STL-DP: Differentially Private Time Series Exploring Decomposition and Compression Methods

Kyunghee Kim[1], Minha Kim[2] and Simon Woo[3,*]

[1]Department of Statistics, Sungkyunkwan University, Seoul, Korea
[2]Department of Artifical Intelligence, Sungkyunkwan University, Suwon, Korea
[3]Department of Applied Data Science, Sungkyunkwan University, Suwon, Korea

Abstract

As time series data is collected and used in a variety of fields, the importance of *preserving privacy* on time series is also on the increase. This paper is a preliminary study of the Differential Privacy (DP) algorithm specially designed to provide privacy to time series data by integrating the time series decomposition technique. In particular, this study extends the Fourier Perturbation Algorithm (FPA) with **S**easonal and **T**rend decomposition using **L**OESS (STL). In this work, we propose STL-DP, which first performs STL decomposition to the original data. Then we apply the FPA only to the core part of the time series, particularly trend or seasonal components, to provide privacy. In this preliminary study, we show that our approach consistently outperforms other baselines in terms of utility according to the experimental results. Our code is available at https://github.com/Privacy-DASH/STL-DP.

Keywords

Differential Privacy, Time Series, Fourier Perturbation Algorithm, STL Decomposition

1. Introduction

Recently the need for providing data privacy has significantly increased, as the quantity of data is growing at an unprecedented speed, and a trend to make such large data accessible to the public is also growing. To share data and use them for multiple tasks, ensuring data privacy is crucial. Therefore, many privacy protection techniques have been proposed and researched, such as Differential Privacy (DP) [1], Homomorphic Encryption [2], and Generative Adversarial Network (GAN) [3].

However, despite the vulnerability of time series data due to their widespread application in various fields, privacy-preserving mechanisms on time series data have not been extensively investigated yet [4]. In this paper, we consider and propose a DP mechanism specially designed to protect the privacy on time series data. One of the unique characteristics of time series is that it exhibits a strong correlation among successive values. Accordingly, if the adversary knows the approximate time information, information leakage can occur through contextual understanding, as shown by other research works [5, 6]. However, existing perturbation methods such as Gaussian Perturbation Algorithm (GPA) [7], and Laplace Perturbation Algorithm (LPA) [8] do not consider the temporal correlations across time.

In this paper, we propose a simple yet intuitive differentially private (DP) mechanism, STL-DP, which can effectively mitigate the aforementioned problem. We assume that time series data can be decomposed into trend and seasonal components, and such information can be considered sensitive by the data providers because they present the overall ups and downs and periodic patterns. Therefore, it is of great importance to maintain such information private throughout the entire data mining process [5]. We explore **S**easonal and **T**rend decomposition using **L**OESS (STL) [9] for decomposing time series which leverages Local regression (LOESS) [10]. Our proposed mechanism STL-DP consists of two stages that effectively hide the trend or seasonality of the time series data. This mechanism enables us to improve the utility while maintaining privacy by concealing the primary components of the time series.

The main contributions of our work are summarized as follows:

- We propose STL-DP that considers the unique characteristics of time series to provide the most suitable privacy protection method for time series.
- We show that STL-DP effectively protects the core parts of the time series data under the same privacy budget, thereby significantly improving utility over the existing methods.

CIKM22-PAS: The 1st International Workshop on Privacy Algorithms in Systems @ CIKM'22, Oct. 17-22, 2022, Atlanta, Georgia, USA
*Corresponding author.
kkh97122647@gmail.com (K. Kim); sunshine01@g.skku.edu (M. Kim); swoo@g.skku.edu (S. Woo)
0000-0002-5129-1325 (K. Kim); 0000-0002-3224-0610 (M. Kim); 0000-0002-8983-1542 (S. Woo)

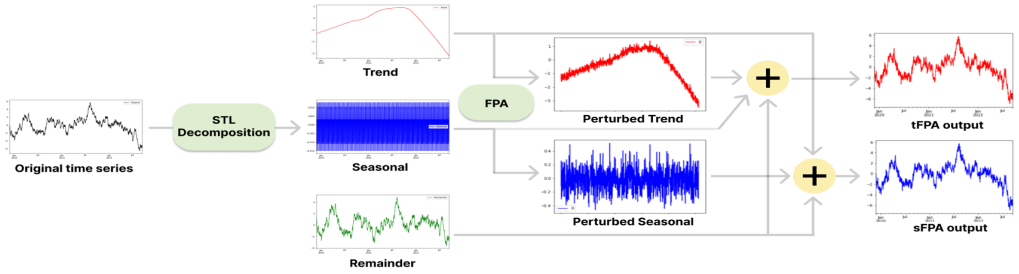

Figure 1: The overview of our proposed STL-DP.

