International Journal of Information Security (2022) 21:1299–1321
https://doi.org/10.1007/s10207-022-00612-8

REGULAR CONTRIBUTION

DFTMicroagg: a dual-level anonymization algorithm for smart grid data

Kayode S. Adewole[1,2] · Vicenç Torra[1]

Published online: 7 September 2022

Abstract
The introduction of advanced metering infrastructure (AMI) smart meters has given rise to fine-grained electricity usage data at different levels of time granularity. AMI collects high-frequency daily energy consumption data that enables utility companies and data aggregators to perform a rich set of grid operations such as demand response, grid monitoring, load forecasting and many more. However, the privacy concerns associated with daily energy consumption data has been raised. Existing studies on data anonymization for smart grid data focused on the direct application of perturbation algorithms, such as microaggregation, to protect the privacy of consumers. In this paper, we empirically show that reliance on microaggregation alone is not sufficient to protect smart grid data. Therefore, we propose DFTMicroagg algorithm that provides a dual level of perturbation to improve privacy. The algorithm leverages the benefits of discrete Fourier transform (DFT) and microaggregation to provide additional layer of protection. We evaluated our algorithm on two publicly available smart grid datasets with millions of smart meters readings. Experimental results based on clustering analysis using k-Means, classification via k-nearest neighbor (kNN) algorithm and mean hourly energy consumption forecast using Seasonal Auto-Regressive Integrated Moving Average with eXogenous (SARIMAX) factors model further proved the applicability of the proposed method. Our approach provides utility companies with more flexibility to control the level of protection for their published energy data.

Keywords Smart meters · Smart grid · Privacy preserving · Discrete Fourier transform · Microaggregation · Load forecasting

1 Introduction

Over the last few decades, the conventional electricity grid has been in existence, which consist of power generation and distribution systems. The conventional grid provides electricity to consumers with monthly billing arrangements. This type of grid is characterized by one-way communication and there is lack of interaction between the customers and the utility company. This leads to different issues that include loss of energy and poor peak load management [1,2]. Nevertheless, the advancement in technology over the years has brought about the rollout of advanced metering infrastructure (AMI) smart meters that improved the traditional energy grid. AMI offers advantages such as effective communication between consumer and utility, increased reliability, resilience and better control of demand response load management [3,4]. With the advancement in smart grid technology, the collection of fine-grained daily electricity usage data with different levels of time granularity has rapidly grown. The fine-grained electricity consumption data has enabled utility companies to perform robust grid operations such as demand response, grid monitoring, consumer profiling, customer segmentation, energy usage prediction, load forecasting and many more [5,6].

Due to the benefits offered by AMI smart meters, the European Union (EU) planned to install 225 million smart meters for electricity and 51 million for gas in the year 2024. In this year, it is expected that almost 77% of European consumers of electricity will have access to smart meters [7]. Similarly, the UK government planned to install 53 million smart meters while the USA plans to roll out 90 million smart meters as of 2020 [1,3]. As part of additional benefits, smart grid also enables consumers to actively manage their energy usage and

✉ Kayode S. Adewole
kadewole@cs.umu.se; adewole.ks@unilorin.edu.ng

Vicenç Torra
vtorra@cs.umu.se

[1] Department of Computing Science, Umeå University, Umeå, Sweden

[2] Department of Computer Science, University of Ilorin, Ilorin, Nigeria

control energy bills. Moreover, besides the use of electricity consumption data by utility companies, these data may be shared with third-party service providers and researchers to provide more insights on electricity consumption. However, fine-grained electricity consumption data has been characterized with privacy-sensitive consumer behaviors, which are capable of revealing general habits and lifestyles of households [4,8]. Consequently, sharing of fine-grained electricity usage data in its original form has been shown to violate the security and privacy of electricity customers.

Fine-grained electricity usage data are valuable and can be sought by many entities including attackers who want to deduce the type of device or appliance that was in use at any given time. There is a specific research field called non-intrusive load monitoring for appliance (NILMA), which relies on electricity consumption data to extract detailed information of consumers based on their domestic appliance usage patterns. The goal of NILMA research is to deduce the types of appliances used in a house along with their energy consumption based on a detailed analysis of the current and voltage of the total load [9,10]. The information obtained through this analysis is useful to third parties like marketers, law enforcement, and criminals [11,12]. For instance, the case of electricity blackout due to hacking has been reported in Ukraine in 2015, 2016 and January 2017 where hackers were able to shut down energy systems that supply heat and light to millions of households [1,13].

As a countermeasure against NILMA and re-identification or de-pseudonymization attacks, different solutions have been proposed, which include cryptographic approach, differential privacy, rechargeable battery for obfuscation of smart meter reading, data aggregation based on trusted third-party (TTP), and data anonymization and perturbation [1–3,6,10,12,14–16]. The cryptographic approach involves the development of encryption protocols to encrypt smart meter data at the point of generation so that it will be difficult to determine the specific household consumption. Cryptographic approach includes both the traditional and homomorphic encryption schemes [17,18]. By traditional encryption we refer to those encryption schemes that do not allow computation on encrypted data. This method can provide high level of security and privacy before transmitting the data to the utility company. However, it is not an efficient method for publishing energy data that are needed for research purposes and complex data analytics as no information is released in the published data for complex statistical analysis [19]. Differentially private (DP) algorithm has been used to publish electricity consumption data [6]. However, previous studies have observed that for high-dimensional time series data, DP often adds too much noise that can lead to unsatisfactory data utility [12,14].

Battery-based load hiding (BLH) has been proposed in [2,20]. The goal of BLH approach is to mask smart meter reading by utilizing a rechargeable battery. This approach has been mainly theoretic and its successful real-world application is yet to be developed [2]. Data aggregation based on TTP was proposed in [10]. This method relies on TTP for aggregation of smart meter reading. The aggregated reading is then transmitted to utility company for workload balancing and statistical analysis. However, as stated by the authors, this approach traded security for privacy; hence, practical application of data anonymization should be extended to improve this method. To provide data anonymization and perturbation of smart meter reading, [12,14] introduce PAD system. PAD directly applied microaggregation using k-ward algorithm to anonymize daily energy consumption data. However, in our study, we empirically show that reliance on microaggregation alone is not sufficient to protect smart grid data against disclosure risk.

In this paper, a dual-level anonymization algorithm, DFT-Microagg, is proposed to reduce the disclosure risk of microaggregation algorithm when used to protect energy data. To achieve this goal, we first conducted an experiment to ascertain the privacy value offered by microaggregation algorithm when used to protect smart grid data. Based on our findings, we extended this model by combining discrete Fourier transform (DFT) and microaggregation to improve privacy. We show that the proposed approach guarantees promising data utility by experimenting with three major data mining tasks based on clustering analysis using k-Means, classification via kNN algorithm and mean hourly load forecasting using SARIMAX model. In addition, we compute information loss (IL) to understand how much information is lost due to the dual-level perturbation process. To the best of our knowledge, this is the first paper to extensively investigate the application of DFT and microaggregation to smart grid data protection. Additionally, we investigate two record linkage attacks based on distance-based record linkage and interval disclosure risk on the protected smart grid data. Summarily, the following are the contributions of this paper:

– Investigate the actual privacy value offered by microaggregation for protecting smart grid data.
– Propose a dual-level anonymization algorithm, which combined DFT with microaggregation.
– Implement two adversarial models using distance-based record linkage and interval disclosure risk. Specifically, we propose distance-based record linkage algorithm which does not only consider the nearest record to the masked data being linked but also the second nearest record.
– Conduct extensive experiments on smart grid data with millions of smart meter readings.

The remaining parts of this paper are organized as follows: Section 2 discusses related works on smart grid data protection. Section 3 provides a detailed information on k-Anonymity and attack model assumed in the previous work for protecting smart grid data. Section 4 presents the proposed approach in this paper as well as the adversarial models considered in our study. Section 5 focuses on experimental setup, and Section 6 presents results and discussion. Finally, Sect. 7 concludes the paper and highlights future research direction.

2 Related work

The literature on privacy-preserving data publishing is vast and different research domains have been extensively studied. [21] presented an algorithm to publish dynamic datasets and compared their results with maximum distance to average vector (MDAV) microaggregation algorithm. Microaggregation procedure has also been extended to time series data in [22] where the authors evaluated the performance of two distance metrics: Euclidean distance and Short Time Series (STS) distance. An empirical comparison of disclosure risk control methods for microdata has been extensively studied [19]. [23] presented the foundation, new development and challenges of data privacy preserving. Nevertheless, in the domain of smart grid, privacy-preserving energy data has been studied from different dimensions. These include methods based on cryptography, differential privacy, BLH, data aggregation based on TTP, data compression, and data anonymization and perturbation [1].

Cryptographic methods involved the development of encryption protocols to encrypt smart meter data at the point of generation so that it will be difficult to determine the specific household consumption from the data. This method can provide some level of security and privacy before transmitting to the utility company. For instance, [16,17] proposed similar approaches based on symmetric encryption algorithms and hashing. In these methods, lightweight cryptographic protocols encrypt smart meter data before transmission to the utility company. Similarly, cryptographic approach that allows computation on encrypted data based on homomorphic schemes have also been studied [18,24]. The major challenge with cryptographic methods when used for privacy-preserving data publishing is that no information is released in the published data for research purposes [19]. Therefore, it is not a suitable method for publishing smart grid data that requires complex statistical analysis.

Differentially private (DP) algorithms have been studied for smart grid data [6,9]. However, previous studies have observed that for high-dimensional time series data, DP often adds too much noise that can lead to unsatisfactory data utility [12,14]. BLH has been proposed in [2,20]. The goal of BLH is to install a battery at the consumer end, which can be charged or discharged to make the electricity meter incapable of precisely obtaining the consumption data of electric appliances and to obfuscate the actual consumption of the electric appliances [25]. This masking method is mainly theoretic and its empirical validation for real-world application is still a major concern [2].

[10] proposed data aggregation method that relies on TTP aggregation of smart meter reading. This approach assumed that utility companies only need to protect data that is collected at high-frequency (HF) without attributing to specific consumers while the low-frequency (LF) smart meter data are transmitted to TTP for aggregation. However, as stated by the authors, this approach traded security for privacy; hence, practical application of data anonymization should be extended to improve this method. A similar assumption was made to evaluate the performance of de-anonymization algorithms in [8,26].

Data compression of smart meter reading has been investigated. The idea is that storage requirement and transmission overhead can be greatly reduced using data compression algorithms. [27] conducted an extensive study of the effect of applying different compression algorithms on smart meter data. The algorithms investigated are wavelet transform, symbolic aggregate approximation (SAX), principal component analysis (PCA), singular value decomposition (SVD), dimensionality reduction via linear regression, Huffman coding and Lempel–Ziv (LZ) algorithm. Nevertheless, this study established that finding an appropriate balance between efficiency and loss ratio is not a trivial issue when applying compression algorithms on smart meter data. Similar findings have also been presented in [28,29] based on smart meter data compression.

Generative adversarial network (GAN) and additive correlated noise have been studied to protect smart meter consumption data [30,31]. One of the benefits of GAN is its ability to model the uncertainties of original data and based on this model a new data is generated, which can be used for grid operations such as planning and scheduling. Two deep neural networks are usually trained: one to capture the distribution of the data and the other to estimate the probability that the input originates from the real data. This approach is promising to protect energy consumption data; however, its capability to prevent disclosure risk attacks is missing in the literature.

Smart grid and building occupancy data publishing system (PAD) was proposed in [12,14]. This approach follows k-anonymity, which is assumed to guarantee some level of privacy. K-anonymity has received a wide range of attention as one of the suitable conditions that data protection algorithms must satisfy to prevent record linkage. In PAD [14], a linear distance metric was learned to determine data user's specific task. A modified version of this approach was presented in [12] where a nonlinear distance metric learning was formulated based on a deep neural network. The goal of PAD is to learn user's specific task by asking data analyst to manually annotate energy data to determine the specific data utility that satisfies the data analyst objective. The annotated data are then passed to k-ward microaggregation algorithm for privacy protection. However, asking data users to manually annotate large time series energy data is not a trivial task. In this study, we show that reliance on microaggregation alone is not sufficient to protect daily energy consumption data against disclosure risk.

3 k-anonymity and attack model assumption

In this section, we briefly present the concept of k-anonymity as well as the attack model that was assumed in the previous work [12] for protecting energy data, which forms the basis for conducting our investigative study.

