can be used for various purposes, including long-term statistical reporting and short-term/long-term forecasting. Our methodology involves several key stages: 1) exploratory analysis to understand the data characteristics; 2) data cleaning by removing irrelevant and unsuitable data points; 3) robust imputation of missing values using our proposed algorithm; 4) comparative analysis to assess the quality of the imputed data; and 5) comprehensive evaluation of the results. The practical value of our work lies in its applicability to analyze and predict wind energy processes in a Big Data environment, even when dealing with incomplete datasets or data aggregation from multiple sources. By effectively handling missing values, our approach enhances the reliability and accuracy of statistical reports and forecasting models for wind energy systems.

Keywords: exploratory analysis; data preprocessing; time series; imputation; missing data; multiple gaps; power system analysis; wind power generators

1. Introduction

The increasing importance of wind energy as a sustainable and renewable power source has led to a growing interest in analyzing wind energy datasets. However, these datasets often suffer from the issue of missing data, which can introduce ambiguity and hinder accurate analysis. Addressing this challenge

requires effective methods for imputing the missing values and generating complete datasets for further analysis and modeling.

Many case studies based on mass surveys have missing values. It leads to a decrease in statistical power. That is, reduce the likelihood of finding real patterns in the data and producing meaningful insights. Processing of missing values is a fairly developed research area with generally accepted terminology and many solutions for various disciplines and specific studies.

Imputation is enough costly approach, especially in terms of time and the required program code, therefore, before delving into its technical details, it makes sense to briefly identify key points.

Multiple imputations are quite applicable to complex processes with complex multiple causalities. This is since that the goal of imputation is not only the prediction of values, but the addition of the data array with the smallest distortions compared to alternative methods [39]. It could be a technical support procedure and diagnostic tool of possibilities and limitations of available empirical data, whose results can be taken into account even when adjusting the goals and objectives of the study.

Considering relevant studies [9], [29], [21], [12] and research [13], [28] it looks quite promising to come through that area and create new opportunities to improve the efficiency of research on the level of business company and government control as well.

In this study, we focus on the task of multiple data imputations in wind energy datasets with time-series data. Our goal is to develop a statistical approach that can handle various types of missing values, including single, double, triple, quadruple, and pentuple gaps in the time-series data. By addressing the problem of multiple missing values, we aim to provide a robust solution that can enhance the reliability and accuracy of data analysis and modeling in the field of wind energy.

Unlike existing approaches, which often rely on simplistic imputation methods or ignore the temporal dependencies in the data, our proposed approach takes into account the time-series nature of the wind energy datasets. We propose an advanced imputation algorithm that leverages statistical techniques to accurately fill in the missing values, considering the correlations and patterns within the data.

To evaluate the effectiveness of our approach, we conduct experiments on real-world wind energy datasets. We compare the imputed datasets generated by our algorithm with the original complete datasets to assess the quality of the imputation results. Furthermore, we demonstrate the practical applications of the imputed datasets, including their usefulness in forming long-term statistical reports and their potential for improving short-term and long-term forecasting models for wind energy systems.

The remainder of this paper is organized as follows. Section 2 provides a comprehensive review of related work in the field of missing data imputation and its applications in wind energy analysis. Section 3 present some primary results of exploratory data analysis in wind energy datasets with time-series data. Section 4 describes the methodology and details our proposed algorithm for multiple data

imputation. Section 5 presents the experimental setup and discusses the results and performance evaluation. Finally, Section 6 concludes the paper with a summary of our findings, contributions, and potential future research directions.

By addressing the challenges of missing data in wind energy datasets, this research aims to contribute to the field by providing a reliable and effective approach for generating complete datasets. The outcomes of this study have implications for improving data analysis, modeling, and decision-making in the domain of wind energy, ultimately promoting the efficient and sustainable utilization of wind resources.

2. Literature Review

Data imputation is a fundamental task in data science. It has been investigating in a wide range of domains, such as clinical trials, ecology, retail, energy generation. For instance, wind turbine power prediction may prevent unexpected losses and failures. In that case substitution of missing values becomes essentially important.

This approach is divided into two major branches of applicable methods and techniques. The first branch includes AI-based and machine learning methods that aim to use neural networks, fuzzy logic [40], and regression models [34].

These few publications represent significant contributions to the field of missing data imputation in wind energy datasets. They cover various techniques, including statistical approaches, deep learning methods, Bayesian networks, and machine learning algorithms, providing researchers and practitioners with valuable insights and approaches for addressing the challenge of missing data in wind energy analysis.

Smith, et al. [26] conducted a comprehensive review of missing data imputation techniques specifically applied to renewable energy datasets. They examined various approaches, including statistical methods, machine learning algorithms, and time-series modeling. The authors provided an in-depth analysis of the strengths and limitations of each technique, highlighting their applicability to renewable energy data analysis. This review serves as a valuable resource for researchers and practitioners working with missing data in renewable energy studies.

Lee, et al. [14] proposed a temporal pattern-based imputation method specifically designed for wind energy datasets. By leveraging the temporal dependencies inherent in the data, their approach effectively imputes missing values by capturing and utilizing the patterns observed in the time-series. Through extensive experiments on real-world wind energy datasets, the authors demonstrated the superiority of their method in accurately reconstructing missing data, thereby contributing to the field of wind energy data analysis and forecasting.

Zhang, et al. [41] proposed a deep learning-based imputation method for handling missing data in wind energy time series. Their approach utilized recurrent neural networks (RNNs) and long short-term memory (LSTM) networks to capture the temporal dependencies in the data and impute missing values. Through comprehensive experiments and comparisons with traditional imputation methods, the authors demonstrated the effectiveness and efficiency of their deep learning

approach, contributing to the advancement of wind energy data analysis and forecasting techniques.

In paper [20] is proposed a combined algorithm, which uses a mix of future extraction, imputation algorithm, and dynamic ANFIS (Adaptive Network-based Fuzzy Inference System) network for the prediction task. The basic idea of work [18] is to use advanced deep learning and transfer learning methods. The authors proposed a bi-directional missing data scheme based on LSTM (Long Short Term Memory). Paper [15] uncovered multiple imputation approach with Gaussian regression models and the expectation-maximization algorithm, which shows a set of advantages comparing with other methods. There is also a novel deep learning method based on conditional generative adversarial networks to tackle missing data issues [22]. All of these developments have proved the base of efficiency evaluation and implementation on wind energy data. The main drawback is a high computational cost of all above-mentioned approaches due to their expensive training stage with a wide range of experiments and different sets of hyperparameters optimization.

Chen, et al. [5] proposed a multiple imputation approach for handling missing wind power data using Bayesian networks. By modeling the relationships among different variables in wind power datasets, their probabilistic model effectively imputes missing values. The authors demonstrated the capabilities of their Bayesian network-based method in handling complex dependencies within wind energy data. This study contributes to the field by providing a novel approach for missing data imputation in wind power analysis.

Liu, et al. [17] investigated missing data imputation for wind speed using machine learning methods. Their study compared several algorithms, including support vector machines (SVM), random forests, and k-nearest neighbors (KNN), to impute missing values in wind speed datasets. Through rigorous evaluation and analysis, the authors provided insights into the performance and suitability of different machine learning techniques for handling missing data in wind energy analysis. This research contributes to the development of robust imputation methods for wind speed data in renewable energy studies.

Tawn, et al. [30] address the impact of missing data on wind power forecasts. It highlights the significance of accurate forecasts in the context of increasing renewable electricity generation and the variability of wind power production. The article emphasizes the need for a better understanding of missing data properties in wind farm operational data and its implications for forecasting. It discusses limitations in existing research and methods for dealing with missing data. The study proposes multiple imputations as a suitable approach for filling in missing data and evaluating forecast errors. Overall, the article contributes to the understanding of missing data in wind power time series and provides insights for improving forecast accuracy.

On the other hand (the second branch), there is still useful and efficient several classic statistical methods, such as [16], [19], [37]. The main idea is to hone missing

data imputation with k nearest neighbors, low-rank matrix, mean, median imputation. However, there is still a goal to deal with multiple gaps.

The following article [11] proposes a novel deterministic approach for imputing missing wind power data based on a wind speed-power model. The study validates the approach using 15 western wind datasets, demonstrating its effectiveness in minimizing the root mean square error. The model-driven imputation approach shows promise in accurately estimating wind power values and can contribute to improved planning, design, and maintenance of wind farms. At the same time, key points of this paper are: wind data is crucial for wind farm planning, design, and maintenance, but missing observations pose a challenge; the article presents a deterministic approach to wind power data imputation based on the wind speed-power model; the proposed approach is validated using 15 western wind datasets with capacities ranging from 25 to 40 MW; the results show that the approach minimizes the root mean square error, indicating its effectiveness in imputing missing wind power values; this model-driven imputation approach has the potential to enhance decision-making and improve the reliability and efficiency of wind power analysis.

