IEICE TRANS. INF. & SYST., VOL.E105–D, NO.7 JULY 2022

LETTER

PRIGM: Partial-Regression-Integrated Generic Model for Synthetic Benchmarks Robust to Sensor Characteristics*

Kyungmin KIM[†], **Jiung SONG**[††], ***and*** **Jong Wook KWAK**[†a)], ***Nonmembers***

SUMMARY We propose a novel synthetic-benchmarks generation model using partial time-series regression, called *Partial-Regression-Integrated Generic Model* (*PRIGM*). *PRIGM* abstracts the unique characteristics of the input sensor data into generic time-series data confirming the generation similarity and evaluating the correctness of the synthetic benchmarks. The experimental results obtained by the proposed model with its formula verify that *PRIGM* preserves the time-series characteristics of empirical data in complex time-series data within 10.4% on an average difference in terms of descriptive statistics accuracy.

key words: *internet of things, synthetic data generation, synthetic benchmarks, ARIMA, time-series analysis*

1. Introduction

To analyze the performance of the IoT or other application devices, researchers are expected to provide various types of sensor data; UCI Repository and KEEL are representative datasets mainly used by AI researchers as data processing sources to evaluate the performance of their techniques. Although many publicly available datasets have emerged, their structures and selection criteria are inconsistent, and a detailed walk-through is not provided. Therefore, selecting an appropriate datasets for performance evaluation becomes another unnecessary burden on researchers [1], [2].

As the use of the data retrieved by the repository has such difficulties and limitations, a tool for generating a standardized synthetic data set exactly matched for the specifics of application sensing data for the convenience of research and verification is necessary. Consequently, many researchers have mainly attempted to generate synthetic benchmarks by implementing a *hidden Markov model* (HMM) or statistical regression. Arlitt et al. used an HMM to generate power data over several months as a synthetic benchmark [3]. Liu et al. proposed regression and probability-based energy consumption time-series data generation [4].

However, existing synthetic benchmark generation methods have several limitations: (1) Existing synthetic data generation studies are limited to generating data on a specific data type, such as energy consumption and meteorological data. That requires additional pre-tuning analysis for specific data, and there are restrictions for automating and using it universally. It means that these cannot be adopted in sensor data of IoT devices with various characteristics and radical sensing value changes. (2) In previous studies, the similarity between empirical data and synthetically generated data has been compared by using only descriptive statistics. However, descriptive statistics cannot compare the difference of time-dependent changes in values. Therefore, the similarity of time-series characteristics cannot be guaranteed by central tendency and measures of variability, especially for data with non-periodic trends.

Therefore, in this letter, we propose a novel synthetic-benchmark generation and verification model that is independent of sensor data characteristics. The proposal takes the strategy of creating parameters by dividing empirical data that have complex time-series characteristics. As a result, without considering specific data target, it can generate various time-series data and automatically generate data with non-periodic tendencies as well as periodic data for complex characteristic data type. In addition, *PRIGM* can quantitatively compare the differences of time-series data characteristics using PoR. Our main contributions are as follows:

(1) *Description of partial-regression-integrated generic model*. We describe a new benchmark generation technique for decomposing, verifying, and evaluating empirical data. Through this process, *PRIGM* can model generic time-series data, regardless of the sensor data characteristics.

(2) *New quantitative evaluation criteria for synthetic benchmarks*. Evaluation using descriptive statistics cannot reflect trends in time-series data, and verifying models using deviations is not suitable for synthetic benchmarks. To address this, we present a new concept of *possibility of reproducibility (PoR)* that interprets individual signals and evaluates the similarity with empirical data through Fourier transform.

(3) *Realization of benchmark ubiquity*. Our model can generate and verify synthetic data using only its parameters, enabling researchers to perform their evaluation and mutual verification by simply using and sharing the model parameters, without considering the experimental environment and empirical data.

Manuscript received December 16, 2021.
Manuscript revised March 15, 2022.
Manuscript publicized April 4, 2022.