3.1 k-anonymity and microaggregation

k-anonymity is not a protection method on its own but a condition that protected data should satisfy to guarantee the privacy of the individual in the masked data. k-anonymity concept was originally proposed in the context of privacy protection for relational databases [32–34]. The goal of k-anonymity is to ensure that each individual in a protected data cannot be identified within a set of k individuals. This means that the dataset is partitioned into a set of at least k indistinguishable records. One way to enforce k-anonymity on the protected data is to use microaggregation algorithm [35].

Generally, microaggregation protects dataset using two steps: k-partition and aggregation. Suppose X represents the input data to be protected and $\hat{X}$ is the protected data after applying microaggregation. The two steps are described as follows:

Step 1 (k-partition): All records in X are partitioned into different clusters, say g, with each consisting of k or more records.

Step 2 (aggregation): Compute a representative (i.e., centroid) for each of the clusters in g and use this centroid to replace the original records in the cluster. This means that all the k records in the cluster are replaced with the same value; hence, k-anonymity is guaranteed.

At the k-partitioning step, it is important to ensure that the in-group distance between cluster element and its centroid is minimized. This is to enforce homogeneity to minimize information loss. To achieve this, the sum of squared error (SSE) criterion in Eq. (1) is minimized. Formally, let u_{ij} describes the clustering of records in X such that $u_{ij} = 1$, if record j is assigned to the ith cluster. Suppose v_i is the centroid of the ith cluster, then homogeneity is enforced by,

$$Minimize\ SSE = \sum_{i=1}^{g}\sum_{j=1}^{n} u_{ij}(d(x_j, v_i))^2 \tag{1}$$

$$Subject\ to\text{:} \sum_{i=1}^{g} u_{ij} = 1\ \forall\ j = 1, 2, \ldots, n$$

$$2k \geq \sum_{j=1}^{n} u_{ij} \geq k\ \forall\ i = 1, 2, \ldots, g$$

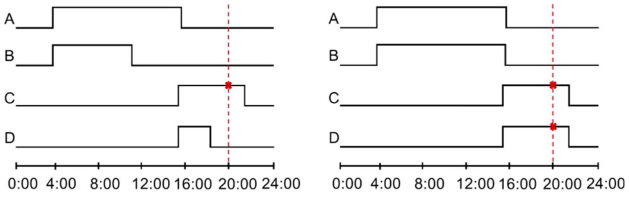
$$u_{ij} \in \{0, 1\}$$

If X is numerical, Euclidean distance is mostly chosen to estimate the distance metric $d(x, v)$ in Eq. (1). Several versions of microaggregation algorithm have been studied in the literature, which includes maximum distance (MD), maximum distance to average vector (MDAV), variable-size maximum distance to average vector (V-MDAV) and k-ward [12,35–37]. In this study, we implemented MDAV as additional layer to DFT due to its performance and wide adoption in the literature [36]. MDAV is described in Algorithm 1 as adapted from [37].

3.2 Attack model assumption

This section presents the attack model assumed in the previous study [12] for protecting energy data. This forms the basis for conducting our investigative study to ascertain the actual privacy value offered by k-anonymity and microaggregation when used to protect energy data. For the sake of clarity, suppose we have energy data where each record (row) is daily energy consumption of a particular household or consumer that has been sampled at a specific time interval (e.g., 1 second, 5 minutes, 1 hour, etc.). Each column depicts the timestamp of the day when the energy was consumed. A household will have multiple records depending on the coverage of the dataset under consideration. As earlier discussed, this data is capable of revealing general habits and lifestyles of a household if published in its original form. By assumption, applying k-anonymity to this data will guarantee indistinguishability of k household with stronger privacy. This attack scenario is presented in Fig. 1. In Fig. 1a, an attacker can infer the privacy of each household by simply studying the unprotected data because the consumption pattern of an individual in the data is different. Whereas in Fig. 1b, where 2-anonymity is applied to protect the data, it will be difficult for an attacker to easily distinguish the consumption traces since we can find two households with the same traces in the protected data.

However, the same household can have very similar energy consumption traces per day, making the 2-anonymous



(a) Original data **(b)** 2-anonymous data

Fig. 1 Illustrating privacy value of k-anonymity for protecting daily energy consumption data [12]

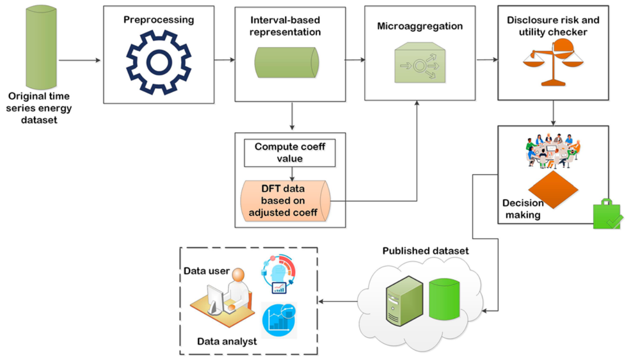

Original time series energy dataset
Preprocessing
Interval-based representation
Microaggregation
Disclosure risk and utility checker
Compute coeff value
DFT data based on adjusted coeff
Decision making
Published dataset
Data user
Data analyst

Fig. 2 Proposed framework for publishing energy data

traces in Fig. 1b point to the same household, thereby leading to successful record linkage. Therefore, it is worth researching the actual privacy value offers by k-anonymity and microaggregation for protecting this type of data. In our study, we empirically show the actual privacy value provided by this protection procedure by considering two types of disclosure risk attacks using distance-based record linkage and interval disclosure. Our findings show that the disclosure risk of k-anonymous energy consumption data with direct application of microaggregation is high and this can be reduced further using the proposed approach in this paper without compromising the utility of the data for research and analytical purposes.

4 Proposed approach

As show in Fig. 2, this paper presents two ways in which energy data can be protected. The time series data are first converted to the form described in Sect. 3.2. This form is termed *interval-based representation* in Fig. 2 for standard representation. The first protection method directly applied microaggregation on the data to produce the masked data. The second approach first applied DFT on the data before microaggregation algorithm. For each case of the protection procedures, we check the utility and privacy values offered by these methods. Based on the outcomes, the utility company decides to publish the protected data for research and analytical purposes. Section 4.2 presents an overview of MDAV algorithm and Sect. 4.3 highlights the detail components of the proposed DFTMicroagg algorithm. In Sect. 6, we show how the proposed DFTMicroagg algorithm reduces disclosure risk while maintaining a high level of data utility.

4.1 Discrete Fourier transform

Discrete Fourier transform (DFT) converts a finite sequence of equally spaced samples of a function into the same length sequence of equally spaced samples coefficients of a finite combination of complex sinusoids, which is a complex-valued function of frequency [38,39]. This property of DFT enables us to efficiently determine the loss and gain of DFT approach by comparing the microaggregated version of DFT anonymized data with the original input data.

An inverse DFT (IDFT) is a Fourier series that uses the DFT samples as coefficients of complex sinusoids at the corresponding DFT frequencies. To provide additional level of masking, instead of producing the original input sequence through IDFT, we modified the coefficients of DFT as described in Sect. 4.3. A fast algorithm for implementing DFT is fast Fourier transform (FFT), which has been widely used in different domains [38]. In this study, we implemented FFT as additional layer to microaggregation algorithm to provide dual-level masking of the energy data.

Formally, a one-dimensional DFT converts a sequence of N complex numbers $\{x_n\} = x_0, x_1, x_2, \ldots, x_{N-1}$ to another sequence of complex numbers $\{\hat{x_k}\} = \hat{x_0}, \hat{x_1}, \hat{x_2}, \ldots, \hat{x}_{N-1}$ such that,

$$\hat{x}_k = \sum_{n=0}^{N-1} x_n.e^{-\frac{i2\pi}{N}kn} \tag{2}$$

The transformation to the complex-valued function of frequency is also denoted as $\hat{x} = F(x)$. The inverse of one-dimensional DFT for a sequence of N complex numbers is given by,

$$x_n = \frac{1}{N}\sum_{k=0}^{N-1} \hat{x}_k.e^{\frac{i2\pi}{N}kn} \tag{3}$$

Suppose n is split into even and odd indexed terms such that $n = 2r$ for even and $n = 2r + 1$ for odd, where $r = 0, 1, \ldots, \frac{N}{2} - 1$. Then Eq. (2) can be computed concurrently in terms of even and odd terms such that,

$$\hat{x}_k = \sum_{r=0}^{\frac{N}{2}-1} x_{(2r)}.e^{-\frac{i2\pi}{N}k(2r)} + \sum_{r=0}^{\frac{N}{2}-1} x_{[2r+1]}.e^{-\frac{i2\pi}{N}k(2r+1)} \tag{4}$$

$$\hat{x}_k = \sum_{r=0}^{\frac{N}{2}-1} x_{(2r)}.e^{-\frac{i2\pi}{N}k(2r)} + e^{-\frac{i2\pi}{N}k} \sum_{r=0}^{\frac{N}{2}-1} x_{[2r+1]}.e^{-\frac{i2\pi}{N}k(2r)} \tag{5}$$

$$\hat{x}_k = \sum_{r=0}^{\frac{N}{2}-1} x_{(2r)}.e^{-\frac{i2\pi}{N/2}k(r)} + e^{-\frac{i2\pi}{N}k} \sum_{r=0}^{\frac{N}{2}-1} x_{[2r+1]}.e^{-\frac{i2\pi}{N/2}k(r)} \tag{6}$$

Similarly, a two-dimensional DFT of discrete sequence $f(x, y)$ of size $M \times N$ is given by,

$$F(u, v) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y).e^{-i2\pi(ux/M+vy/N)} \tag{7}$$

where $F(u, v)$ is the frequency component of the discrete function $f(x, y)$, u and v are the frequency variables in DFT, and x and y are the spatial variables in the input space. The inverse of Eq. (7) is given by,

$$f(x, y) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} F(u, v).e^{i2\pi(ux/M+vy/N)} \tag{8}$$

4.2 MDAV microaggregation

As discussed in Sect. 3, there are several algorithms for microaggregation. However, this study has adapted MDAV [37] as additional layer to DFT due to its performance and wide adoption in the literature [36]. Algorithm 1 describes the stages involved in MDAV.

4.3 DFTMicroagg

4.3.1 Overview of DFTMicroagg

In this study, we propose DFTMicroagg (see Algorithm 2) to improve privacy guarantees of microaggregation algorithm without violating the utility of the protected data. The proposed algorithm aims to improve the privacy value offered by the protection method presented in Fig. 1b. The algorithm takes as input the original energy data X to be masked, an integer number representing the anonymity level and the desired coefficient value which is computed according to Eq. (9). X is a matrix representing daily energy consumption time series data as described in Sect. 3.2. The algorithm produces as output the masked dataset with k-anonymity guaranteed. The parameter *coeff* in the algorithm controls the degree of compression. The proposed algorithm applies a low-pass filtering as an anonymization step before the microaggregation algorithm (see Algorithm 2). This provides a two-level anonymization for the protected energy data and stronger privacy guarantees.

Algorithm 1: MDAV Microaggregation

```
Input: $X$ :- original dataset to protect, $k$: integer, representing
       anonymity level
Output: $\hat{X}$ :- the protected dataset with k-anonymity
begin
    $C = \emptyset$;
    while $|X| \geq 3k$ do
        $\tilde{x} \leftarrow$ average record of all records in $X$;
        $x_r \leftarrow$ most distant record from $\tilde{x}$;
        $x_s \leftarrow$ most distant record from $x_r$;
        $C_r \leftarrow$ cluster around $x_r$ (with $x_r$ and the $k-1$ closest
        records to $x_r$);
        $C_s \leftarrow$ cluster around $x_s$ (with $x_s$ and the $k-1$ closest
        records to $x_s$);
        Remove records in $C_r$ and $C_s$ from $X$;
        $C = C \cup \{C_r, C_s\}$;
    end
    if $|X| \geq 2k$ then
        $\tilde{x} \leftarrow$ average record of all records in $X$;
        $x_r \leftarrow$ most distant record from $\tilde{x}$;
        $C_r \leftarrow$ cluster around $x_r$ (with $x_r$ and the $k-1$ closest
        records to $x_r$);
        $C_s \leftarrow X \setminus C_r$ (form another cluster with the rest of
        the records);
        $C = C \cup \{C_r, C_s\}$;
    else
        $C = C \cup \{X\}$;
    end
    return ($C$);
end
```

First, the variable *no timestamps* in the algorithm represents the total number of columns which corresponds to the timestamps of the day when the energy was consumed. The algorithm tests if the parameter *coeff* is even or odd. Based on the outcome of the test, the indices for the real and imaginary components to be used during FFT are then computed using the function *sequence*. This function takes three parameters. The first parameter is the *start* position of the sequence to be generated, the second is the *stop* position which signifies the end of the interval. The third parameter is the *step* value which indicates the spacing between values in the generated sequence. So, the function *sequence* can be seen as equivalent to numpy.arange() function in Python. The generated real and imaginary indices are used for the FFT computation. Inverse FFT takes as input the computed DFT and the *no timestamps* to produce the transformed data. This is passed as input to MDAV along with the value of k to generate the final masked dataset $\hat{X}$.