2. Preliminaries

2.1. ϵ - Differential Privacy

Differential Privacy [1] ensures no significant change in the query response, whether a particular individual is in a database or not [11].

Definition. There are two databases $D, D^{'}$ which satisfy $\|D - D^{'}\|_1 \leq 1$. D denotes composed data of U individual users, i.e., $D = \cup_{i=1}^{U} D_i$, and the data of any single user can be put as D_i. Let us denote M and ϵ as some randomized function and a privacy budget, respectively. M guarantees ϵ-privacy if and only if it satisfies the following Eq. (1):

$$P[M(D) \in S] \leq e^{\epsilon} \times P[M(D^{'}) \in S], \forall S \in Range(M) \quad (1)$$

DP mechanism aims to keep the query response for each $D, D^{'}$ the same, despite having one or fewer non-overlapping individuals. Specifically, the smaller the ϵ is, the higher the privacy protection of the data becomes [12].

2.2. DP Algorithms for Time Series

Laplace Perturbation Algorithm (LPA). LPA [8] adds independent noise generated from the Laplace distribution [1]. LPA is renowned for its simplicity but it is unsuitable for protecting time series because of its independent noise injection.

Fourier Perturbation Algorithm (FPA). FPA is a compression-based method that first applies the Discrete Fourier Transform (DFT) to the true query answers, then performs LPA to Fourier coefficients [8]. The perturbed coefficients undergo the inverse DFT (IDFT) to obtain the resulting perturbed sequence. The entire process can be expressed as $perturbed\ f(D) = IDFT(LPA(DFT(f(D))))$, where f is a function that maps each individual $D_1, D_2, \ldots, D_U$ to numbers. The DFT and IDFT for the j^{th} element of the series is defined as (2):

$$DFT(f(D))_j = \sum_{i=1}^{n} e^{\frac{2\pi\sqrt{-1}}{n} ji} f(D)_i, \quad (2)$$

$$IDFT(f(D))_j = \frac{1}{n} \sum_{i=1}^{n} e^{-\frac{2\pi\sqrt{-1}}{n} ji} f(D)_i$$

As compression methods convert the series from time to frequency domain, noises injected in the frequency domain are no longer independent but are correlated. For this reason, FPA is better suited for perturbing time series [13], and we extend the FPA-based method in our work.

2.3. Seasonal and Trend decomposition using LOESS (STL)

There are various time series decomposition methods such as classical decomposition [14], X11 [15], and STL [9]. The classical method is simple to implement but is inapplicable since some data from both ends of the sequence are lost. X11 successfully tackled the problem of data loss but is still limited in use as it can only handle monthly or quarterly data. On the other hand, STL effectively handles the problems mentioned above. STL is a flexible and robust time series decomposition method that leverages local regression (LOESS).

3. Our Approach

We propose STL-DP to protect core information of the time series while improving utility within a predefined privacy budget. Refer to Figure 1 for a glance at our proposed STL-DP.

The main difference of STL-DP with the existing methods is the integration of STL decomposition. First, by incorporating the decomposition phase, we can identify the core components, such as the trend and seasonality of

Table 1
Euclidean distance between the original and the perturbed time series.

Algorithm	Zone	epsilon1	epsilon2	epsilon3	epsilon4
LPA	Zone1	1.3418E+4	2.6102E+3	1.3052E+3	2.6694E+2
FPA		3.0351E+0	4.7537E+0	2.2658E+0	7.6124E-8
sFPA		9.6445E+0	**1.2354E-7**	**1.4956E+0**	**2.3439E-8**
tFPA		**6.1856E-7**	**5.2730E-7**	1.8135E+0	7.6125E-8
LPA	Zone2	1.3004E+4	2.7051E+3	1.3226E+3	2.6037E+2
FPA		4.9405E+0	1.3781E+0	1.7282E+0	0.3781E+0
sFPA		**1.6598E-7**	1.9342E+0	**6.4095E-7**	**4.4947E-8**
tFPA		**4.7581E-7**	**1.0069E-7**	1.2109E+0	0.3782E+0
LPA	Zone3	1.2928E+4	2.6534E+3	1.3374E+3	2.7253E+2
FPA		**2.7932E-7**	3.8466E+0	1.1964E-6	**2.5183E-7**
sFPA		1.3435E+1	**2.4369E-7**	1.6377E+0	0.2554E+0
tFPA		2.1689E+1	2.2007E+0	**9.7750E-8**	**2.5183E-7**