[†]The authors are with the Department of Computer Engineering, Yeungnam University, Gyeongsan, Korea.
[††]The author is with the Biometrics and Statistics, SYMYOO CO., LTD., Yongsan-gu, Seoul, South Korea.

*This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1F1A1065788).

a) E-mail: kwak@yu.ac.kr (Corresponding author)

DOI: 10.1587/transinf.2021EDL8113

2. Motivation

We have previously addressed the problem of modeling data from acceleration sensors or gyroscopes using specific parameters. These problems arise from the following observations. Figure 1 presents the sensor data mainly utilized in IoT devices based on their characteristics. Figure 1 (a-1) shows the temperature data, Fig. 1 (a-2) shows the appliance power consumption, Fig. 1 (b-1) shows the acceleration sensor, and Fig. 1 (b-2) shows the Dau Index shares of stock. The characteristics of the sensor data shown in Fig. 1 are as follows:

Frequency: Frequency is the reciprocal of the number of data inputs during the same period. A high frequency amplifies variability and regularity by providing a large number of values per unit of time. In contrast, a low frequency mitigates the effect of variability and regularity.

Variability: Variability encompasses the sensor's bit depth and the magnitude of the temporal variation of the sensor value. High variability means that the sensor's range of values is broad, and value changes are rapid and frequent. Low variability indicates that the sensor unit has a low and gradual change degree.

Figures 1 (a-1) and 1 (a-2) show a low frequency and variability. In previous studies, data, such as climate data and energy grids, have been examined, and these can easily define patterns through statistical modeling or machine learning. Figures 1 (b-1) and 1 (b-2) show a high frequency and variability and they include activity recognition or stock indices, the data of which volume is rapidly increasing. We observed the reason why modeling based on previous studies is difficult for data with high variability, such as these types. The main reason is that the time-series characteristics and critical patterns of data are inconsistently scattered in a specific section rather than the entire input dataset. However, the proposed model, *PRIGM*, is used as a primary motive to find a specific section with consistent time-series characteristics and divide it into several segments to generate a composite benchmark from such complex data.

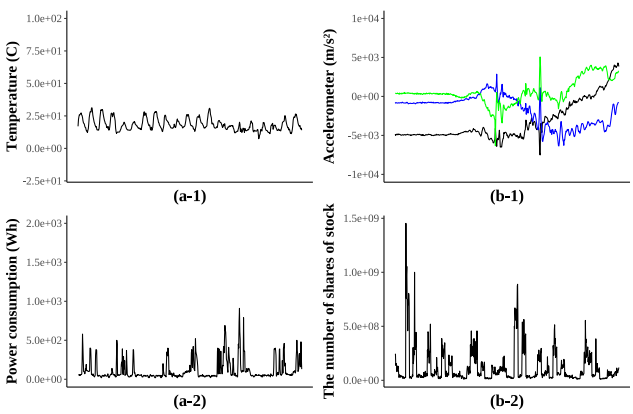


Fig. 1 Various empirical sensor data.

3. Partial-Regression-Integrated Generic Model

To generate a synthetic benchmark, *PRIGM* employs three processes: segmentation, modeling, and verification. Segmentation divides the *empirical data set* (EDS) into *empirical segments* (ES_i) based on moving average and variance changes. In the modeling process, for each ES_i, *PRIGM* defines the ARIMA model parameters for a synthetic benchmark using the Hyndman–Khandakar algorithm. The parameters generated are recorded for each *synthetic segment class* (SSC_i). *Synthetic definition class* (SDC) reserves metadata, such as the number of SSC_i or categories of benchmarks. *PRIGM* creates a *segment set* (SS_i) by configuring the parameters defined in SSC_is into the ARIMA model. To verify that the generated data are appropriate for the intend of the modeling process, each SS_i corresponding to ES_i is decomposed by applying Fourier transform for frequency components, and the result is quantified to measure the similarity. Finally, the combination of SS_i constitutes a *synthetic benchmark* (SB). Algorithm 1 shows the overall procedure.