4.3.2 Use case of the proposed approach

Suppose we have a time series dataset $D = \{SM_{cid}, timestamp, value\}$ that was collected from AMI smart meters daily by the utility company. In this dataset, SM_{cid} denotes the identifier of households based on the smart meters used. The high-frequency (HF) data (i.e., $value$) from the smart meters denotes the energy consumption of the households at a particular $timestamp$ of the day. As discussed earlier, the HF data can reveal the consumption patterns of the households and this can be explored by attackers even if SM_{cid} is pseudonymized. Utility company wants to protect the privacy of the households in this data so that it will be difficult for an attacker to re-identify a particular household record. At the same time, the protected data should be useful for research and analytical purposes. To protect D via microaggregation, first, the data need to be converted to what we termed $interval - based\ representation$ or standard format X where $t_1, t_2, \ldots, t_n \in T$ represent the number of attributes ($timestamps$) in X along with attributes $Date$ and SM_{cid}. Each row in X denotes the time series daily energy consumption data recorded as SM_{cid},$Date$, and T. Each $t_i \in T$ is a numeric attribute corresponding to the actual energy consumption value at a time t_i and its value needs to be masked to protect the privacy of the households in X. In addition, SM_{cid} is pseudonymized before publishing the data by the utility company to hide the true identities of the households. Each $t_i \in T$ is a $quasi - identifier$ and combination of t_i can be used to re-identify a specific household. It is assumed that a specific t_i or a subset of t_i which is in the possession of an attacker is considered as $confidential$ attribute(s). Therefore, before publishing X, each $t_i \in T$ must be masked to avoid privacy leakage.

To achieve this goal, as stated in Sect. 4, we provide two ways in which X can be protected. The first is to directly apply microaggregation on X to obtain the masked data $\hat{X}$. The second approach is to apply the proposed DFTMicroagg algorithm to protect X. For the sake of clarity, the number of coefficients used for each test case of DFTMicroagg is given by,

$$coeff = \frac{T}{i} \tag{9}$$

where T is the total number of $timestamps$ in X and i is a constant that is to be chosen by the utility company for privacy control. We evaluated with different values of i as presented in Sect. 5.2.5. The motivation is that instead of continuously increasing the value of k to a large number during microaggregation, which can lead to significant information loss, we provide additional layer to microaggregation that offers suitable masking with specific consideration on the shape of the time series. We empirically show that this approach reduces disclosure risk without compromising the data utility of the protected data for research and analytical purposes.

Algorithm 2: DFTMicroagg

```
Input: $X$ :- original dataset to protect, $k$: integer, representing
        anonymity level,
        $coeff$ : integer, representing the number of DFT
coefficients to keep.
Output: $\hat{X}$ :- the protected dataset with k-anonymity
begin
    $no\text{-}timestamps \leftarrow$ number of timestamps in $X$;
    if $is\text{-}even(coeff)$ then
        $Re\text{-}index = sequence(1, coeff, 2)$;
        $Im\text{-}index = sequence(2, coeff, 2)$;
        $dft\text{-}fft \leftarrow$ compute DFT on $X$ using $Re\text{-}index$ and
        $Im\text{-}index$;
    else
        $Re\text{-}index = sequence(1, coeff, 2)$;
        $Im\text{-}index = sequence(2, coeff + 1, 2)$;
        $dft\text{-}fft \leftarrow$ compute DFT on $X$ using $Re\text{-}index$ and
        $Im\text{-}index$;
    end
    $X\text{-}transform \leftarrow$ compute inverse FFT using $dft\text{-}fft$
    and $no\text{-}timestamps$;
    $\hat{X} \leftarrow MDAV(X\text{-}transform, k)$;
    return ($\hat{X}$);
end
```

4.4 Adversarial model

In this paper, we consider an adversary whose goal is to launch two types of record linkage attacks (distance-based record linkage and interval disclosure) to link the records in the masked dataset with an external data that the intruder has obtained through an external knowledge. The external data usually contain the key attributes such as the one in the masked data. When testing a record linkage model, the original dataset is used to represent the intruder external data. For each case of the attack model, we check the privacy values of microaggregation and DFTMicroagg for protecting energy data.

4.4.1 Distance-based record linkage

The goal of an attacker with distance-based record linkage is to use a distance metric to link each record in the masked dataset with its corresponding record in the original. [19] gives a brief description of how a robust distance-based record linkage algorithm for a typical case of microaggregation protection should be developed. For each record in the masked dataset, the distance to every record in the original dataset is computed. Thereafter, the 'closest' and 'second closest' records in the original dataset are considered. A record in the masked dataset is labeled as 'linked' when the closest record in the original dataset is the corresponding original record. Similarly, a record in the masked dataset is labeled as 'linked to 2nd closest' when the second closest record in the original dataset turns out to be the corresponding original record. In all other cases, a record in the masked dataset is labeled as 'not linked.' The percentage of disclosure risk is computed based on the number of 'linked' and 'linked to 2nd closest' records to the overall records in the masked dataset. Based on this description, we propose a robust distance-based record linkage algorithm in Algorithm 3, which does not only consider the closest record but also the second closest to the masked record being linked. This algorithm can also be generalized to evaluate the privacy value of other anonymization methods. Algorithm 3 uses a list comprehension method to compute the distances from each record in the masked dataset to every records in the original dataset. Note also how the closest and second closest distances were computed after the distance computation. The algorithm assumed the maximum knowledge attacker could have regarding the original data.

Algorithm 3: Distance-based Record Linkage

```
Input: $X$ :- original dataset, $\hat{X}$ :- masked dataset
Output: disclosure risk percentage
$linkCount \leftarrow 0$;
begin
    for $i, row\text{-}i\ in\ \hat{X}.rows()$ do
        $d \leftarrow$
        $[Euclidean\text{-}dist(row\text{-}i, row\text{-}j)$ for $j, row\text{-}jinX.rows()]$;

        //compute the closest and second closest;
        $first\text{-}closest\text{-}dist \leftarrow system.maxintsize()$;
        $first\text{-}closest\text{-}index \leftarrow system.maxintsize()$;
        $second\text{-}closest\text{-}dist \leftarrow system.maxintsize()$;
        $second\text{-}closest\text{-}index \leftarrow system.maxintsize()$;
        for index,val in enumerate(d) do
            if $val < first\text{-}closest\text{-}dist$ then
                $second\text{-}closest\text{-}dist \leftarrow$
                $first\text{-}closest\text{-}dist$;
                $second\text{-}closest\text{-}index \leftarrow$
                $first\text{-}closest\text{-}index$;
                $first\text{-}closest\text{-}dist \leftarrow val$;
                $first\text{-}closest\text{-}index \leftarrow index$;
            end
            else if $val < second\text{-}closest\text{-}dist$ and $val \neq$
            $first\text{-}closest\text{-}dist$ then
                $second\text{-}closest\text{-}dist \leftarrow val$;
                $second\text{-}closest\text{-}index \leftarrow index$;
            end
        end
        //increament linkCount based on closest status;
        if $first\text{-}closest\text{-}index =$
        $i$ or $second\text{-}closest\text{-}index = i$ then
            $linkCount \leftarrow linkCount + 1$;
        end
    end
    return $(linkCount/len(\hat{X}) \times 100)$;
end
$Euclidean\text{-}dist(vector1, vector2)$
  return $sqrt(sum((vector1 - vector2)^2))$;
```

4.4.2 Interval disclosure risk

The second adversarial model considered in this study is interval disclosure risk [19], which is an attribute inference attack that tries to infer the smart meter values. Formally, for each record r in the masked dataset $\hat{X}$, an attacker computes rank interval based on the following procedures. First, each attribute in $\hat{X}$ is ranked independently to define a rank interval around the value the attribute takes on each record. Second, the ranks of values within the interval for an attribute around record r should differ less than p percent of the total number of records and the rank in the center of the interval should correspond to the value of the attribute in record r. If true, the proportion of original values that fall into the interval centered around their corresponding masked value is a disclosure risk measure. A 100 percent proportion indicates that an attacker is completely certain that the original value falls in the interval around the masked value. This leads to interval disclosure of the record in the original data. In the case of the daily energy consumption dataset, each attribute is taken as a particular $timestamp$ of the day. A quantitative measure is then computed to quantify the interval disclosure risk for the protected data $\hat{X}$. We implemented interval disclosure via sdcMicro package. Algorithm 4 provides the procedural steps to achieve this goal. In this algorithm, n is the total number of records in $\hat{X}$ and the parameter p can be used to enlarge or down scale the interval.

5 Experimental setup

All experiments have been implemented in Python programming language on a Dell Laptop computer running Windows operating system with 1TB HDD and 32GB RAM. As stated in Sect. 4.4.2, we implemented interval disclosure risk using sdcMicro. sdcMicro is a statistical disclosure control methods for anonymization of data and risk estimation package in R. However, we use *rpy2* package in Python to access sdcMicro.

5.1 Datasets description

We evaluated the efficacy of the proposed approach based on two publicly available datasets. The first dataset 'EnerNOC GreenButton Data,' hereafter refers to as Dataset 1, is a time series energy usage data collected at 5-minute resolution for 100 commercial/industrial sites in the year 2012. The data is available for download at https://open-enernoc-data.s3.amazonaws.com/anon/index.html. The second dataset 'Low Carbon London Electric Vehicle Load Profiles Data,' hereafter refers to as Dataset 2, is a time series data relating to load profiles for electric vehicle charging. This is part of the Low Carbon London (LCL) project delivered by UK Power Net-

Algorithm 4: Interval Disclosure Risk

```
Input: $X$ :- original dataset, $\hat{X}$ :- masked dataset, $n$ :- total
       records in $\hat{X}$, $p$ :- percentage of the deviation
Output: disclosure risk percentage
$linkCount \leftarrow 0$;
begin
    for $rowIndex \ in \ \hat{X}.rows()$ do
        for $attrib \ in \ \hat{X}.columns()$ do
            $rankInterval \leftarrow$ rank attribute $attrib$ by
            computing rank interval;
            for $rankValue \ in \ rankInterval$ do
                if $rankValue < p/100 * n$ then
                    $midRank \leftarrow median(rankInterval)$;
                    if $midRank \leftarrow \hat{X}[rowIndex, attrib]$
                    then
                        if
                        $X[rowIndex, attrib] in rankInterval$
                        then
                            $linkCount \leftarrow linkCount + 1$;
                            $break$;
                        end
                    end
                end
            end
        end
    end
    return $(linkCount/size(\hat{X}) \times 100)$;
end
```

works. The dataset spans two years from 2013 to 2014 with 53 commercial and 70 residential trials. The data is available for download at https://data.london.gov.uk/dataset/low-carbon-london-electric-vehicle- load-profiles. Table 1 summarizes the datasets.

5.2 Utility measures

This section discusses the various ways in which the utility of the proposed approach has been validated. This is to ascertain the usefulness of the protected data for different grid operations such as consumer clustering, consumer profiling, customer segmentation, household daily usage classification, time series forecasting and so on. To capture different tasks and application domains that would be beneficial to data analysts, we conducted clustering analysis using k-Means algorithm, classification based on kNN and time series forecasting using SARIMAX model. In addition, we implemented information loss as described in Sect. 5.2.1 to check the loss of the proposed approach.