In addition, we plan to consider some countries based on which the authors want to show some examples of using wind energy datasets with time-series data. Nowadays in the globalized world, there is an ongoing process of evaluation of development level of different countries [4], [10], [7], [6], creating enough wider possibilities for exchange knowledge processes and increasing competitiveness of these subjects on both domestic and foreign markets [23].

In this paper, we investigated a task of multiple data imputation. Unlike the prior works, we consider a statistical approach for defining multiple missing values applied on wind energy datasets with time-series data [3].

3. Primary (exploratory) data analysis for Wind Energy Datasets

For this stage, we took some well-known and free opened datasets, which connected to wind energy directly. Based on these examples we could show a few achievements that can make enough easy with clean datasets and to obtain great results of primary or exploratory data analysis [33], [27].

The first dataset [36] is presented data values of a wind turbine [38] in Turkey (this country was been selected only for an example) in 2018 and about the following indicators: wind speed, wind direction, generated active power for 10 minutes intervals. Generally, it has 50 530 observations. Common distribution in 3D-space for these three indicators (wind speed (m/s), wind direction (degrees), generated power (kW)) is presented in Figure 1.

As you can see, two noticeable crests exist in Figure 1, which have two directions about 45-55° (Nord-East) and about 195-205° (South-South-West). These crests show a sharp increase of wind speed and, as a consequence, an increase of generated power [2]. In addition, we can suggest these wind directions are prevailing directions in this specific place of Turkey. Additionally, after some mathematical transformations, we can obtain the following monthly and quarterly dependencies

(Figure 2a and Figure 2b). Based on these few first stages we can obtain some primary information about distributions in our datasets.

Analysis of these distributions can say that the biggest values of generated active power can be obtained in November and March and, generally, in the first quarter of the year, thus this dataset confirms some seasonality, which can be to explore in the future.

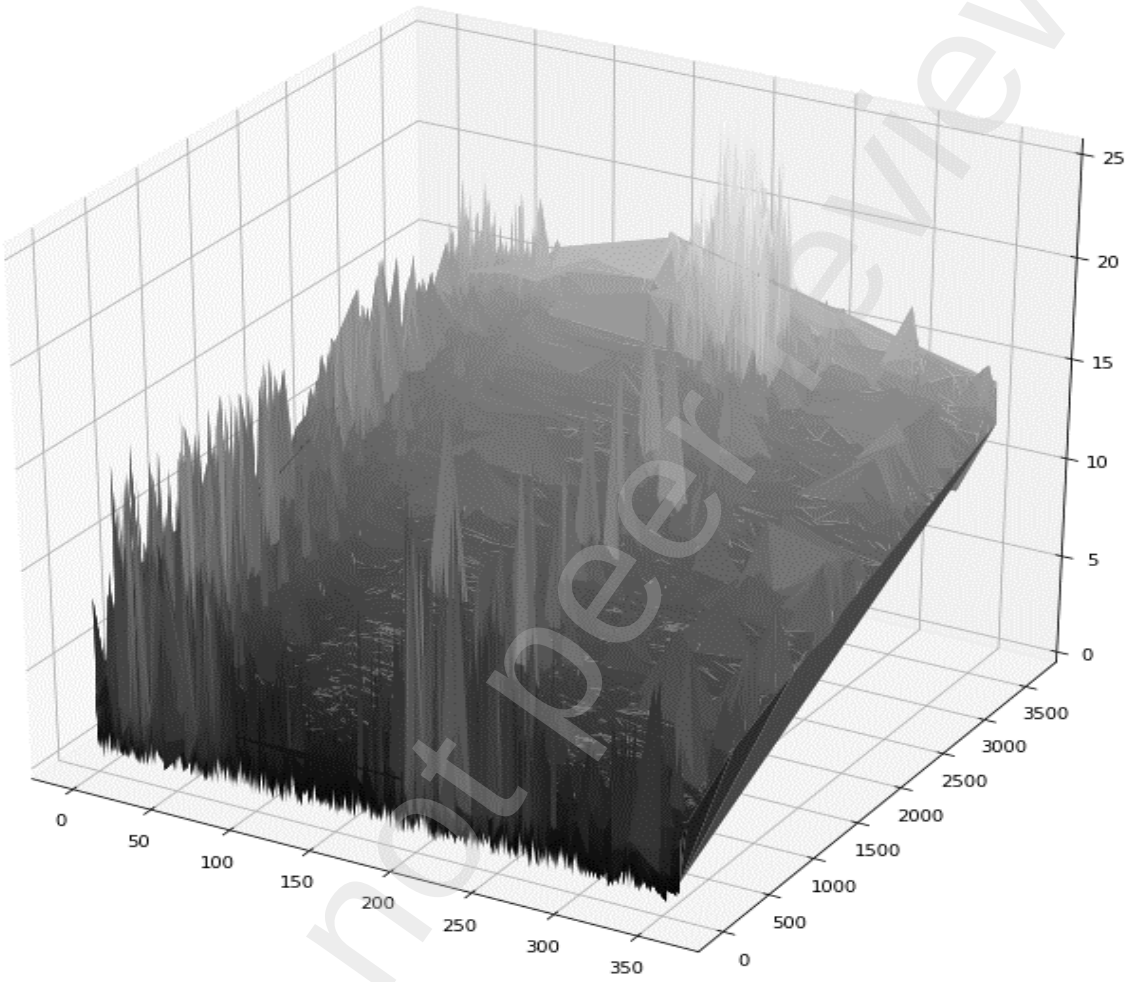


Figure 1. 3D-distribution for three indicators: wind speed (0 ÷ 25), wind direction (0 ÷ 360), generated active power (0 ÷ 4000)