First, in the segmentation process, *PRIGM* aims to generate synthetic benchmarks by abstracting different types of time-series data into the same interface. Through segmentation, *PRIGM* divides EDS into ES_i set to preserve each section's time-series characteristics for highly variable data to allow each segment to reproduce a time-series model (line 3). The segmentation process divides the interval based on the mean-variance gap of the time-series data. If the segmentation process cannot determine an appropriate parame-

Algorithm 1 PRIGM Procedure

Ensure: The minimum *possibility of reproducibility* required is p, the segmentation group number is i;
Input: *empirical data set* (EDS)
Output: *synthetic benchmark* (SB)
1: do
2: //Devide EDS with the change point in segmentation.
3: $ES_i \leftarrow$ Segmentation(EDS)
4: for i to length(ES_i) then
5: //Modeling parameters for synthetic data generation.
6: $SSC_i \leftarrow$ Modeling(ES_i);
7: //Test model for time-series consistency.
8: $SS_i \leftarrow$ ARIMA(SSC_i.AR, SSC_i.I, SSC_i.MA);
9: //Calculate PoR value in verification process.
10: PoR $\leftarrow$ Verification(ES_i, SS_i);
11: if PoR $< p$ then
12: //Subdivide the current segment recursively.
13: Segmentation(ES_i);
14: else
15: //Assign the metadata of the segment class.
16: assign SSC_i metadata (ID, seq, len ...);
17: SDC.len $\leftarrow$ SDC.len + 1;
18: end if
19: end for
20: //Assign the metadata of the definition class
21: assign SDC metadata (ID_{SDC}, ID_{type}, Segment ...);
22: SB $\leftarrow$ ARIMA for each SDC segment;
23: Verification(EDS, SB) $< p$ while

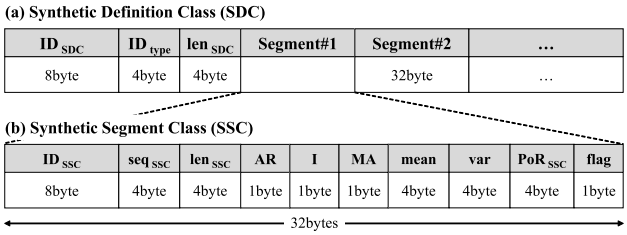


Fig. 2 Data structure for handling SDC and SSC.

ter in the subsequent modeling process, *PRIGM* performs the segmentation process recursively, and all intervals or the mean-variance difference are within the margin of error (line 10–13).

Second, in the modeling process, *PRIGM* automatically defines that the ARIMA parameters describe the segmented ES_i section in the empirical data set using the Hyndman–Khandakar algorithm (line 6). Each ES_i creates an SSC_i composed of the corresponding ARIMA model and SDC for managing the set of SSC_i corresponding to EDS (line 8). The structures of SDC and SSC are shown in Fig. 2. SDC is a set of SSC_is and corresponds to a unique empirical benchmark. It is an interface for generating synthetic benchmarks and consists of metadata for each category, such as an identifier for structural composition (ID_{SDC}), a category ID for classifying values in the synthetic benchmark generation service afterward (ID_{type}), and the segment length included in SDC, meaning the number of SSC_i included (len_{SDC}). Each set of SSC_is, SSC_i is a 32 byte record consisting of SSC identifiers for data management (ID_{SSC}), sequences to store relative position in SDC (seq_{SSC}), empirical data lengths of ES_i to be modeled (len_{SSC}), parameters used for ARIMA modeling (AR, I, MA, mean, and variance), the highest PoR value measured in the model (PoR), and flags to set temporary characteristics (line 16–17).

In the *PRIGM* process for generating synthetic data, the data subset from using ARIMA parameters stored in SSC_i corresponding to ES_i is called SS_i. As a set of these SS_is, the final SB is constructed using the SDC corresponding to the EDS (line 21–22).

Finally, we propose the *possibility of reproducibility* (PoR), which is the criterion for evaluation in the verification process. The main motivation of PoR is based on the following two requirements. (1) The performance comparison of the generated synthetic benchmark is based on the similarity with the empirical data. However, the descriptive statistics presented in previous studies cannot compare the differences in time-series data variation of professional benchmark data with long measurement intervals. (2) When the appropriate number of SSC_i is not determined during the segmentation process, it damages the complexity and stability of the entire model. As the input data characteristics are different for each benchmark and sensor type, a standardized method is required to adapt and respond to the input data.