5.2.1 Information loss

Measuring information loss (IL) is a crucial step to evaluate a protection procedure in terms of utility–privacy trade-off. IL defines loss of data utility and the goal of the protection procedure is to minimize this loss while reducing the risk of disclosure to an acceptable level. In this study, IL metric that computes the distance between the original dataset X and the masked dataset $\hat{X}$ is implemented as,

$$IL(X, \hat{X}) = \frac{1}{TN} \sum_{j=1}^{T} \sum_{i=1}^{N} \frac{|x_{ij} - \hat{x}_{ij}|}{\sqrt{2\sigma_j}} \tag{10}$$

where T is the total number of $timestamps$; N is the number of daily energy profiles in the dataset; x_{ij} and $\hat{x}_{ij}$ are the values before and after perturbation for $timestamp \ j$ and profile i, respectively; σ_j is the standard deviation of $timestamp \ j$ in X [43]. The higher the value of IL, the higher the information loss.

5.2.2 Clustering analysis

To test the utility of the protected data in terms of clustering of daily energy profile, we implemented k-Means algorithm with k-Means++ heuristic algorithm for initializing the clusters' centroids [44]. k-Means is a popular clustering algorithm that partitions data into k clusters around the nearest centroids (mean of the cluster centers). To avoid confusing the k hyperparameter in k-Means algorithm with microaggregation, we represent k in k-Means as c, where c is the number of clusters to generate from the data. To measure clustering quality, we use Silhouette score as a cluster validity measure that checks how similar a daily energy profile is to its cluster (cohesion) compared to the daily energy profile in another clusters (separation). Silhouette is defined as a method of interpretation and validation of consistency within clusters of data. The silhouette value measures how similar an object is to its own cluster compared to other clusters. Silhouette score ranges from -1 to $+1$, where a high score indicates that the daily energy profile is well grouped to its cluster. -1 indicates poor clustering and 0 indicates overlapping clusters.

5.2.3 kNN classification

To further test the utility of the protected data, we conducted a classification task where each daily energy profile is categorized based on the household or consumer type. For Dataset 1, the classes are 1, 2, 3 and 4 representing the commercial property, education, food sales and storage, and light industrial buildings, respectively, as described in the dataset. There are two classes in Dataset 2 where 1 is used for residential consumers and 2 for commercial consumers. We train kNN algorithm to classify the profile in both the original and masked datasets. kNN is a supervised machine learning algorithm, which can be used for both classification and regression tasks. During the classification stage, an unlabeled

Table 1 Datasets description

Name	Date	Size	Attribute	Resolution	Data owner
EnerNOC Green-Button Data	2012	100 commercial/industrial sites. 10,531,288 readings (after merging) (Mean: 49.83, Std: 79.13).	Timestamp, dttm_uct, value (kW), estimated, anomaly, lat, long.	5-minute	EnerNOC
Low Carbon London Electric Vehicle Load Profiles Data	2013-2014	123 customers, 53 commercial and 70 residential trials. 4,472,919 readings (after merging) (Mean: 0.57, Std: 2.45).	Energy (kWh) for different measurement.	10-minute	UK Power Networks

test sample is classified by assigning the class which is most frequent among the k training samples nearest to that test sample. For a similar reason, as stated in Sect. 5.2.2, we use c to represent the number of nearest neighbors for kNN. To evaluate the classification performance of kNN, accuracy and F1-score were used. Accuracy measures the percentage of the correct predictions for the test samples while F1-score is calculated from recall and precision.

5.2.4 Forecasting model

The most common task on time series data is forecasting and to test the usefulness of the protected data based on the proposed approach, we developed SARIMAX model to perform mean hourly energy consumption forecast. SARIMAX is an extension of Auto-regressive Integrated Moving Average (ARIMA) model that comprises two parts: auto-regressive part (AR) and the moving average (AR) part. The Integrated (I) component of ARIMA is for differencing purposes. ARIMA model has been widely used for time series forecasting as it provides promising models on time series data. However, the main issue with ARIMA is that it cannot handle seasonality. Seasonal ARIMA (SARIMA) is provided to handle this drawback. SARIMA component is given by $SARIMA(p, d, q)(P, D, Q)_m$, where p is the non-seasonal AR order; d is the non-seasonal differencing; q is the non-seasonal MA order; P is the seasonal AR order; D is the seasonal differencing; Q is the seasonal MA order, and m is the length of repeating seasonal pattern. Using the seasonal components, SARIMA solves the problem of seasonality. SARIMAX extends this model by providing the capability to handle exogenous attributes. For further reading on ARIMA model, the reader is referred to [45]. To efficiently determine the values of the SARIMAX model parameters, we perform a grid search method to obtain the optimal values for modeling SARIMAX.

The performance of the forecast model is evaluated using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) metrics. MSE is defined in Eq. (11). RMSE is the square root of MSE.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \tag{11}$$

where N is the total number of data points; Y_i is the observed value for data i in the time series data; and $\hat{Y}_i$ is the equivalent forecasted value.

5.2.5 Hyperparameter settings

The different parameter settings for each of the methods discussed in the previous sections are summarized in Table 2.

6 Results and discussion

In this section, we discuss the results obtained from the different experiments conducted in this study to test the efficacy of the proposed method. This section is divided into two. The first section shows the results obtained when applying microaggregation as the protection mechanism. The second section presents the results of microaggregation alongside DFTMicroagg results. The two sections focus on utility and privacy results computed for the two datasets that we have considered in this study.

6.1 Microaggregation results

This section presents the results of applying microaggregation (see Algorithm 1) as a privacy protection mechanism on Dataset 1 and Dataset 2. Each daily profile in Dataset 1 was sampled at an equal time interval of 5 minutes, so, there

Table 2 Hyperparameters settings

Algorithm/Model	Hyperparameter	Value
MDAV	k	$k = 2, 3, 4, 5.$
DFTMicroagg	k, *coeff*	$k = 2, 3, 4, 5$ and *coeff* is based on Eq. (9), where $i = 2, 3, 4, 5, 6.$
k-Means	c	$c = 2, 3, 4.$
kNN	c	$c = 2, 3.$
SARIMAX	$(p, d, q)(P, D, Q)_m$	Optimal value using grid search. (1, 0, 1)(1, 1, 1, 12).

are a total of 288 timestamps per day (12 * 24h = 288 samples). It is important to mention that there is a total of 36,401 records in Dataset 1 and 29,597 records in Dataset 2 after merging and preprocessing based on interval representation format. The unit of measurement of the two datasets varies and The level of sparsity of Dataset 2 is higher than Dataset 1 as shown in 4. Similarly, Dataset 2 has a total timestamps of 144 per day, which was sampled at 10-minute resolution (i.e., 6 * 24h = 144 samples).

Figure 3 shows the result of applying micrpaggregation on Dataset 1. It can be seen that microaggregation algorithm computes the mean of the similar daily profiles that were clustered alongside the first time series investigated in these plots. The series in the figure represent a full day consumption. It can be seen that at around 12:50pm to 7:00pm (i.e., timestamps 150-228), the consumers experience a significant increase in energy usage for that day, which is similar to the usage pattern of the first time series investigated. This is usually due to the use of energy-hungry appliances that consume a significant amount of energy for that period. Figures show that different consumers have very similar daily energy usage patterns. Similarly, Fig. 4 shows the application of microaggregation algorithm to protect the records in Dataset 2. The figures demonstrate similar usage patterns in energy consumption of the consumers in the dataset when charging their electric vehicles. The series in the figures represent a full day consumption. It can be seen that at some periods of the day, the energy usage of some consumers increases for a longer period. This shows the consumption habit of the consumers in the dataset when charging their electric vehicles. Microaggregation aims to protect this consumption habit by generating k-anonymous records that are indistinguishable.

As discussed earlier, k-anonymity is one way to provide privacy protection at individual level, as microaggregation is an algorithm that produces data that is compliant with the k-anonymity privacy model. Using microaggregation on the daily energy profile data, we can provide k-anonymous profiles. For instance, when k = 2, there will be at least 2 daily energy profiles with exactly the same values (i.e., the output of the algorithm will be two exact copies of the same time series generated from two different, but similar, individual time series).

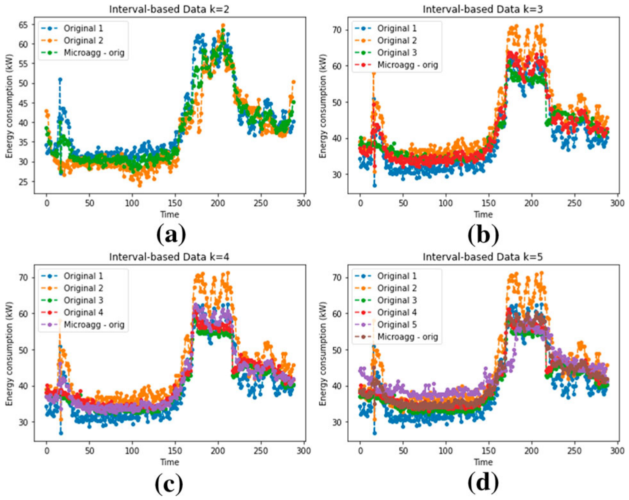


Fig. 3 Microaggregation on Interval-based daily energy consumption data for the first time series in Dataset 1. **a** k = 2; **b** k = 3; **c** k = 4 and **d** k = 5

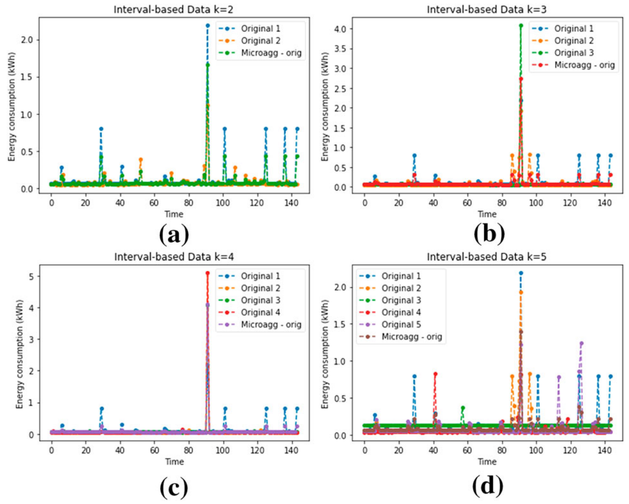


Fig. 4 Microaggregation on Interval-based daily energy consumption data for the first time series in Dataset 2. **a** k = 2; **b** k = 3; **c** k = 4 and **d** k = 5

6.1.1 Utility

Information Loss: Table 3 and Table 4 show the information loss (IL) of applying microaggregation directly on Dataset 1 and Dataset 2, respectively. As expected, the higher the

Table 3 Information loss for each level of k for interval-based representation on Dataset 1

k-Anonymity Level	Information Loss (IL)
k = 2	0.1171
k = 3	0.1421
k = 4	0.1784
k = 5	0.1992

Table 4 Information loss for each level of k for interval-based representation on Dataset 2

k-Anonymity level	Information loss (IL)
k = 2	0.0440
k = 3	0.0634
k = 4	0.0759
k = 5	0.0824

value of k, the higher the information loss. Therefore, we check the utility of the microaggregated masked data when use for different data analyses.

Clustering analysis: Tables 5 and 6 show the result of applying k-Means clustering on the original and microaggregated interval-based data using Dataset 1 and Dataset 2, respectively. The results obtained show that the microaggregated data is very useful for clustering tasks. Most importantly, the clustering process of microaggregation affects the k-Means clustering step as it can be seen clearly when the number of clusters from k-Means is 4 in Dataset 1. This produces more divergence on the clustering analysis. For each case of clustering on the microaggregated data based on Dataset 1, Silhouette score was above 0.7 and higher than the result obtained with the original data, which shows the quality of the clusters formed.

Similarly for Dataset 2, for each case of the clustering on the microaggregated data, Silhouette score was above 0.6, which shows the quality of the clusters formed.

Classification: Recall that in Sect. 5.2.3, we stated that there are four classes in Dataset 1 representing the commercial property, education, food sales and storage, and light industrial buildings, respectively, as described in the dataset. In this section, we check the performance of the microaggregated data for classification of these consumers type based on their daily energy consumption. As shown in Table 7, microaggregated data achieved close results in terms of accuracy and F1-score when compare with the original data. The accuracy and F1-score of the microaggregation dropped to 79.18% when two nearest neighbors were used with $k = 5$. For three nearest neighbors, the accuracy and F1-score maintained 80.41% with $k = 5$. This result confirmed the utility of the microaggregated data for classification of consumers' daily consumption profiles on Dataset 1.