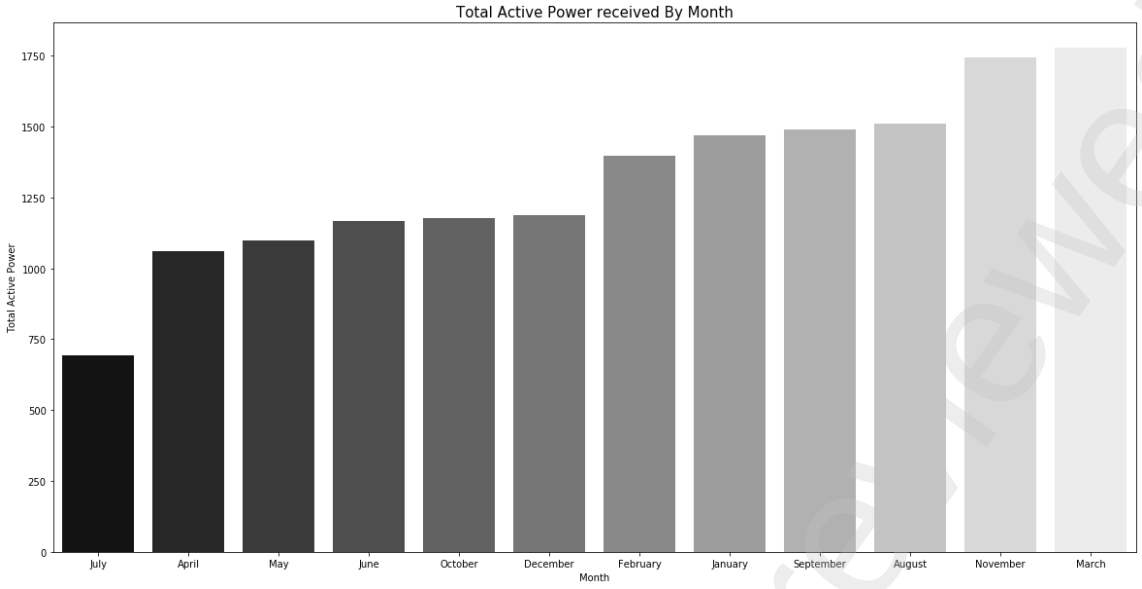


Figure 2a. Distributions for the monthly dependency of generated active power

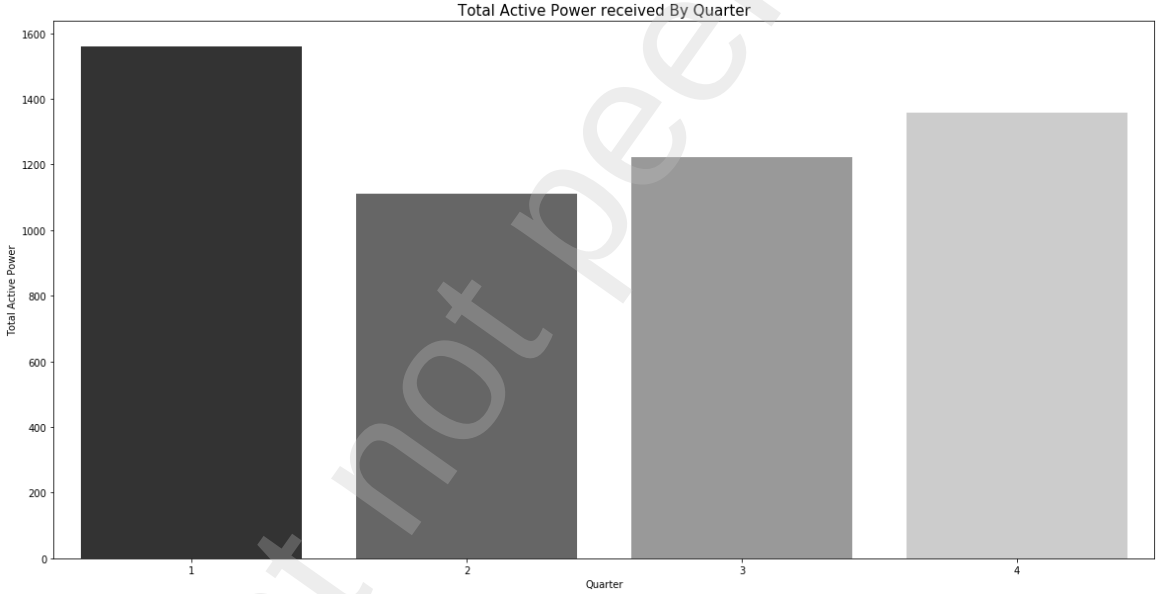


Figure 2b. Distributions for the quarterly dependency of generated active power

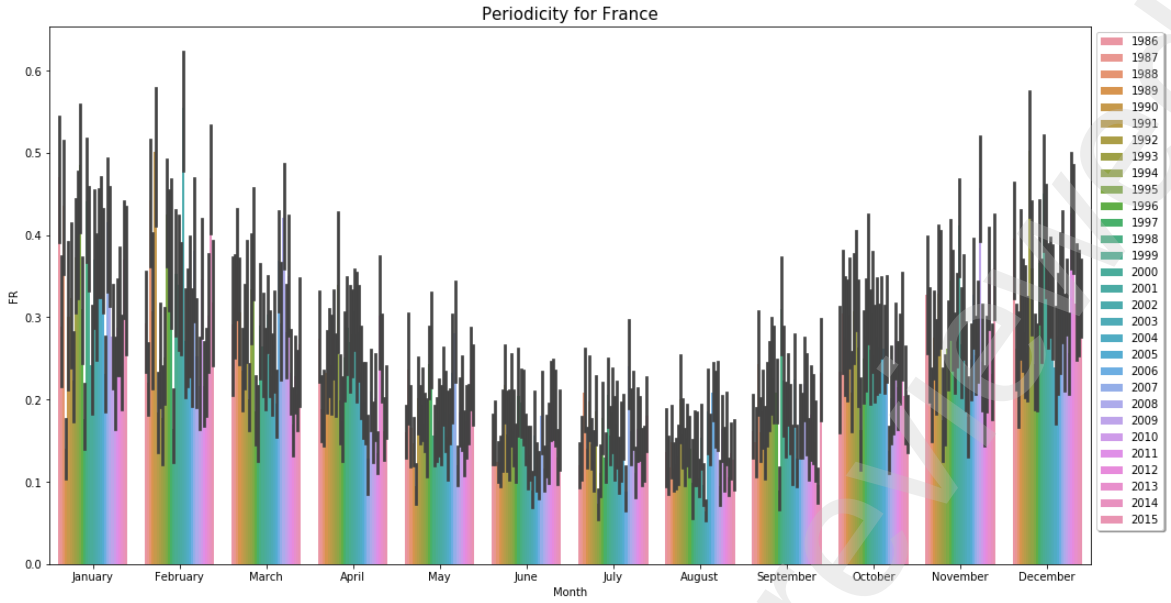


Figure 3. Distributions of the periodicity of active power value for France

Figure 4 shows the distribution of wind power energy value in hourly sectional for Estonia.

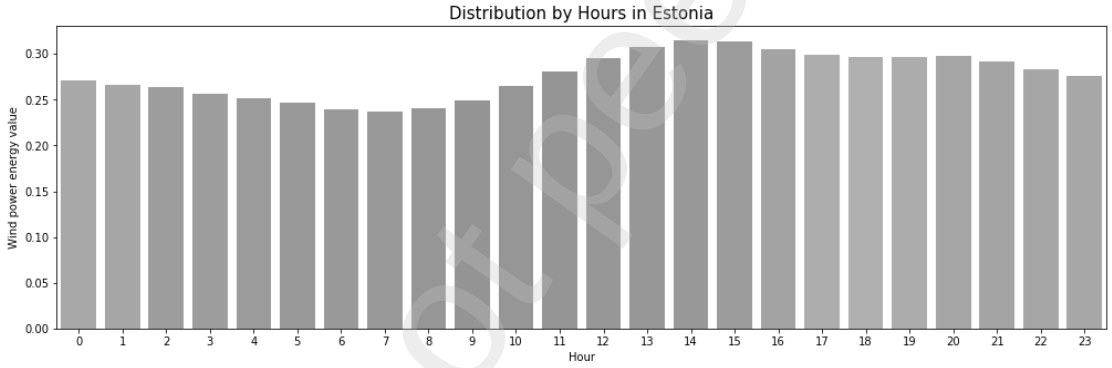


Figure 4. Distributions of wind power energy value in hourly sectional for Estonia

The second dataset [36] is presented data values of generated active power for hour intervals. It is opened and free access dataset that shows active power value (as a percent of the theoretical possible value at the normal working) for all EU countries (36 countries). It has 262 968 observations, but after hour grouping, we obtained 10 957 observations (by days). This set of observations covers the time period. This distribution confirms that during night time the value is lowest that at day, it indicates about the best possible energy efficiency, perhaps, it can be connected with a location of turbines from 1986 to 2015. This dataset describes only one indicator, but for a few dozens of countries and enough biggest timeline. Next step. We will show some results from different countries.

For example, Figure 3 shows us a distribution of periodicity of active power value for France (here and further the authors will show obtained results based on different countries of EU, it allows us to embrace as many countries as possible). As you can see, the unfavorable time period is the middle of the year, approximately,

from April to September, at the same with a low variation country. This similar primary analysis allows us to choose optimal parameters (time of usage, duration, etc.) for applying to the turbines and, thereby to improve energy efficiency.

Next Figure 5 can confirm some seasonality during a few years, which is equal to about 5 years. It can be an excellent background and start point for the next researches.

Figure 6 shows several distributions for the Federal Republic of Germany (DE), Iceland (IS), and Bulgaria (BG) quarterly and connecting with week number of year.