Table 2
Comparison of ▵MAPE for each mechanism; ▵MAPE = |MAPE (DP algorithm) - MAPE (Original Series)|

		epsilon1	epsilon2	epsilon3	epsilon4		epsilon1	epsilon2	epsilon3	epsilon4
LPA	Linear	0.7476	0.0081	0.0125	0.0068	SimpleDNN	2.2842	1.6252	0.5990	0.2236
FPA		**0.0013**	**0.0002**	0.0002	0.0001		1.5972	**0.3159**	0.7612	**0.1213**
tFPA		0.0017	0.0005	0.0010	**0.0000**		**0.0602**	1.0390	1.0390	0.2152
sFPA		0.0094	0.0014	**0.0001**	**0.0000**		1.5158	1.7738	**0.1083**	0.2468
LPA	RNN	1.4877	0.0194	0.0437	0.0084	Transformer	1.1464	0.5750	0.4180	0.2477
FPA		0.0169	0.0089	**0.0004**	0.0094		**0.0738**	0.1268	**0.0860**	0.0941
tFPA		**0.0072**	0.0081	0.0042	**0.0004**		0.2325	**0.0776**	0.2083	**0.0271**
sFPA		0.0117	**0.0070**	0.0007	0.0009		0.0864	0.4175	0.1468	0.0832

time series data, which may contain critical information and are prone to attacks. One of the STL-DP mechanisms is referred to as *sFPA*, which is a method that injects noises only to the seasonal part of the decomposed series. Similarly, the approach of performing perturbations only on the trend is named *tFPA*. Lastly, the perturbed components from the seasonal or trend parts are combined with the rest of the unperturbed components to reconstruct the form of the sequence.

4. Experimental Results

Methods. We demonstrate the effectiveness of the proposed STL-DP by comparing the utility of our sFPA and tFPA with two baselines, LPA and FPA. Herein, we introduce two different metrics to quantify utility. The first metric is the Euclidean distance between the original and the perturbed series. Nextly, the original and the perturbed data are each fed into the forecasting model to evaluate the respective Mean Absolute Percentage Error (MAPE), and the difference between the two MAPEs is used as the second metric.

Dataset. We used the power consumption data from 2017-01-01 to 2017-12-31 of three zones of Tetouan city located in northern Morocco [16]. The properties of the dataset are summarized as follows:

- Prediction variables : Power consumption of zone 1, 2, and 3 of Tetouan city with additional information, including temperature, humidity, wind speed, general diffuse flows, and diffuse flows.
- Data information : Aggregated from 550,374 inhabitants according to Morocco Census [16].

Throughout the experiment, we set the privacy budget $\epsilon 1 - \epsilon 4$ as 0.48, 2.4, 4.8, and 24, respectively, and the sensitivity as 48, which are experimental settings taken from Günther, et al. [17].

Euclidean distance results. The degree of closeness between the original and the perturbed series under the same privacy budget can be interpreted as the level of utility. As shown in Table 1, LPA yields a greater distance than other methods, confirming that FPA-based methods are better than LPA. Furthermore, our tFPA and sFPA are consistently ranked as the best algorithm in terms of Euclidean distance. These results indicate the superiority of STL-DP over the baselines.

Comparison on forecasting performances. Recall that our objective is to generate noise-injected series that minimize the performance drop of the forecasting model. As shown in Table 2, we used four models, from a simple feed-forward neural network to advanced models such as LSTM and Transformer as our forecasting model. The models were trained to predict the upcoming 20 timesteps given the past 60 timesteps. In most cases, as ϵ grew, the forecasting error difference (▵MAPE) decreased. Not

surprisingly, both sFPA and tFPA outperformed other DP mechanisms for a majority of models in terms of $\triangle$MAPE.