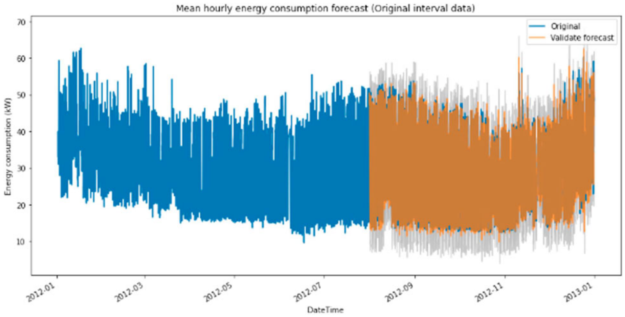


Fig. 5 Utility based on mean hourly time series forecasting on original data for consumer 6

For Dataset 2, as shown in Table 8, microaggregation achieved close results in terms of accuracy and F1-score when compared with the original data. The accuracy and F1-score of the microaggregation were above 80% for each value of k. This results confirmed the usefulness of microaggregated data for classification of consumers energy consumption in Dataset 2 as either residential or commercial profile.

Time series forecasting:We conducted mean hourly time series forecasting on the original and microaggregated data using the two datasets. The procedure to achieve this using Dataset 1 as an example is as follows. First, in order to align with the specific time series data requirement format for SARIMAX model, we converted the interval-based data to the form discussed earlier in Sect. 4.3. This conversion generated over 10 million samples (see Sect. 5.1). Second, we generated mean hourly load data, which was used to develop the forecast model. The goal of SARIMAX model is to predict the value of hourly energy consumption for a particular consumer and timestamp. This aligned with the demand response service that can be rendered by the utility company. We use data from September 1, 2012, to December 31, 2012, as the test data to validate the SARIMAX model based on Dataset 1. Recall that Dataset 1 covers a period between January 2012 and December 2012. Due to the space constraint, Figs. 5 and 6 showed the visualizations of mean hourly forecasting on original and microaggregated data for consumer with identity 6 in Dataset 1, respectively.

Table 9 shows the MSE and RMSE results of the forecast model for both the original and microaggregated data for Dataset 1. For Dataset 1, we noticed that the MSE and RMSE of forecasting for the microaggregated data reduced than the original data. It can also be seen when the value of k increases.

Similarly, Table 10 presents the MSE and RMSE of mean hourly load forecasting on Dataset 2. As discussed earlier, the level of sparsity of Dataset 2 is higher that Dataset 1. This may account for the reduction in MSE when compare with the results obtained with Dataset 1. However, microaggregation maintained a consistent results across the two datasets when

Table 5 Utility based on clustering analysis on original and microaggregated interval-based representation on Dataset 1

	Silhouette score (c = 2)	Silhouette score (c = 3)	Silhouette score (c = 4)
Original	0.7963	0.7567	0.6120
Microaggregation (k = 2)	0.7970	0.7584	0.7171
Microaggregation (k = 3)	0.7970	0.7591	0.7186
Microaggregation (k = 4)	0.7979	0.7595	0.7192
Microaggregation (k = 5)	0.7984	0.7599	0.7197

Table 6 Utility based on clustering analysis on original and microaggregated interval-based representation on Dataset 2

	Silhouette score (c = 2)	Silhouette score (c = 3)	Silhouette score (c = 4)
Original	0.6685	0.6760	0.6772
Microaggregation (k = 2)	0.6683	0.6729	0.6818
Microaggregation (k = 3)	0.6712	0.6745	0.6878
Microaggregation (k = 4)	0.6503	0.6719	0.6853
Microaggregation (k = 5)	0.6723	0.6725	0.6850

Table 7 Utility based on kNN classification on original and microaggregated interval-based representation for Dataset 1

	kNN (c = 2)		kNN (c = 3)	
	Acc	F1-Score	Acc	F1-Score
Original	0.8290	0.8290	0.8247	0.8247
Microaggregation (k = 2)	0.8151	0.8151	0.8144	0.8144
Microaggregation (k = 3)	0.8111	0.8111	0.8152	0.8152
Microaggregation (k = 4)	0.8000	0.8000	0.8079	0.8079
Microaggregation (k = 5)	0.7918	0.7918	0.8041	0.8041

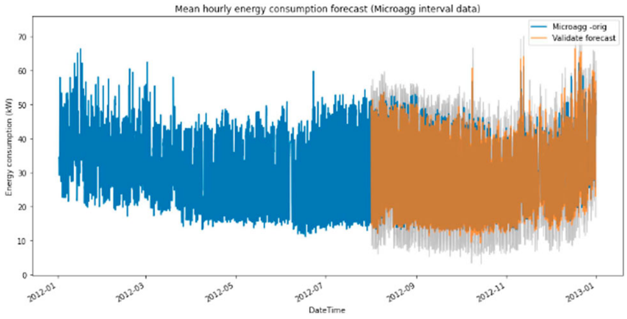


Fig. 6 Utility based on mean hourly time series forecasting on microaggregated data for consumer 6

Table 9 Utility based on mean hourly load forecasting on original and microaggregated data for Dataset 1

	MSE	RMSE
Original	56.9108	7.5439
Microaggregation (k = 2)	50.3921	7.0987
Microaggregation (k = 3)	48.3414	6.9528
Microaggregation (k = 4)	44.3132	6.6568
Microaggregation (k = 5)	41.5235	6.4439

compared with the original for each level of k. This shows the applicability of the masked data for forecasting energy load.

Table 8 Utility based on kNN classification on original and microaggregated interval-based representation for Dataset 2

	kNN (c = 2)		kNN (c = 3)	
	Acc	F1-Score	Acc	F1-Score
Original	0.8796	0.8796	0.8708	0.8708
Microaggregation (k = 2)	0.8599	0.8599	0.8607	0.8607
Microaggregation (k = 3)	0.8453	0.8453	0.8474	0.8474
Microaggregation (k = 4)	0.8305	0.8305	0.8282	0.8282
Microaggregation (k = 5)	0.8227	0.8227	0.8220	0.8220

Table 10 Utility based on mean hourly load forecasting on original and microaggregated data for Dataset 2

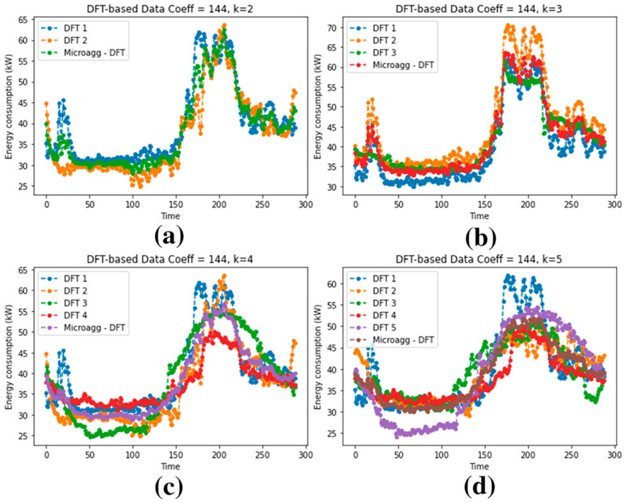
	MSE	RMSE
Original	2.5402	1.5938
Microaggregation (k = 2)	2.3918	1.5465
Microaggregation (k = 3)	2.2890	1.5129
Microaggregation (k = 4)	2.1946	1.4814
Microaggregation (k = 5)	2.1264	1.4582

6.1.2 Disclosure risk

This section discusses the privacy value of microaggregation when used to protect Dataset 1 and Dataset 2. Table 11 shows the results obtained using the proposed distance-based record linkage algorithm (see Algorithm 3) and interval disclosure risk based on Dataset 1. To provide a detailed analysis on record linkage, we evaluated two scenarios. The first being the case when the masked record was linked to the closest record in the original dataset and the second scenario was when the masked record was linked to both the closest and second closest records (see Algorithm 3 description). The results based on the second scenario were presented inside the parentheses. The one without parentheses represents the results of the first scenario. For interval disclosure, we passed both the original and masked datasets to the disclosure risk measure of sdcMicro. The result of this experiment was also shown in Table 11. Recall that we have a total of 36,401 and 29,597 records in Dataset 1 and Dataset 2, respectively, after merging and preprocessing based on interval representation format.

From the results obtained, it can be seen that the disclosure risk of microaggregation for protecting Dataset 1 is on the high side. For instance, when $k = 2, 48.67\%$ was linked to the closest records while 87.26% was linked to both the closest and second closest records in the original with distance-based record linkage. For interval disclosure, an attacker was 71.86% sure that the original value lies in the interval constructed around the masked value. The lowest disclosure risk when $k = 5$ produced 17.35% and 34.04% for the two scenarios of distance-based record linkage, respectively, and 47.87% for interval disclosure. The goal of DFTMicroagg is to further reduce this disclosure risk to a certain extent while still maintaining high level of data utility with minimal loss without the need to increase the value of k. In Sect. 6.2, we empirically show that this is achievable with the application of DFTMicroagg.

Similarly, Table 12 shows the disclosure risk of applying microaggregation as a protection procedure for Dataset 2. As mentioned earlier, the goal of DFTMicroagg is to lower the disclosure risk while ensuring the usability of the masked data for research and analytic purposes.



(a) (b) (c) (d)

Fig. 7 k-anonymity satisfaction via DFTMicroagg algorithm (*coeff* = 144) for the first time series in Dataset 1. **a** k = 2; **b** k = 3; **c** k = 4 and **d** k = 5

6.2 Microaggregation and DFTMicroagg

In this section, we discuss the results obtained when DFTMicroagg was applied as a protection method. Due to the space constraint, we present the results of the upper and lower value of the coefficients that were used by DFTMicroagg algorithm. The upper value corresponds to 144 while the lower value is 48 for Dataset 1 (see Eq. (9)). These are equivalent to 72 and 24, respectively, for Dataset 2. These two values will vary according to the dataset as earlier discussed. In addition, for simplicity and clear comparison, we present microaggregation results alongside DFTMicroagg in this section. Despite the fact that the two datasets were collected using different units of measurement, microaggregation and DFTMicroagg produced consistent results across the two datasets as will be seen in the results obtained.

Figures 7 and 8 show the outcome of applying DFTMicroagg with 144 and 48 coefficients values, respectively, on Dataset 1. It can be seen from the figures that DFTMicroagg algorithm maintained the consumption patterns similar to what was obtained when microaggregation was directly applied (see Fig. 3). The patterns of households with similar energy consumption have been preserved and through the application of microaggregation as additional layer, k-anonymity was enforced on the data for privacy protection.

Figures 9 and 10 show the outcome of applying DFTMicroagg with 72 and 24 coefficients values, respectively, on Dataset 2. Similarly, DFTMicroagg maintained the consumption patterns similar to what was obtained when microaggregation was directly applied (see Fig. 4). The patterns of households with similar energy consumption have been protected based on k-anonymity. The consumption values of all the similar energy profiles including that of the first time

Table 11 Disclosure risk for each level of k based on microaggregated data for Dataset 1

	Disclosure Risk (%) Distance-based record linkage	Interval disclosure
Microaggregation (k = 2)	48.67 (87.26)	71.86
Microaggregation (k = 3)	31.61 (62.59)	61.54
Microaggregation (k = 4)	22.91 (44.87)	51.61
Microaggregation (k = 5)	17.35 (34.04)	47.87

Table 12 Disclosure risk for each level of k based on microaggregated data for Dataset 2

	Disclosure Risk (%) Distance-based record linkage	Interval disclosure
Microaggregation (k = 2)	47.30 (80.85)	41.02
Microaggregation (k = 3)	28.83 (54.97)	34.93
Microaggregation (k = 4)	19.14 (35.66)	31.31
Microaggregation (k = 5)	14.05 (26.11)	28.98

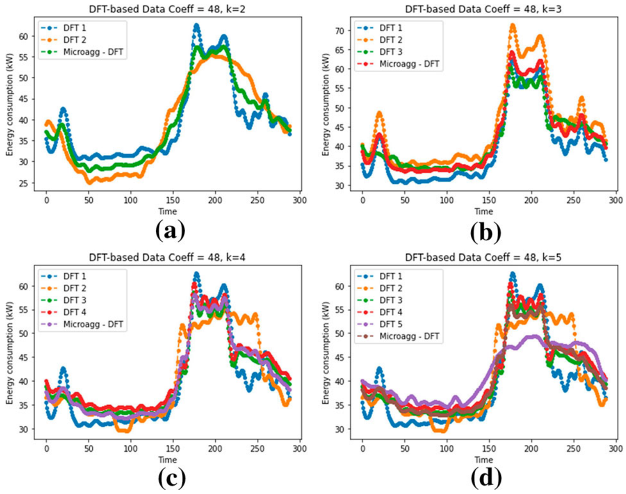


Fig. 8 k-anonymity satisfaction via DFTMicroagg algorithm (*coeff* = 48) for the first time series in Dataset 1. **a** k = 2; **b** k = 3; **c** k = 4 and **d** k = 5