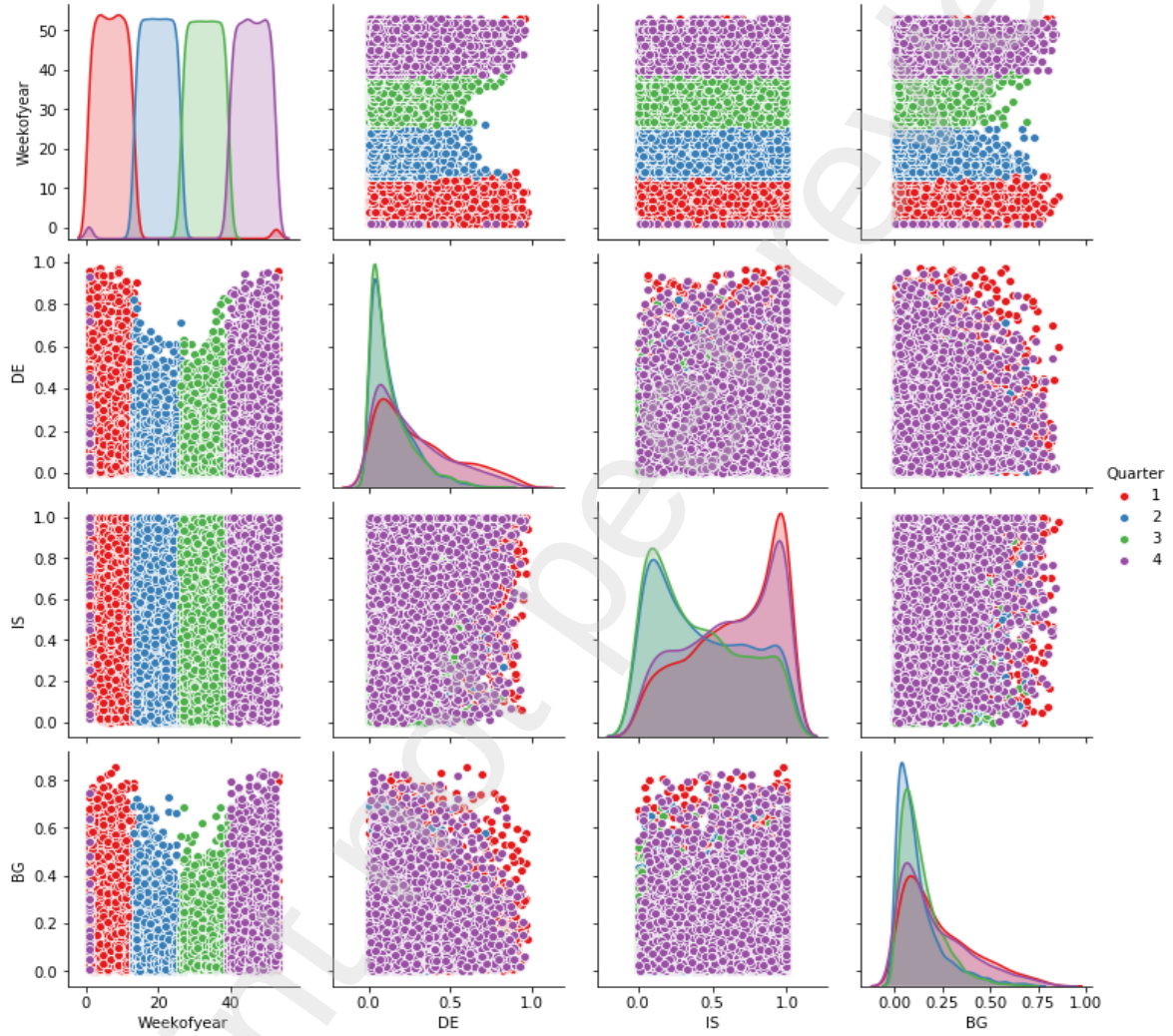


Figure 6. Distributions of active power value at quarterly for the Federal Republic of Germany, Iceland, and Bulgaria

As you can see on Figure 6, for instance, in the Federal Republic of Germany during the middle of the year, i.e., throughout about 15-43 weeks of the year, the active power value is subsiding a little (also as Bulgaria), so we can suggest that in this time the government of this country should be more careful with power consumption. From the other hand, Iceland hasn't this similar problem, because a level of active power value is constantly during all year. Moreover, we can observe some imbalance of active power value during the year by quarters, for example, the Federal Republic of Germany and Bulgaria have an imbalance, in the first part of the

year (the first quarter) they have more active power value than the end of the year, thus perhaps, in this case, they need to accumulate energy for the future periods.

In addition, the authors carried out research and calculated correlations among all 36 countries, the results of which are presented in Figure 7.

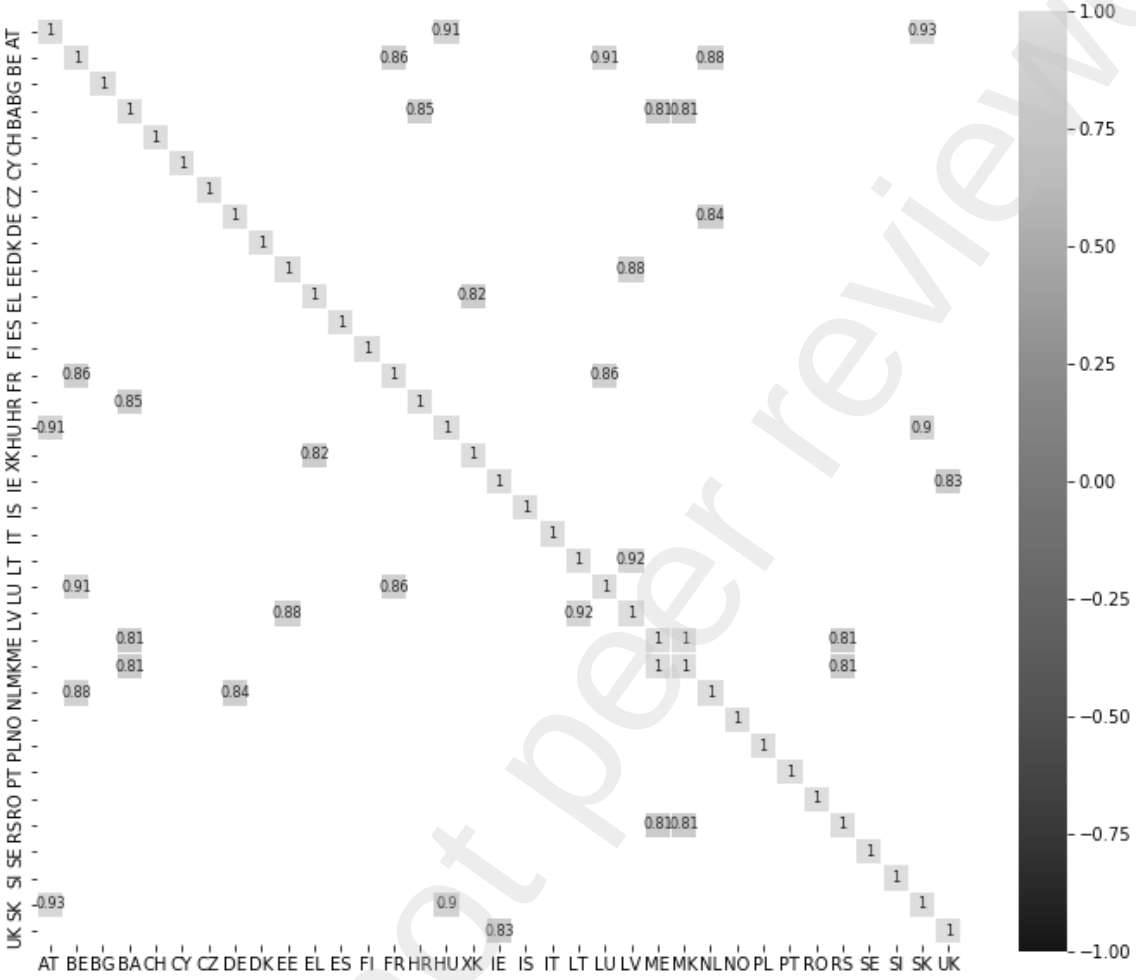


Figure 7. Correlation matrix of wind power energy value for all countries with highlighting of the bigger and lower values

From this matrix (Figure 7) you can see the most positive correlations which more 0.8 (was selected by authors). Among other countries, which have enough big correlations (as positive and negative) we can select the following: Austria (AT), Belgium (BE), Slovakia (SK), Hungary (HU), etc.

This distribution (Figure 7) can help to understand which countries correlate (or possible depend) each from other. In addition, it can prompt to which agglomerate this or that country belongs, or which area of neighboring countries is covering the common development. For instance, we can see that Makedonia (MK) and Montenegro (ME) have the mutual correlation is 1.0, so, probably these countries have a common and general way (strategic plan) of development. Also, we can observe the following common features for: Lithuania (LT) and Latvia (LV) – 0.92; Slovakia (SK) and Austria (AT) – 0.93; Austria (AT) and Hungary (HU) – 0.91;

Luxembourg (LU) and Belgium (BE) – 0.91; Hungary (HU) and Slovakia (SK) – 0.9, etc.

Finally, the authors calculated all distributions for all countries of a density of active power value (Figure 8). These results indicate to us that the most countries have the same distribution of density, except for the following: Norway (NO), Ireland (IE), the United Kingdom (UK), and, probably, Finland (FI). The most specific among them is Norway (NO), perhaps the usage and reproduction of energy puts on a priority basis.