5. Conclusion and Future Work

As preliminary research, we introduce an effective DP mechanism, STL-DP, specially designed for generating privacy-protected time series data. As the experiment results suggested, the distance between the original series and the perturbed series from sFPA and tFPA was much closer than the other baselines. Also, the difference between the MAPE of the original and the perturbed series was significantly lower for our proposed sFPA and tFPA than for other perturbation algorithms. Therefore, we showed that considering the unique property of time series data improves the utility under the same privacy budget. For future works, we plan to extend our research by designing a more advanced mechanism FPA_k, that uses only the k ($<n$) Fourier coefficients as targets of the perturbation.

Acknowledgments

The work was supported by the affiliated institute of ETRI [2022-075]. Also, this work was partially supported by the Basic Science Research Program through National Research Foundation of Korea (NRF) grant funded by the Korean Ministry of Science and ICT (MSIT) under No. 2020R1C1C1006004 and Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean MSIT: (No. 2022-0-01199, Graduate School of Convergence Security at Sungkyunkwan University), (No. 2022-0-01045, Self-directed Multi-Modal Intelligence for solving unknown, open domain problems), (No. 2022-0-00688, AI Platform to Fully Adapt and Reflect Privacy-Policy Changes), (No. 2021-0-02068, Artificial Intelligence Innovation Hub), (No. 2019-0-00421, AI Graduate School Support Program at Sungkyunkwan University), and (No. 2021-0-02309, Object Detection Research under Low Quality Video Condition).

References

[1] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of cryptography conference, Springer, 2006, pp. 265–284.

[2] R. L. Rivest, L. Adleman, M. L. Dertouzos, et al., On data banks and privacy homomorphisms, Foundations of secure computation 4 (1978) 169–180.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).

[4] S. Papadimitriou, F. Li, G. Kollios, P. S. Yu, Time series compressibility and privacy, in: Proceedings of the 33rd international conference on Very large data bases, Citeseer, 2007, pp. 459–470.

[5] Y. Zhu, Y. Fu, H. Fu, On privacy in time series data mining, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2008, pp. 479–493.

[6] Y. Zhu, Y. Fu, H. Fu, A new class of attacks on time series data mining, Intelligent Data Analysis 14 (2010) 405–418.

[7] N. U. Sheikh, H. J. Asghar, F. Farokhi, M. A. Kaafar, Do auto-regressive models protect privacy inferring fine-grained energy consumption from aggregated model parameters, IEEE Transactions on Services Computing (2021).

[8] V. Rastogi, S. Nath, Differentially private aggregation of distributed time-series with transformation and encryption, in: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, 2010, pp. 735–746.

[9] R. B. Cleveland, W. S. Cleveland, J. E. McRae, I. Terpenning, Stl: A seasonal-trend decomposition, J. Off. Stat 6 (1990) 3–73.

[10] W. G. Jacoby, Loess:: a nonparametric, graphical tool for depicting relationships between variables, Electoral studies 19 (2000) 577–613.

[11] W. Huang, S. Zhou, T. Zhu, Y. Liao, Improving utility of differentially private mechanisms through cryptography-based technologies: a survey, arXiv preprint arXiv:2011.00976 (2020).

[12] J. Wang, S. Liu, Y. Li, A review of differential privacy in individual data release, International Journal of Distributed Sensor Networks 11 (2015) 259682.

[13] H. Wang, Z. Xu, Cts-dp: publishing correlated time-series data via differential privacy, Knowledge-Based Systems 122 (2017) 167–179.

[14] J. Cohen, W. Gorr, C. Durso, Estimation of crime seasonality: a cross-sectional extension to time series classical decomposition, H. John Heinz III Working Paper (2003).

[15] A. Sutcliffe, X11 time series decomposition and sampling errors, Australian Bureau of Statistics, 1993.

[16] A. Salam, A. El Hibaoui, Comparison of machine learning algorithms for the power consumption prediction:-case study of tetouan city–, in: 2018 6th International Renewable and Sustainable Energy Conference (IRSEC), IEEE, 2018, pp. 1–5.

[17] G. Eibl, K. Bao, P.-W. Grassal, D. Bernau, H. Schmeck, The influence of differential privacy on short term electric load forecasting, Energy Informatics 1 (2018) 93–113.