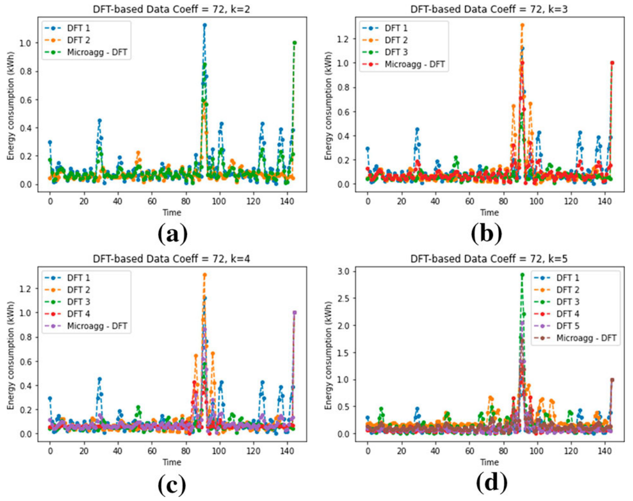


Fig. 9 k-anonymity satisfaction via DFTMicroagg algorithm (*coeff* = 72) for the first time series in Dataset 2. **a** k = 2; **b** k = 3; **c** k = 4 and **d** k = 5

series investigated in the plots have been replaced with the centroid that was computed using microaggregation layer. The subsequent sections present the utility and disclosure risk of applying DFTMicroagg to protect the two datasets.

6.2.1 Utility

Information Loss: Table 13 shows the information loss of applying DFTMicroagg with the IL of microaggregation based on Dataset 1. It was observed that the higher the value of coefficient, the lower the information loss. The subsequent sections show the benefits of incurring this loss as a good trade-off for privacy preserving of individual household consumption. It can also be seen in the subsequent sections that despite this loss, DFTMicroagg maintained a high level of data utility for research and analytic purposes. Therefore, utility company has the flexible option of choosing the actual coefficient value that suits their data publication policy. Similarly, for Dataset 2, Table 14 shows the IL of both microaggregation and DFTMicroagg.

Clustering analysis: Tables 15 and 16 further confirmed the applicability of the proposed DFTMicroagg for clustering analysis on Dataset 1 and Dataset 2, respectively. For Dataset 1, the clustering result of DFTMicroagg, even when $k = 5$, was above the result of the direct application of microaggregation algorithm and the original. Nevertheless, for each value of k, DFTMicroagg maintained Silhouette score that was above 0.7. For Dataset 2, similar to the result obtained when microaggregation was applied, DFTMicroagg maintained a high level of utility for clustering analysis. In all cases, the algorithm produced results that slightly improved

Table 13 Information loss (IL) of microaggregation and DFTMicroagg on Dataset 1

k-Anonymity Level	IL(Microagg)	IL(DFTMicroagg Coeff = 144)	IL(DFTMicroagg Coeff = 48)
k = 2	0.1171	0.1388	0.1628
k = 3	0.1421	0.1592	0.1807
k = 4	0.1784	0.1882	0.2016
k = 5	0.1992	0.2056	0.2152

Table 14 Information loss of microaggregation and DFTMicroagg on Dataset 2

k-Anonymity Level	IL(Microagg)	IL(DFTMicroagg Coeff = 72)	IL(DFTMicroagg Coeff = 24)
k = 2	0.0440	0.0732	0.1169
k = 3	0.0634	0.0849	0.1230
k = 4	0.0759	0.0932	0.1274
k = 5	0.0824	0.0995	0.1307

Table 15 Utility based on clustering analysis on original, microaggregation and DFTMicroagg using Dataset 1

	Silhouette score (c = 2)	Silhouette score (c = 2, Coeff = 144)	Silhouette score (c = 2, Coeff = 48)
Original	0.7963	–	–
Microaggregation (k = 2)	0.7970	–	–
DFTMicroagg (k = 2)	–	0.7973	0.7976
DFTMicroagg (k = 3)	–	0.7973	0.7979
DFTMicroagg (k = 4)	–	0.7981	0.7982
DFTMicroagg (k = 5)	–	0.7985	0.7987

over the original and microaggregation datasets based on Dataset 2.

Classification: Tables 17 and 18 further confirmed the applicability of the proposed DFTMicroagg for classification task on Dataset 1 and Dataset 2, respectively. For Dataset 1, the accuracy of DFTMicroagg when $k = 2$ was close to the original and slightly higher than the accuracy result of microaggregation algorithm. When the coefficient was 48 and $k = 2$, we noticed a slight increase in the accuracy value when compared with the result of the original data (see Table 17). Similarly for Dataset 2, the accuracy of DFTMicroagg during classification when $k = 2$ was also close to the microaggregation result. In all cases, DFTMicroagg produced an accuracy that was above 80% on Dataset 2.

Time series forecasting: The patterns of results obtained in the previous section can also be seen in Table 19 where we notice a reduction in MSE and RMSE of DFTMicroagg for mean hourly load forecasting on Dataset 1. Similarly,

Table 16 Utility based on clustering analysis on original, microaggregation and DFTMicroagg using Dataset 2

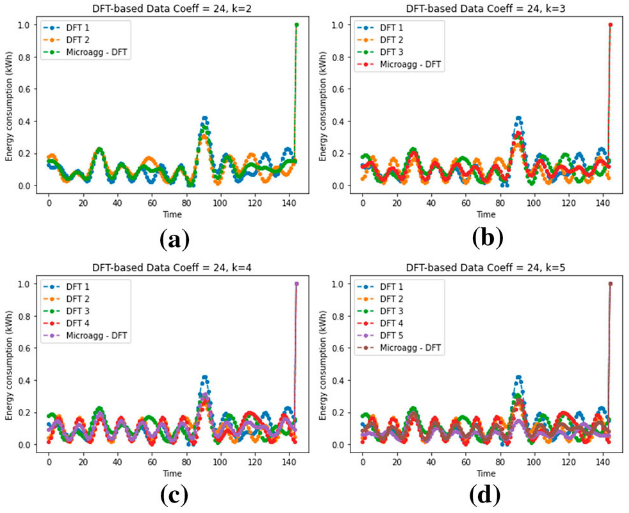
	Silhouette score (c = 2)	Silhouette score (c = 2, Coeff = 72)	Silhouette score (c = 2, Coeff = 24)
Original	0.6685	–	–
Microaggregation (k = 2)	0.6683	–	–
DFTMicroagg (k = 2)	–	0.6729	0.6840
DFTMicroagg (k = 3)	–	0.6719	0.6827
DFTMicroagg (k = 4)	–	0.6564	0.6836
DFTMicroagg (k = 5)	–	0.6763	0.6852

Table 17 Utility based on kNN classification of daily energy consumption profiles on original, microaggregation and DFTMicroagg using Dataset 1

	kNN (c = 2)	kNN (c = 2, Coeff = 144)	kNN (c = 2, Coeff = 48)
Original	0.8290	–	–
Microaggregation (k = 2)	0.8151	–	–
DFTMicroagg (k = 2)	–	0.8221	0.8310
DFTMicroagg (k = 3)	–	0.8119	0.8244
DFTMicroagg (k = 4)	–	0.8089	0.8149
DFTMicroagg (k = 5)	–	0.8039	0.8046

Table 18 Utility based on kNN classification of daily energy consumption profiles on original, microaggregation and DFTMicroagg using Dataset 2

	kNN (c = 2)	kNN (c = 2, Coeff = 72)	kNN (c = 2, Coeff = 24)
Original	0.8796	–	–
Microaggregation (k = 2)	0.8599	–	–
DFTMicroagg (k = 2)	–	0.8569	0.8393
DFTMicroagg (k = 3)	–	0.8362	0.8242
DFTMicroagg (k = 4)	–	0.8312	0.8197
DFTMicroagg (k = 5)	–	0.8276	0.8134



(a) (b) (c) (d)

Fig. 10 k-anonymity satisfaction via DFTMicroagg algorithm (*coeff* = 24) for the first time series in Dataset 2. **a** k = 2; **b** k = 3; **c** k = 4 and **d** k = 5

as shown in Table 20, we notice a reduction in MSE and RMSE as compared with the original and microaggregated data based on Dataset 2. This also confirmed the consistency of the proposed approach as a protection method with generalization feature.

6.2.2 Disclosure risk

This section discusses the privacy value of DFTMicroagg and compares it with microaggregation result on both datasets. For Dataset 1, Table 21 presents the disclosure risk of DFTMicroagg alongside the result of microaggregation in terms of distance-based record linkage. As shown in the table, when $k = 2$ and $Coeff = 144$, DFTMicroagg prevented approximately additional 6,541 records in the masked dataset against record linkage attack by considering the closest records when compare with microaggregation result while it was 7,265 records for the second scenario of distance-based record linkage. Similarly, the result based on coefficient value of 48 prevented 8,295 records for the first case and 8,761 records for the second scenario. For both first and second scenarios of record linkage attack, and for each level of k-anonymity, DFTMicroagg outperformed the direct application of microaggregation algorithm for privacy protection of energy consumption data using Dataset 1. This gives the utility company a flexible option to control the privacy level of energy data to be published by adjusting the coefficient value of DFTMicroagg while still maintaining a high level of data utility without increasing the value of k to avoid significant information loss.

Similarly, for Dataset 2, Table 22 shows the disclosure risk of DFTMicroagg with the result of microaggregation in terms of distance-based record linkage. According to the results in this table, when $k = 2$ and $Coeff = 72$, DFTMicroagg prevented approximately additional 7,950 records in the masked dataset against record linkage attack by considering the closest records when compare with microaggregation result while it was 13,869 records for the second scenario of record linkage. Recall that Dataset 2 has a total of 29,597 records after merging and preprocessing based on interval representation format. Similarly, the result based on coefficient value of 24 and $k = 2$ prevented 11,131 records for the first scenario and 19,220 records based on the second scenario of record linkage attack when compared with microaggregation results. For both first and second scenarios of record linkage attack and for each level of k-anonymity, DFTMicroagg outperformed

Table 19 Utility based on time series forecasting on original, microaggregation and DFTMicroagg using Dataset 1

	MSE	RMSE
Original	56.9108	7.5439
Microaggregation (k = 2)	50.3921	7.0987
Microaggregation (k = 3)	48.3414	6.9528
Microaggregation (k = 4)	44.3132	6.6568
Microaggregation (k = 5)	41.5235	6.4439
DFTMicroagg (k = 2, Coeff = 144)	50.3821	7.0980
DFTMicroagg (k = 3, Coeff = 144)	48.2224	6.9442
DFTMicroagg (k = 4, Coeff = 144)	44.5270	6.6729
DFTMicroagg (k = 5, Coeff = 144)	41.2078	6.4193
DFTMicroagg (k = 2, Coeff = 48)	50.0428	7.0741
DFTMicroagg (k = 3, Coeff = 48)	47.0064	6.8561
DFTMicroagg (k = 4, Coeff = 48)	43.2996	6.5802
DFTMicroagg (k = 5, Coeff = 48)	40.2921	6.3476

Table 20 Utility based on time series forecasting on original, microaggregation and DFTMicroagg using Dataset 2

	MSE	RMSE
Original	2.5402	1.5938
Microaggregation (k = 2)	2.3918	1.5465
Microaggregation (k = 3)	2.2890	1.5129
Microaggregation (k = 4)	2.1946	1.4814
Microaggregation (k = 5)	2.1264	1.4582
DFTMicroagg (k = 2, Coeff = 72)	2.3709	1.5398
DFTMicroagg (k = 3, Coeff = 72)	2.2655	1.5052
DFTMicroagg (k = 4, Coeff = 72)	2.1773	1.4756
DFTMicroagg (k = 5, Coeff = 72)	2.1226	1.4569
DFTMicroagg (k = 2, Coeff = 24)	2.0949	1.4474
DFTMicroagg (k = 3, Coeff = 24)	2.0034	1.4154
DFTMicroagg (k = 4, Coeff = 24)	1.9359	1.3914
DFTMicroagg (k = 5, Coeff = 24)	1.8893	1.3745

the direct application of microaggregation algorithm for privacy protection of energy consumption data. The algorithm also maintained a high level of data utility when compared with the original and direct application of microaggregation.