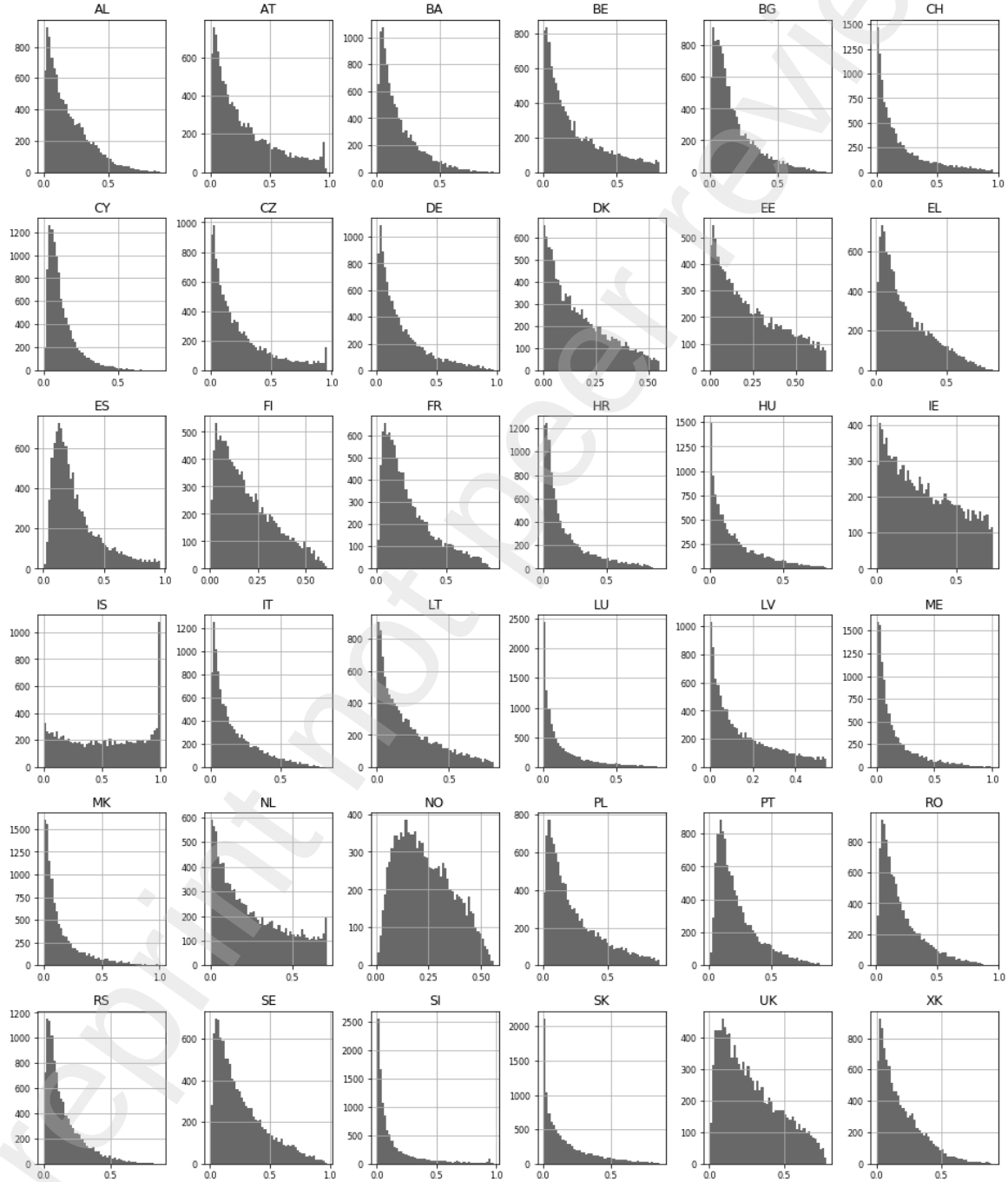


Figure 8. Histograms of distributions of power energy value density for all countries

There is just an example of what kind of dependencies and distributions we can obtain from datasets, if they are fully and clearly.

4. Imputation stage of missing data based on proposed new methods

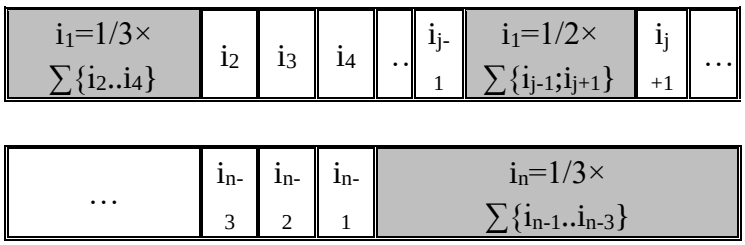
Missing data introduces an element of ambiguity in data analysis. They can affect the properties of statistical estimators, such as deviations or percentages, resulting in loss of performance and inaccurate (false) conclusions. There are several methods for replacing missing ones (gaps). This process is usually called the "imputation of missing data".

Analysis of existed methods of imputation has been made by authors earlier [12]. However, anyone from them cannot to full power in our case with different numbers of gaps [25]. Unfortunately, as it has shown a practice, existed methods of imputation can work only for separated and simplest cases. As it became known the well-known imputation methods, in particular, the hot deck method (chosen as the best alternative [12]) may be ineffective in filling large gaps. Therefore, author's imputation methods have been developed for each of the five cases (by a number of gaps) in which the following gap locations are distinguished: 1. gap in the left in the begin of the row, i.e. data gaps begin with the first-row index; 2. the gap on the right at the end of the row, i.e. data gaps end with the last-row index; 3. a gap inside of the row, i.e. cell gaps between the first and the last indices of the row; it's the hardest case.

Evaluating a model with many missing data sets can significantly affect the quality of the result. Some algorithms assume that the entire sample is complete and it only works if there are no data gaps. The Missing Data Study was formalized by Rubin [24] with the concept of a missing mechanism, in which the missing data are random variables and have a distribution. There are four general "mechanisms of absence" that go from the simplest to the most general. Rubin [24] identified missing data based on three mechanisms of absence.

4.1. Imputation of the single gaps

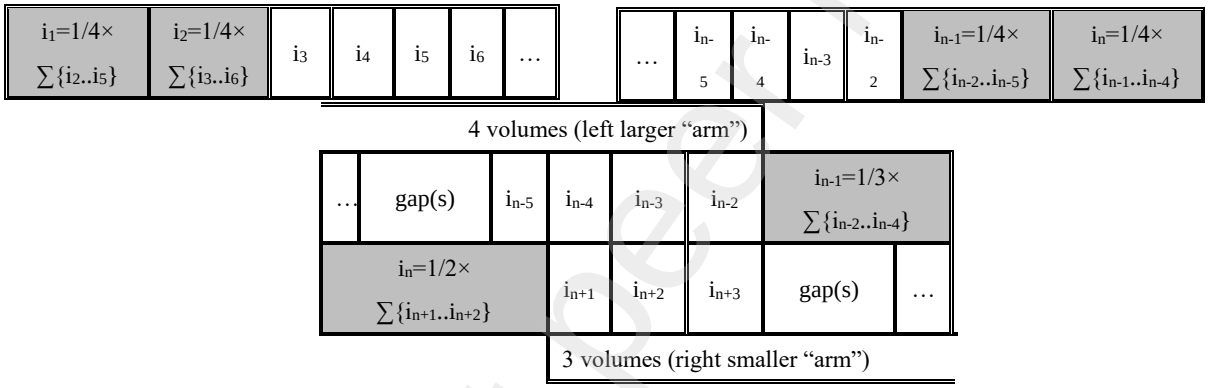
In order to use as much more input data as possible, the lines with a maximum number of gaps size up to five cells are stored in the study. The above-known imputation methods, in particular, the hot deck method (was been chosen as the best alternative in [25]) may be ineffective in filling large gaps. The following rules our own imputation methods for each gap size in the case of a location at each cell: 1. if the location of the single gap is to the left (the first cell in the row), the missing values are calculated as the arithmetic means of the three subsequent cells; 2. if the location of the single gap is in the middle (any inner cell in the row), the missing values are calculated as the arithmetic means of the adjacent two cells; 3. if the location of the single gap is to the right (the last cell in the row), the missing values are calculated as the arithmetic means of the three previous cells concerning the current one. For a closer look, a diagram of our proposed method of the imputation of the single gaps is given (Figure 9). Hereafter, the illuminated cells obtain the values based on other (existed) data (cells).



345 **Figure 9.** Diagram of the imputation of the single gaps

346 *4.2. Imputation of the double gaps*

347 If the location of the double gaps is to the left (the first two cells in the row), the
348 missing values for both cells are calculated as the arithmetic mean of four
349 subsequent cells in relation to the current one. However, these values have
350 calculated in the mode "step-by-step", in order from the right to the left alternately
351 (Figure 10, up).
352



353 **Figure 10.** Diagram of the imputation of the double gaps
354

355 If the location of the double gaps is to the right (the last two cells in the row), the
356 missing values for both cells are calculated as the arithmetic mean of the four
357 previous cells in relation to the current one. However, these values have calculated
358 in the mode "step-by-step", in order from the left to the right alternately (Figure 10,
359 in the center).
360 If the location of the double gaps is in the middle (any two inner arranged row
361 cells), the weight of the "arms" (the number of non-empty cells at the left and at the
362 right of the gap considered until the first gap appeared) is estimated. The missing
363 values for the cell next to the larger "arm" are calculated as the arithmetic mean of
364 the three closest cells in the array of this "arm". The missing values for the cell next
365 to the smaller "arm" are calculated as the arithmetic mean of the two closest cells in
366 the array of this (smaller) "arm" (Figure 10, down).