For Dataset 1, Table 23 also confirmed the applicability of the proposed approach by lowing the chances of an attacker to accurately construct the interval value around the masked value in the dataset. DFTMicroagg reduced the disclosure risk while keeping the k-anonymity level in the range of 2 to 5. For instance, when $k = 2$, the proposed approach reduced disclosure risk from 71.86% to 58.43% when the coefficient value was 144. This can go as low as 48.31% when the coefficient value is 48 and $k = 2$. The results presented for both cases of the disclosure risk show that 2-anonymous daily energy profile is susceptible to disclosure risk as against the attack model assumption in [12]. For each level of k, DFTMicroagg reduced the disclosure risk.

Similarly, for Dataset 2, Table 24 also confirmed the applicability of the proposed approach based on the results of the interval disclosure risk. DFTMicroagg lower the percentage of correctly predicting the interval value around the masked value in the protected dataset.

6.3 Order of households and sampling rate

In this section, we provide the results obtained based on the order of households of rows in matrix X as well as using a different sampling rate.

6.3.1 Order of households

In the previous results, the rows of matrix X were arranged in chronological order based on the consumption day for individual households. Therefore, for each day (e.g., 01/02/2012), the first row contains energy consumption for household 1, the second row contains energy consumption for household 2 and so on. This pattern was used for another day's consumption (e.g., 02/02/2012). The columns of matrix X are the actual time of the day in which the consumptions were recorded. In this section, we check the impact of sorting matrix X in ascending and descending order based on household's number.

Utility: In this section, we check the utility of the proposed approach based on order of households using information loss and clustering analysis. The goal is to ascertain the impact of ordering households before applying microaggregation and DFTMicroagg algorithms.

Ascending and Descending order of households: We obtained the same result as those presented in Sect. 6.2.1 for information loss (IL) measure for both microaggregation and DFTMicroagg algorithms when the datasets were sorted in ascending and descending order (see Tables 13 and 14). This shows that the proposed approach does not depend on the order of households in terms of the IL metric used. However, there is a slight change in the clustering results for both Dataset 1 and Dataset 2, respectively. This can be attributed to the random selection of initial cluster centroids in k-Means algorithm since the ordering of the records in both datasets has changed. Nevertheless, microaggregation and DFTMicroagg produced consistent results and guaranteed utility of the protected data. Silhouette score based on ascending and descending order of households is not less than 70% and 60% for both Dataset 1 and Dataset 2, respectively, which is similar to the result obtained for the original data without the application of privacy protection mechanisms. For instance, considering Dataset 1 in ascending order, the Silhouette score for the original dataset when $k = 2$ is 0.7408 and for microaggregation is 0.7394. However, DFTMicroagg

Table 21 Distance-based record linkage disclosure risk for each level of k based on microaggregation and DFTMicroagg for Dataset 1

Methods	Disclosure risk (%) k = 2	k = 3	k = 4	k = 5
Microaggregation	48.67 (87.26)	31.61 (62.59)	22.91 (44.87)	17.35 (34.04)
DFTMicroagg (Coeff = 144)	30.70 (67.30)	20.55 (49.30)	11.79 (32.19)	9.50 (24.92)
DFTMicroagg (Coeff = 48)	25.88 (63.19)	18.44 (43.40)	10.15 (28.11)	6.39 (20.46)

Table 22 Distance-based record linkage disclosure risk for each level of k based on microaggregation and DFTMicroagg for Dataset 2

Methods	Disclosure risk (%) k = 2	k = 3	k = 4	k = 5
Microaggregation	47.30 (80.85)	28.83 (54.97)	19.14 (35.66)	14.05 (26.11)
DFTMicroagg (Coeff = 72)	20.44 (33.99)	12.67 (22.23)	8.92 (15.81)	6.86 (12.44)
DFTMicroagg (Coeff = 24)	9.69 (15.91)	6.60 (11.44)	4.98 (8.83)	4.11 (7.39)

Table 23 Interval disclosure risk of microaggregation and DFTMicroagg on Dataset 1

k-Anonymity Level	Interval disclosure risk (%) Microaggregation	DFTMicroagg Coeff = 144	DFTMicroagg Coeff = 48
k = 2	71.86	58.43	48.31
k = 3	61.54	51.11	45.00
k = 4	51.61	47.66	43.19
k = 5	47.87	45.24	41.52

produced 0.7399 and 0.7412 for coefficient of 144 and 48, respectively. For descending order, Silhouette score for the original dataset is 0.7142 and for microaggregation is 0.7174 while DFTMicroagg produced 0.7177 and 0.7183 for coefficient of 144 and 48, respectively.

When considering Dataset 2 in ascending order, the Silhouette score for the original dataset when $k = 2$ is 0.6550 and for microaggregation is 0.6593. However, DFTMicroagg produced 0.6615 and 0.6712 for coefficient of 72 and 24, respectively. For descending order, Silhouette score for the original dataset is 0.6926 and for microaggregation is 0.6890. DFTMicroagg produced 0.6942 and 0.6940 for coefficient of 72 and 24, respectively. In all cases of the clustering analysis, the proposed approach slightly outperformed the direct application of microaggregation algorithm based on the Silhouette scores obtained.

Privacy: Similarly, we obtained the same results (see Tables 21, 22, 23 and 24) as discussed in Sect. 6.2.2 for both record linkage and interval disclosure risks when both datasets were sorted in ascending and descending order. These results confirmed that sorting the datasets in ascending or descending order of households does not have any impact on the privacy results of the proposed approach as presented in the previous section.

6.3.2 Sampling rate

Recall that Dataset 1 and Dataset 2 were originally sampled at 5 and 10 minutes resolutions, respectively (see Table 1). In this section, we check the impact of re-sampling the datasets on utility and privacy using a different sampling rate. For Dataset 1, we re-sampled the energy consumptions of the individual households using 10 minutes sampling rate while 20 minutes was used for Dataset 2. Based on this sampling rate, the total columns of matrix X for Dataset 1 becomes 144 while that of Dataset 2 is 72. The upper and lower values

Table 24 Interval disclosure risk of microaggregation and DFTMicroagg on Dataset 2

k-Anonymity Level	Interval disclosure risk (%) Microaggregation	DFTMicroagg Coeff = 72	DFTMicroagg Coeff = 24
k = 2	41.02	29.69	19.13
k = 3	34.93	27.38	18.50
k = 4	31.31	25.88	18.07
k = 5	28.98	24.98	17.88

of coefficients used for Dataset 1 based on 10 minutes re-sampling are 72 and 24, respectively, while that of Dataset 2 based on 20 minutes re-sampling are 36 and 12, respectively. We conducted utility and privacy check based on the newly re-sampled datasets. The results obtained are summarized as follows.

Utility: As presented in Tables 25 and 26, both microaggregation and DFTMicroagg algorithms provided consistent results similar to those obtained in Sect. 6.2.1 for IL metric despite the fact that the original datasets have been re-sampled. Similarly, the proposed approach demonstrated consistent results with high level of data utility based on the clustering analysis for the re-sampled datasets. Again, the silhouette scores obtained is not less than 70% and 60% for the re-sampled Dataset 1 and Dataset 2, respectively.

For instance, considering Dataset 1 based on 10 minutes sampling rate, the Silhouette score for the original dataset when $k = 2$ is 0.7963 and for microaggregation is 0.7969. However, DFTMicroagg produced 0.7974 and 0.7982 for coefficient of 72 and 24, respectively. For Dataset 2 based on 20 minutes sampling rate, the Silhouette score for the original dataset when $k = 2$ is 0.6702 and for microaggregation is 0.6682. However, DFTMicroagg produced 0.6756 and 0.6947 for coefficient of 36 and 12, respectively. In all cases of the clustering analysis, the proposed approach outperformed the direct application of microaggregation algorithm based on the Silhouette scores. These results also confirmed the consistency of the proposed approach as a promising privacy protection mechanism.

Privacy: Again, Tables 27 and 28 further shows the effect of DFT introduced as additional layer for privacy protection. It can be seen that for each value of k, DFTMicroagg provides improved privacy guarantees over the direct application of microaggregation algorithm as a privacy protection mechanism. Similar to the results obtained in Sect. 6.2.2, DFTMicroagg algorithm prevents a significant amount of records from being linked based on the two scenarios investigated for the distance-based record linkage algorithm. The results of the privacy protection based on the re-sampled datasets improved over the previous results. For instance, as shown in Table 27, for the re-sampled Dataset 1, when $k = 2$ and $Coeff = 72$, DFTMicroagg prevented approximately additional 6,843 records in the masked dataset against record linkage attack based on the closest records when compare with direct application of microaggregation algorithm. For the second scenario of the distance-based record linkage attack, DFTMicroagg prevented additional 7,990 records. Similarly, when $k = 2$ and $Coeff = 24$, DFTMicroagg prevented additional 9,911 records based on the first scenario of the distance-based record linkage and 10,297 based on the second scenario. For both first and second scenarios of the distance-based record linkage attack and for each level of k-anonymity, DFTMicroagg outperformed the direct application of microaggregation algorithm.

Similarly, for the re-sampled Dataset 2, when $k = 2$ and $Coeff = 36$, DFTMicroagg prevented approximately additional 7,669 records in the masked dataset against record linkage attack based on the closest records when compare with direct application of microaggregation algorithm and 14,357 records based on the second scenario. Also, when $k = 2$ and $Coeff = 12$, DFTMicroagg prevented 11,513 records for the first scenario of the distance-based record linkage attack and 20,445 records for the second scenario. For each scenario of the distance-based record linkage attack and for each level of k-anonymity, the proposed approach outperformed microaggregation algorithm by preventing a significant number of records from being linked. These results show that DFTMicroagg algorithm can provide promising privacy guarantee as an effective privacy-preserving method over the direct application of microaggregation algorithm.

Based on interval disclosure risk attack, DFTMicroagg produced consistent results similar to those obtained earlier. We observed that the higher the value of k, the lower the disclosure risk based on this attack. Also, the lower the coefficient, the lower the disclosure risk. For both the re-sampled datasets 1 and 2, DFTMicroagg produced an improved result over the direct application of microaggregation algorithm (see Tables 29 and 30). The results based on interval disclosure risk using the re-sampled datasets also further confirmed the applicability of the proposed approach as a promising privacy-preserving mechanism for smart grid data.

7 Conclusion

In this paper, we demonstrate the possibility of estimating the utility–privacy trade-off of microaggregation and the proposed DFTMicroagg algorithm that is based on DFT and microaggregation to provide additional layer of privacy for protecting smart grid data. We evaluated the privacy values offered by microaggregation algorithm and based on our findings, we propose a dual-level anonymization method, which leverages the capability of DFT and microaggregation to enforce k-anonymity protection on time series daily energy consumption profiles. We analytically show that the proposed approach maintains a high level of utility with minimal information loss. The applicability of the proposed approach for different data mining tasks, such as clustering analysis, classification and energy load forecasting on the protected data have been discussed. We show that the proposed approach can provide the utility company with a more flexible option for dual-level masking of the energy data to be published. To ascertain the privacy improvement of the proposed approach over direct application of microaggregation algorithm, we

Table 25 Information loss (IL) of microaggregation and DFTMicroagg on Dataset 1 based on 10 minutes sampling rate

k-Anonymity Level	IL(Microagg)	IL(DFTMicroagg Coeff = 72)	IL(DFTMicroagg Coeff = 24)
k = 2	0.1163	0.1479	0.1848
k = 3	0.1424	0.1682	0.2005
k = 4	0.1781	0.1930	0.2168
k = 5	0.1994	0.2088	0.2282

Table 26 Information loss of microaggregation and DFTMicroagg on Dataset 2 based on 20 minutes sampling rate

k-Anonymity Level	IL(Microagg)	IL(DFTMicroagg Coeff = 36)	IL(DFTMicroagg Coeff = 12)
k = 2	0.0405	0.0929	0.1714
k = 3	0.0587	0.1014	0.1741
k = 4	0.0705	0.1076	0.1760
k = 5	0.0797	0.1120	0.1773

Table 27 Distance-based record linkage disclosure risk for each level of *k* based on microaggregation and DFTMicroagg for Dataset 1 based on 10 minutes sampling rate

	Disclosure risk (%)			
Methods	k = 2	k = 3	k = 4	k = 5
Microaggregation	48.63 (88.62)	31.56 (62.45)	22.89 (44.78)	17.40 (33.97)
DFTMicroagg (Coeff = 72)	29.83 (66.67)	20.29 (48.51)	11.51 (31.49)	9.35 (23.57)
DFTMicroagg (Coeff = 24)	21.40 (60.33)	15.45 (41.04)	9.87 (24.03)	5.16 (17.68)

Table 28 Distance-based record linkage disclosure risk for each level of *k* based on microaggregation and DFTMicroagg for Dataset 2 based on 20 minutes sampling rate

	Disclosure risk (%)			
Methods	k = 2	k = 3	k = 4	k = 5
Microaggregation	44.41 (77.80)	27.17 (51.23)	18.41 (34.12)	14.04 (26.06)
DFTMicroagg (Coeff = 36)	18.50 (29.29)	11.71 (19.96)	8.49 (14.96)	6.71 (12.04)
DFTMicroagg (Coeff = 12)	5.51 (8.72)	4.01 (6.71)	3.24 (5.62)	2.77 (4.86)

Table 29 Interval disclosure risk of microaggregation and DFTMicroagg on Dataset 1 based on 10 minutes sampling rate

	Interval disclosure risk (%)		
k-Anonymity Level	Microagg	DFTMicroagg Coeff = 72	DFTMicroagg Coeff = 24
k = 2	73.81	56.79	47.37
k = 3	63.97	51.69	45.02
k = 4	54.81	49.05	43.53
k = 5	50.96	46.89	42.50

Table 30 Interval disclosure risk of microaggregation and DFTMicroagg on Dataset 2 based on 20 minutes sampling rate

	Interval disclosure risk (%)		
k-Anonymity Level	Microagg	DFTMicroagg Coeff = 36	DFTMicroagg Coeff = 12
k = 2	52.60	42.21	33.29
k = 3	47.39	40.57	32.97
k = 4	44.48	39.63	32.72
k = 5	42.66	38.83	32.60

implement two attack models using distance-based record linkage and interval disclosure. The results obtain further confirm the efficacy of the proposed method. In future, we plan to investigate a suitable protection framework to protect smart grid data with multi-level smart meter readings, such as a dataset from utility company that has the total consumption aggregate as well as the consumption for each appliance used by the consumers at different levels of resolutions. In addition, we would like to investigate the case where DFT is applied after MDAV microaggregation algorithm to check the impact on the results both in terms of the utility and privacy guarantee.

Acknowledgements This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The first author is supported by the Kempe foundation.

Funding Open access funding provided by Umea University.

Data availability The two datasets that were used in this study are publicly available for download. The first dataset, 'EnerNOC GreenButton Data,' which was referred to as Dataset 1 is available for download at https://open-enernoc-data.s3.amazonaws.com/anon/index.html. The second dataset, 'Low Carbon London Electric Vehicle Load Profiles Data,' referred to as Dataset 2, is available for download at https://data.london.gov.uk/dataset/low-carbon-london- electric-vehicle-load-profiles. Table 1 summarizes the datasets.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Ethical approval This research does not involve human or animal. However, it involved anonymization of dataset which are publicly available. The dataset used in this research are publicly available online and the links to them have been provided in the article.

Informed consent There are no human subjects in this article and informed consent is not applicable.

References

1. Armoogum, S., Bassoo, V.: "Privacy of energy consumption data of a household in a smart grid," In: *Smart Power Distribution Systems* (Q. Yang, T. Yang, W. Li, eds.), 163–177, Academic Press, 2019
2. Chin, J.-X., De Rubira, T.T., Hug, G.: Privacy-protecting energy management unit through model-distribution predictive control. IEEE Trans. Smart Grid **8**(6), 3084–3093 (2017)
3. Karopoulos, G., Ntantogian, C., Xenakis, C.: Masker: masking for privacy-preserving aggregation in the smart grid ecosystem. Comput. Secur. **73**, 307–325 (2018)
4. Mashima, D., Serikova, A., Cheng, Y., Chen, B.: Towards quantitative evaluation of privacy protection schemes for electricity usage data sharing. ICT Expr. **4**(1), 35–41 (2018)
5. Kapoor, S., Sturmberg, B., Shaw, M.: "A review of publicly available energy data sets," Wattwatchers' My Energy Marketplace (MEM)(The Australian National University, p. 2020. Canberra, Australia (2020)
6. Soykan, E.U., Bilgin, Z., Ersoy, M.A., Tomur, E.: "Differentially private deep learning for load forecasting on smart grid," in *2019 IEEE Globecom Workshops (GC Wkshps)*,1–6, IEEE, 2019
7. Commission, E.: "Smart grids and meters." https://ec.europa.eu/energy/topics/markets-and-consumers/smart-grids-and-meters_en?redir=1, 2021
8. Tudor, V., Almgren, M., Papatriantafilou, M.:"A study on data de-pseudonymization in the smart grid," In: Proceedings of the Eighth European Workshop on System Security, 1–6, 2015
9. Dong, R., Ratliff, L.J.: "Energy disaggregation and the utility-privacy tradeoff," In: Big data application in power systems (R. Arghandeh and Y. Zhou, eds.), 409–444, Elsevier, 2018
10. Efthymiou, C., Kalogridis, G.: "Smart grid privacy via anonymization of smart metering data," In: 2010 first IEEE International conference on smart grid communications, 238–243, IEEE, 2010
11. Cleemput, S., Mustafa, M.A., Marin, E., Preneel, B.:"De-pseudonymization of smart metering data: Analysis and countermeasures," In: 2018 Global internet of things summit (GIoTS), 1–6, IEEE, 2018
12. Sangogboye, F.C., Jia, R., Hong, T., Spanos, C., Kjærgaard, M.B.: A framework for privacy-preserving data publishing with enhanced utility for cyber-physical systems. ACM Trans. Sens. Netw. (TOSN) **14**(3–4), 1–22 (2018)
13. BBCNews, "Ukraine power cut 'was cyber-attack'." https://www.bbc.com/news/technology-38573074, 2017
14. Jia, R., Sangogboye, F.C., Hong, T., Spanos, C., Kjærgaard, M.B.: "Pad: protecting anonymity in publishing building related datasets," In: Proceedings of the 4th ACM International conference on systems for energy-efficient built environments, 1–10 (2017)
15. Thouvenot, V., Nogues, D., Gouttas, C.: "Data-driven anonymization process applied to time series," In: SIMBig, 80–90 (2017)
16. Zhang, L., Zhao, L., Yin, S., Chi, C.-H., Liu, R., Zhang, Y.: A lightweight authentication scheme with privacy protection for smart grid communications. Fut. Generat. Comput. Sys. **100**, 770–778 (2019)
17. Ali, W., Din, I.U., Almogren, A., Guizani, M., Zuair, M.: A lightweight privacy-aware iot-based metering scheme for smart industrial ecosystems. IEEE Trans. Ind. Info. **17**(9), 6134–6143 (2020)
18. Alharbi, K.N., Lin, X., Shao, J.: A privacy-preserving data-sharing framework for smart grid. IEEE Intern. Things J. **4**(2), 555–562 (2016)
19. Domingo-Ferrer, J., Torra, V.: "A quantitative comparison of disclosure control methods for microdata," Confidentiality, disclosure and data access: theory and practical applications for statistical agencies, 111–134 (2001)

20. Bovornkeeratiroj, P., Iyengar, S., Lee, S., Irwin, D., Shenoy, P.: "Repel: A utility-preserving privacy system for iot-based energy meters," In: 2020 IEEE/ACM Fifth International conference on internet-of-things design and implementation (IoTDI), 79–91, IEEE,(2020)
21. Salas, J., Torra, V.: "A general algorithm for k-anonymity on dynamic databases," In: Data privacy management, cryptocurrencies and blockchain technology
22. Nin, J., Torra, V.: "Extending microaggregation procedures for time series protection," In: International conference on rough sets and current trends in computing, 899–908, Springer, Berlin, 2006
23. Torra, V.: Data privacy: foundations, new developments and the big data challenge, 1st edn. Springer, Berlin (2017)
24. Romdhane, R.B., Hammami, H., Hamdi, M., Kim, T.-H.: "A novel approach for privacy-preserving data aggregation in smart grid," in *2019 15th international wireless communications & Mobile Computing Conference (IWCMC)*, 1060–1066, IEEE, 2019
25. Cao, H., Liu, S., Zhao, R., Gu, H., Bao, J., Zhu, L.: "A privacy preserving model for energy internet base on differential privacy," In: 2017 IEEE International conference on energy internet (ICEI), pp. 204–209, IEEE, 2017
26. Tudor, V., Almgren, M., Papatriantafilou, M.:"Analysis of the impact of data granularity on privacy for the smart grid," In: Proceedings of the 12th ACM workshop on workshop on privacy in the electronic society, 61–70 (2013)
27. Wen, L., Zhou, K., Yang, S., Li, L.: Compression of smart meter big data: a survey. Renew. Sustain. Energy Rev. **91**, 59–69 (2018)
28. Huang, X., Hu, T., Ye, C., Xu, G., Wang, X., Chen, L.: Electric load data compression and classification based on deep stacked auto-encoders. Energies **12**(4), 653 (2019)
29. Plenz, M., Dong, C., Grumm, F., Meyer, M.F., Schumann, M., McCulloch, M., Jia, H., Schulz, D.: Framework integrating lossy compression and perturbation for the case of smart meter privacy. Electronics **9**(3), 465 (2020)
30. Feng, X., Lan, J., Peng, Z., Huang, Z., Guo, Z.: "A novel privacy protection framework for power generation data based on generative adversarial networks," In: 2019 IEEE PES Asia-Pacific power and energy engineering conference (APPEEC), pp. 1–5, IEEE, 2019
31. Khwaja, A.S., Anpalagan, A., Naeem, M., Venkatesh, B.: Smart meter data obfuscation using correlated noise. IEEE Intern. Things J. **7**(8), 7250–7264 (2020)
32. Samarati, P., Sweeney, L.: "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," *SRI Intl. Tech. Rep.*, 1998
33. Samarati, P.: Protecting respondents identities in microdata release. IEEE Trans. Knowl. Data Eng. **13**(6), 1010–1027 (2001)
34. Sweeney, L.: k-anonymity: a model for protecting privacy. Int J Uncertain Fuzzin. Knowl.-Bas. Sys. **10**(05), 557–570 (2002)
35. Torra, V., Navarro-Arribas, G.: Data privacy. Wiley Interdisciplinary Reviews: data Mining and Knowledge Discovery. vol. 4, issue 4, pp. 269–280, Wiley, Hobroken, 2014
36. Alarte Aleixandre, J.: "Application of clustering techniques to privacy protection," Thesis, Universitat Oberta de Catalunya, 2018
37. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Min. Knowl. Discov. **11**(2), 195–212 (2005)
38. Cooley, J.W., Lewis, P.A., Welch, P.D.: The fast fourier transform and its applications. IEEE Trans. Educat. **12**(1), 27–34 (1969)
39. Wang, Z.: Fast algorithms for the discrete w transform and for the discrete fourier transform. IEEE Trans. Acoust., Speech, Signal Process. **32**(4), 803–816 (1984)
40. Watson, A.B.: Image compression using the discrete cosine transform. Math. J. **4**(1), 81 (1994)
41. Liao, X., Li, K., Yin, J.: Separable data hiding in encrypted image based on compressive sensing and discrete fourier transform. Multim. Tool. Appl. **76**(20), 20739–20753 (2017)
42. Weinstein, S., Ebert, P.: Data transmission by frequency-division multiplexing using the discrete fourier transform. IEEE Trans. Commun. Technol. **19**(5), 628–634 (1971)
43. Yancey, W.E., Winkler, W.E., Creecy, R.H.: "Disclosure risk assessment in perturbative microdata protection," In: Inference control in statistical databases, pp. 135–152, Springer, Berlin, 2002
44. Arthur, D., Vassilvitskii, S.: "k-means++: The advantages of careful seeding," In: Eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 1027–1035, ACM, 2007
45. Hyndman, R.J., Athanasopoulos, G.: Forecasting: principles and practice. OTexts, Melbourne, Australia (2018)