367 *4.3. Imputation of the triple gaps*

368 If the location of the triple gaps is to the left (the first three cells in the row), the
369 missing values for all three cells are calculated as the arithmetic means of five
370 subsequent cells in relation to the current one. However, these values have

calculated in the mode “step-by-step”, in order from the right to the left alternately (Figure 11, up).

If the location of the triple gaps is to the right (the last three cells in the row), the missing values for all three cells are calculated as the arithmetic mean of the five previous cells in relation to the current one. However, these values have calculated in the mode “step-by-step”, in order from the left to the right alternately (Figure 11, in the center).

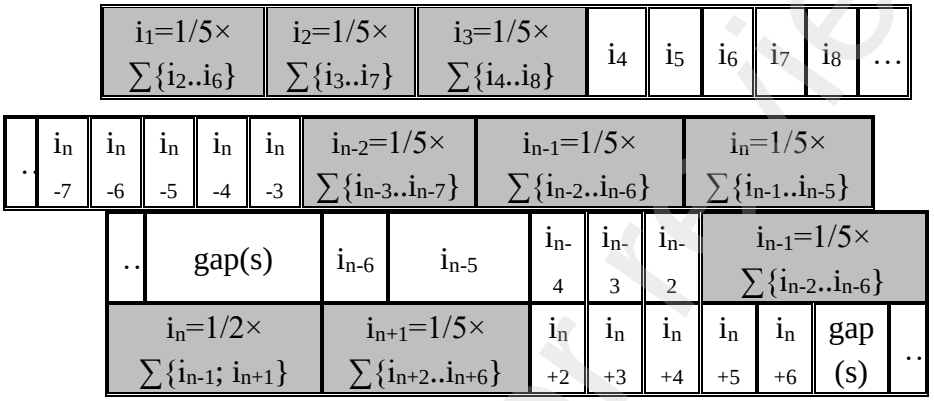


Figure 11. Diagram of the imputation of the triple gaps

If the location of the triple gaps is in the middle (any three inner consecutive cells in the row), the missing values are calculated as follows: 1. on the first step: the value of the left (the first) one cell of the gap is calculated as the arithmetic means of the five previous cells; 2. on the second step: the value of the right (the third) one cell of the gap is calculated as the arithmetic means of the five next cells in relation to the current one; 3. on the third step: the value of the middle (the second) one cell of the gap is calculated as the arithmetic means of the both previous (the first) and subsequent (the third) cells. i.e., as for the imputation of the single gaps (Figure 11, down).

4.4. Imputation of the quadruple gaps

If the location of the quadruple gaps (i.e., consists of four consecutive cells) is to the left (the first four cells in the row), the missing values for all four cells are calculated as the arithmetic means of the five following cells in relation to the current one. However, these values have calculated in the mode “step-by-step”, in order from the right to the left alternately (Figure 12, up).

If the location of the quadruple gaps (i.e., consists of four consecutive cells) is to the right (the last four cells in the row), the missing values for all four cells are calculated as the arithmetic mean of the five previous cells in relation to the current one. However, these values have calculated in the mode “step-by-step”, in order from the left to the right alternately (Figure 12, in the center).

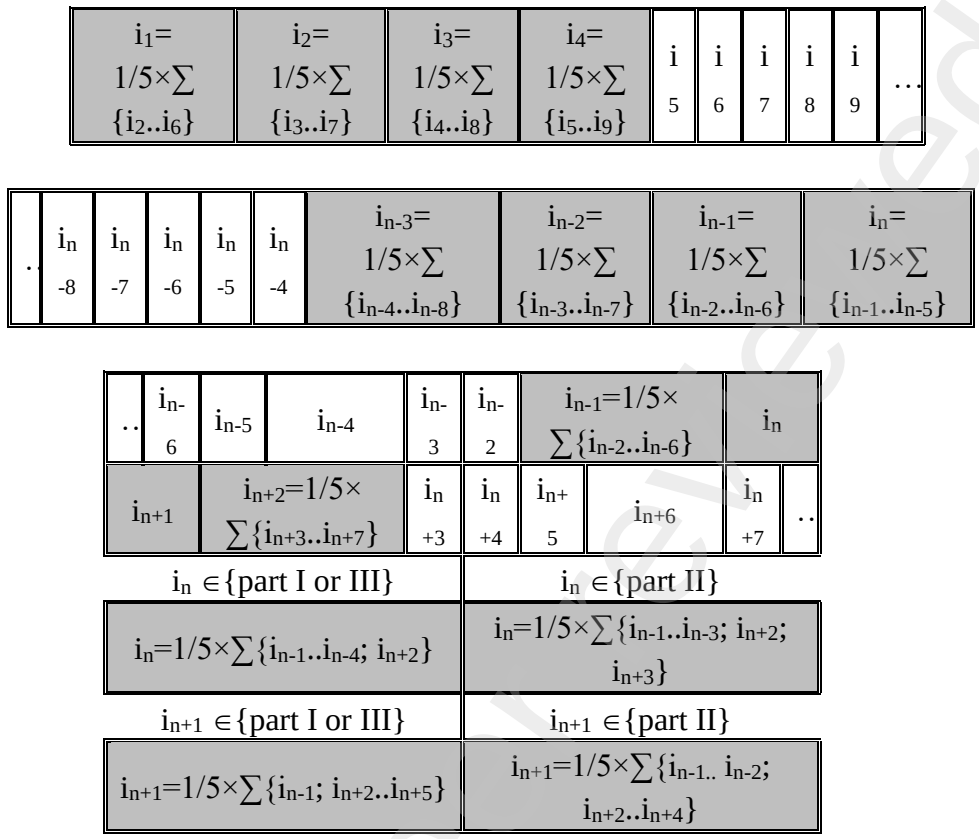


Figure 12. Diagram of the imputation of the quadruple gaps

If the location of the quadruple gaps (i.e. consists of four consecutive cells) is in the middle (any four internally consecutive cells in the row), the missing values for all cells are calculated as follows (Figure 12, down): 1. the missing values of the left (the first cell in our quadruple gaps space) gap cell are calculated as the arithmetic means of the five previous cells in relation to the current one; 2. the missing values of the right (the fourth cell in our quadruple gaps space) gap cell are calculated as the arithmetic means of the five following cells in relation to the current one; 3. the third number (one third) is located, in which the inner few cells of the quadruple gaps stand (in our case, there're the second and third cells). The whole row is divided into three thirds, i.e. (26 / 3 ≈ 9): cells with numbers 1-9 make the 1st part (I) of the third, cells with numbers 10-18 make the 2nd part (II) of the third, cells with numbers 18-26 make the 3rd part (III) of the third:

a) if the inner left (the second cell on account) gap cell is located in part I or III, its missing values are calculated as the arithmetic mean of the three cells before the gap cell, the 1st and 4th cells;

b) if the inner left (the second cell on account) gap cell is located in part II, its missing values are calculated as the arithmetic mean of the two cells before the gap cell, the 1st, 4th cells and one cell after the gap cell;

c) if the inner right (the third cell on account) gap cell is located in part I or III, its missing values are calculated as the arithmetic mean of the three cells after the gap cell, the 1st and 4th cells;

d) if the inner right (the third cell on account) gap cell is located in part II, its missing values are calculated as the arithmetic mean of one cell before the gap cell, the 1st, 4th cells, and two cells located after the gap cell.

4.5. Imputation of the pentuple gaps

If the location of the pentuple gaps (i.e., consists of five consecutive cells) is to the left (the first five cells in the row), the missing values for all five cells are calculated as the arithmetic means of the five following cells in relation to the current one. However, these values have calculated in the mode "step-by-step", in order from the right to the left alternately (Figure 13, up).

If the location of the pentuple gaps (i.e., consists of five consecutive cells) is to the right (the last five cells in the row), the missing values for all five cells are calculated as the arithmetic means of the five previous cells in relation to the current one. However, these values have calculated in the mode "step-by-step", in order from the left to the right (Figure 13, in the center).

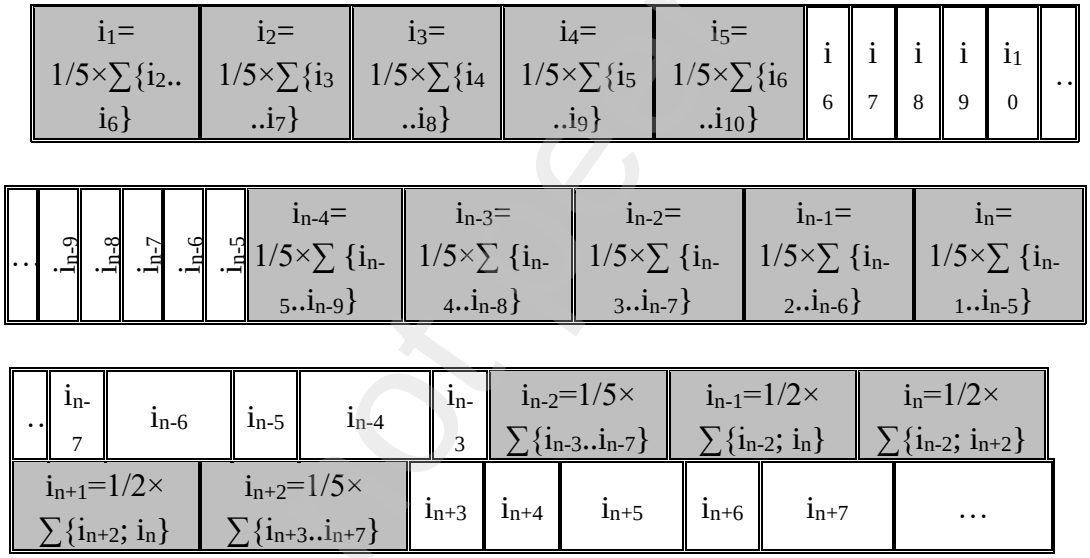


Figure 13. Diagram of the imputation of the pentuple gaps

If the of the pentuple gaps (i.e. consists of five consecutive cells) is in the middle (any inner five consecutive cells in the row), the missing values for the cells are calculated as follows (Figure 13, down): 1. the missing values of the left (the first cell in our pentuple gaps space) gap cell are calculated as the arithmetic mean of the five previous cells in relation to the current one; 2. the missing values of the right (the fifth cell in our pentuple gaps space) gap cell are calculated as the arithmetic mean of the five following cells in relation to the current one; 3. the missing values of the inner middle (the third cell in our pentuple gaps space) gap cell are calculated as the arithmetic mean of the left (the first cell in our pentuple gaps space) and the right (the fifth cell in our pentuple gaps space) gap cells; 4. the missing values of the inner left (the second cell in our pentuple gaps space) gap cell are calculated as the arithmetic mean of the left (the first cell in our pentuple gaps space) and inner middle (the third cell in our pentuple gaps space) gap cells; 5. the missing values of

the inner right (the fourth cell in our pentuple gaps space) gap cell are found as the arithmetic mean of the right (the fifth cell in our pentuple gaps space) and inner mean (the third cell in our pentuple gaps space) gap cells.

As a result of the data imputation stage, we validated our approach on EMHIRES dataset [8] and compared the value of mean error after reconstruction with forwarding and backward imputation.

5. Results

EMPHIRES is a dataset with wind power generation European Meteorological derived HIgh resolution RES generation time series for present and future scenarios [8].

It was conducted 10 experiments (cross-validation approach) with 100 randomly corrupted data positions. The obtained results with the average error by all experiments are represented in Table 1. It was explored single and multiple gaps with shoulder's size range from 1 to 5.

TABLE 1. Quantitative results (average error for missing data reconstruction, %) with data imputation

Gap's number	Shoulder's size	Proposed approach	Backward imputation		Forward imputation	
			Current value	Improv. percentage	Current value	Improv. percentage
1	1	13.449	25.660	47.59	22.893	41.25
2	2	20.33	22.65	10.3	19.58	-3.79
2	3	16.464	30.436	45.91	34.723	52.59
3	3	32.310	50.707	36.28	53.736	39.87
4	5	51.85	44.29	-17.05	45.4	-14.19
5	5	63.537	72.427	12.27	82.014	22.53

Table 1 shows the priority of our approach to other alternatives. Some selected cases have been shown in the columns "Gap's number" and "Shoulder's size". Improvement percentage is calculating as relation between results from well-known method and our proposed approach. As can be seen, the first experiment with a single gap demonstrates outperforming in 47.59% compared with backward imputation and 41.25% with forwarding imputation respectively. Second and third cases outperform at least 36%. Experiment with 5 gaps shows better results than other techniques in about 12% of improvement. To sum up our validation results we can conclude that mean error improvement is 37.29%.

6. Discussions

The calculated author's results based on available datasets [8] and [36] are enough plausible. However, the authors also agree that those results can require further validation and clarification by conducting new tests and their corresponding

statistical confirmation. Even that success, it is undeniable that the team of authors suggested the solving algorithm the existed problem of selected objects in the Big Data environment.

In addition, obtained results and proposed authors approaches can be used at forming some different reports, such as the following [32], [35], [31].

7. Conclusions

The conducted analysis of the literature sources confirms the relevance of the formed research to develop multiple imputation method in countries ranking working with long-term time-series data. It was revealed the need for further research and development of new methods and approaches for the statistical processing of multimodal data.

The scientific novelty of this work is an improved approach of data imputation by developing the appropriate algorithm, which can work with single, double, triple, quadruple, pentuple gaps in time-series data.

The practical value lies in the possible application of the developed method to analyze and predict wind energy processes in the Big Data environment with incomplete data or for preparing long-term statistical reports with aggregation from different sources and handling blank values.

The prospect of further study is to investigate the opportunity of the categorical time series research and extend the proposed missing data imputation approach with implementation on images, text data etc.

Acknowledgments: Authors would like to thank the anonymous referees. Moreover, we will be glad of any kind of questions, proposals, and comments. Based on the author opinion, it would be great to get some support to this topic from a side of other scientists and researchers. Besides, author would like to thank for any offers about possible further collaboration or research in this or some other similar areas.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contributions: Conceptualization, S. Kavun and A. Zamula; methodology, S. Kavun; software, S. Kavun; validation, S. Kavun, A. Zamula; formal analysis, S. Kavun; investigation, A. Zamula; resources, S. Kavun; data curation, A. Zamula; writing—original draft preparation, A. Zamula; mathematical analysis, S. Kavun; visualization, S. Kavun; supervision, S. Kavun. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Adam, A.; Shapiai, M. I.; Chew, L.; Ibrahim, Z.; Jau, L.; Khalid, M.; Watada, J. A Two-Step Supervised Learning Artificial Neural Network for Imbalanced Dataset Problems. *International Journal of Innovative Computing, Information and Control (IJICIC)* **2012**, *Volume 8*, No. 5(A), pp. 3163-3172.

[2] Al-Addous, M.; Al Hmidan, S.; Jaradat, M.; Alasis, E.; Barbana, N. Potential and Feasibility Study of Hybrid Wind–Hydroelectric Power System with Water-Pumping Storage: Jordan as a Case Study. *Appl. Sci.* **2020**, 10, 3332. https://doi.org/10.3390/app10093332.

[3] Ancuti, M.-C.; Musuroi, S.; Sorandaru, C.; Dordescu, M.; Erdodi, G.M. Wind Turbines Optimal Operation at Time Variable Wind Speeds. *Appl. Sci.* **2020**, 10, 4232. https://doi.org/10.3390/app10124232

[4] Bandura, R. Measuring Country Performance and State Behavior: A Survey of Composite Indices. UNDP/ODS Background Paper, Office of Development Studies, New York, 2005. www.thenewpublicfinance.org/ background/measuring.pdf.

[5] Chen, L., Wang, Y., & Zhang, M. (2017). Multiple imputation of missing wind power data using Bayesian networks. Journal of Renewable and Sustainable Energy, 9(2), 023304. https://doi.org/10.1063/1.4977309.

[6] Daradkeh, Y.; Guryanova, L.; Kavun, S.; Klebanova, T. Forecasting the cyclical dynamics of the development territories: Conceptual approaches, models, experiments. *European Journal of Scientific Research* **2012**, *Volume 74(1)*, pp. 5-20.

[7] Farmer, J.; Foley, D. The economy needs agent-based modelling. *Nature* **2009**, *Volume 460(7256)*, pp. 685–686. https://doi.org/10.1038/460685a

[8] Gonzalez, A.I; Zucker, A.; Careri, F.; Monforti, F.; Huld, T.; Badger, J. EMHIRES dataset. Part I: Wind power generation European Meteorological derived HIgh resolution RES generation time series for present and future scenarios. (2016) EUR 28171 EN; https://doi.org/10.2790/831549. Available: https://setis.ec.europa.eu/EMHIRES-datasets.

[9] Gorokhovatskyi, V.A.; Zamula, A.A. Employment of Intelligent Technologies in Multiparametric Control Systems. *Telecommunications and Radio Engineering* **2016**, *Volume 75*, No 19, pp. 1775–1785. https://doi.org/10.1615/TelecomRadEng.v75.i19.60.

[10] Hajirahimova, M.; Aliyeva, A. Big data initiatives of developed countries. *Problems of Information Society* **2017**, *Volume 08*, pp. 10-19. https://doi.org/10.25045/jpis.v08.i1.02.

[11] Jha, S., Wang, J., Marina, N. (2022). A novel model-driven deterministic approach to wind power imputation. Sustainable Computing: Informatics and Systems. 36. 100818. https://doi.org/10.1016/j.suscom.2022.100818.

[12] Kavun S.; Zamula, A.; Miziurin, V. Intelligent Evaluation Method for Complex Systems in The Big Data Environment, 2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON), July 2-6, Lviv, Ukraine, 2019, pp. 951-957. https://doi.org/10.1109/ UKRCON.2019.8880024.

[13] Kavun, S. Conceptual fundamentals of a theory of mathematical interpretation. *Int. J. Computing Science and Mathematics* **2015**, *Volume 6*, No. 2, pp. 107–121. https://doi.org/10.1504/IJCSM.2015.069459.
[14] Lee, S. H., Park, Y. H., & Kim, W. (2019). Temporal pattern-based imputation for wind energy datasets. IEEE Transactions on Sustainable Energy, 10(2), 865-875. https://doi.org/10.1109/TSTE.2018.2852093.
[15] Liu, T.; Wei, H.; Zhang, K. Wind power prediction with missing data using Gaussian process regression and multiple imputation. *Applied Soft Computing* **2018**, *Volume 71*, pp. 905-916. https://doi.org/10.1016/j.asoc.2018.07.027.
[16] Liu, X.; Zheng, Z.; Zhang, Z.; Cao, Z. A statistical learning framework for the intelligent imputation of offshore wind farm missing SCADA data. Proceedings of the 8th Renewable Power Generation Conference (RPG 2019), Shanghai, China, 24-25 Oct. 2019. https://doi.org/10.1049/cp.2019.0615.
[17] Liu, Z., Jiang, H., & Du, P. (2016). Missing data imputation for wind speed using machine learning methods. Energy Procedia, 100, 428-433. https://doi.org/10.1016/j.egypro.2016.10.157.
[18] Ma, J.; Cheng, J. C.P.; Jiang, F.; Chen, W.; Wang, M.; Zhai. C. A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data. *Energy and Buildings* **2020**, *Volume 216*, Article 109999. https://doi.org/10.1016/j.enbuild.2020.109941.
[19] Martinez-Luengo, M.; Shafiee, M.; Kolios, A. Data management for structural integrity assessment of offshore wind turbine support structures: data cleansing and missing data imputation. *Ocean Engineering* **2019**, *Volume 173*, pp. 867-883. https://doi.org/10.1016/j.oceaneng.2019.01.003.
[20] Morshedizadeh, M.; Kordestani, M.; Carriveau, R.; Ting, D. S.-K.; Saif, M. Application of imputation techniques and Adaptive Neuro-Fuzzy Inference System to predict wind turbine power production. *Energy* **2017**, *Volume 138*, pp. 394-404. https://doi.org/10.1016/j.energy.2017.07.034.
[21] Panchenko, V.; Zamula, A.; Kavun, S.; Mikheev, I. Intelligent management of the enterprise personnel security system. 2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT), Kiev, 2018, pp. 469-474, https://doi.org/10.1109/DESSERT.2018.8409179.
[22] Qu, F.; Liu, J.; Ma, Y.; Zang, D.; Fu, M. A novel wind turbine data imputation method with multiple optimizations based on GANs. *Mechanical Systems and Signal Processing* **2020**, *Volume 139*, Article 106610. https://doi.org/10.1016/j.ymssp.2019.106610.
[23] Rogers, J.; Chong, H.-Y.; Preece, C.; Lim, C.C.; Jayasena, H.S. *BIM Development and Trends in Developing Countries: Case Studies*, Bentham Science Publishers, Sharjah, U.A.E., 2015.
[24] Rubin, D.B. Inference and missing data. *Biometrika* **1976**, 63(3), pp. 581–592. https://doi.org/10.1093/biomet/63.3.581.
[25] Sahri, Z.; Yusof, R.; Watada, J. FINNIM: Iterative Imputation of Missing Values in Dissolved Gas Analysis Dataset. *IEEE Transactions on Industrial*

Informatics **2014**, *Volume 10*, no. 4, pp. 2093-2102. https://doi.org/10.1109/TII.2014.2350837.

[26] Smith, J. D., Johnson, A. B., & Anderson, C. R. (2018). A comprehensive review of missing data imputation techniques in renewable energy datasets. Applied Soft Computing, 42, 112-125. https://doi.org/10.1016/j.asoc.2017.08.023.

[27] Smith-Miles, K. *Exploratory Data Analysis*. In: Lovric M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg, 2011.

[28] Tan, S. C.; Watada, J.; Ibrahim, Z.; Khalid, M. Evolutionary Fuzzy ARTMAP Neural Networks for Classification of Semiconductor Defects. *IEEE Transactions on Neural Networks and Learning Systems* **2015**, Volume 26, no. 5, pp. 933-950. https://doi.org/10.1109/TNNLS.2014.2329097.

[29] Tan, S.C.; Wang, S.; Watada, J. A self-adaptive class-imbalance TSK neural network with applications to semiconductor defects detection. *Information Sciences* **2018**, Volume 427, pp. 1-17. https://doi.org/10.1016/j.ins.2017.10.040.

[30] Tawn, R., Browell, J., Dinwoodie, I. (2020). Missing data in wind farm time series: Properties and effect on forecasts, vol. 189, 106640. https://doi.org/ 10.1016/j.epsr.2020.106640.

[31] The changing nature of work. International Bank for Reconstruction and Development / The World Bank, 1818 H Street NW, Washington, DC 20433, 2019.

[32] The Global Competitiveness Report 2019, World Economic Forum, 91-93 route de la Capite, CH-1223, Cologny, Geneva, Switzerland.

[33] Tukey, J.W. *Exploratory data analysis*. Addison-Wesley, Reading, MA, 1977.

[34] Vnukova, N.; Kavun, S.; Kolodiziev, O.; Achkasova, S.; Hontar, D. Indicators-Markers for Assessment of Probability of Insurance Companies Relatedness in Implementation of Risk-Oriented Approach. *Economic Studies* **2020**, issue 1, pp. 151-173. Available from: https://www.iki.bas.bg/node/4115/?width=600& height=400&iframe=true&ajax=1.

[35] Wendling, Z. A.; Emerson, J. W.; Esty, D. C.; Levy, M. A.; de Sherbinin, A., et al. 2018 Environmental Performance Index. New Haven, CT: Yale Center for Environmental Law & Policy, 2018. https://epi.yale.edu.

[36] Wind Turbine Scada Dataset. 2018 Scada Data of a Wind Turbine in Turkey. Dataset owner: Berk Erisen. Available: https://www.kaggle.com/berkerisen/wind-turbine-scada-dataset.

[37] Xie, Z.; Sun, X. Imputation of missing wind speed data based on low-rank matrix approximation. 2017 2nd International Conference on Power and Renewable Energy (ICPRE), Chengdu, China, 2017, pp. 397-401. https://doi.org/10.1109/ICPRE.2017.8390566.

[38] Yang, D.; Jin, E.; You, J.; Hua, L. Dynamic Frequency Support from a DFIG-Based Wind Turbine Generator via Virtual Inertia Control. *Appl. Sci.* **2020**, 10, 3376. https://doi.org/10.3390/app10103376.

[39] Zamula, A.; Kavun, S. Complex systems modeling with intelligent control elements. *International Journal of Modeling, Simulation, and Scientific Computing* **2017**, *Volume 8*, No. 1. https://doi.org/10.1142/S179396231750009X.
[40] Zamula, A.; Kavun, S.; Serdukov, K. Binary Recommender System with Artificial Intelligence Aids, 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), Kharkiv, Ukraine, 2019, pp. 251-255. https://doi.org/10.1109/PICST47496.2019.9061502.
[41] Zhang, Q., Wang, C., & Liu, Y. (2020). Deep learning-based imputation for missing data in wind energy time series. Renewable Energy, 147(Part 1), 872-882. https://doi.org/10.1016/j.renene.2019.09.044.