In *PRIGM*, to verify these time-series data, SS_is are substituted with the sum of the frequency components. The values of the empirical and frequency components generated are converted into quantified indices to check the similarity between them. The input sensor data are converted into discrete data through quantization. Therefore, ES_i, the frequency decomposition section of the partial empirical data, and SS_i, the synthesized data generated by regressive modeling, are frequency-decomposed into discrete Fourier transform. The Fourier transform of the sampled data for discrete time t is given by Eq. (1).

$$f(t) = \sum_{j=0}^{N-1} c_j e^{ijt2\pi/N}, \quad t = 0, \ldots, N-1. \tag{1}$$

$$\mathrm{PoR}(\mathrm{seg}, i) = \frac{\sum_{i=0}^{N} \mathrm{FSS}_i^2}{\sum_{i=0}^{len_{SSC}} (\mathrm{FSS}_i - \mathrm{FSE}_i)^2 + \sum_{i=0}^{N} \mathrm{FSS}_i^2} \tag{2}$$

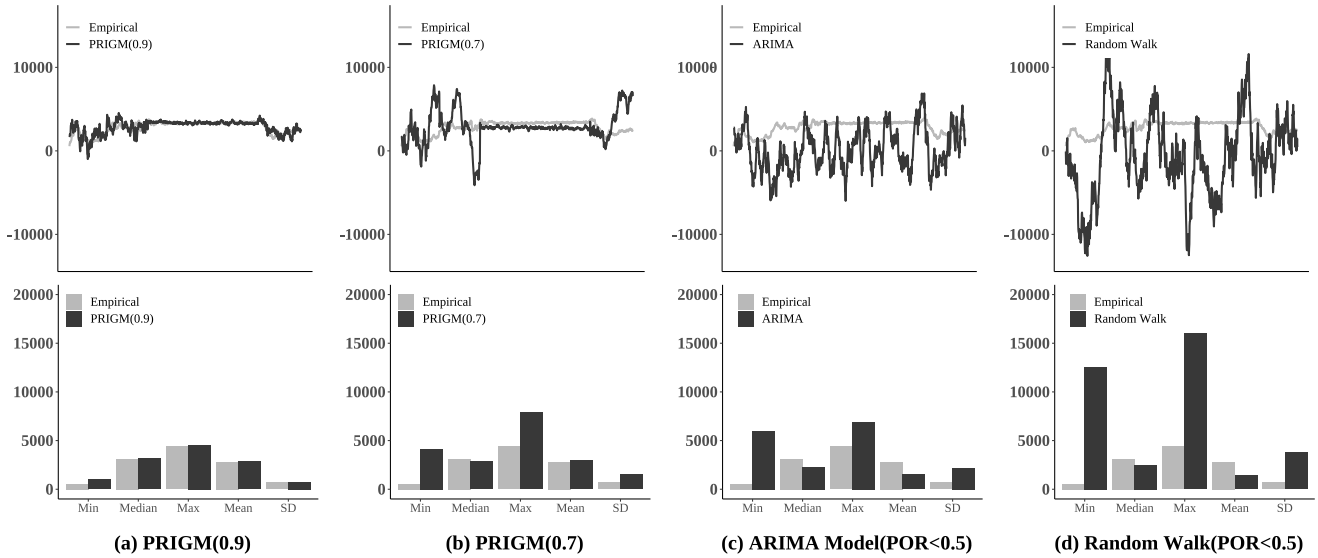
Let the Fourier spectra of ES_i and SS_i obtained by using Eq. (1) be FSE_i and FSS_i, respectively. $\mathrm{PoR}(\mathrm{seg}, i)$ for each $\mathrm{segment}_i$ is defined as Eq. (2), which shows that if the spectral difference between FSE_i and FSS_i is enlarged, $\mathrm{PoR}(\mathrm{seg}, i)$ converges to 0. However, if the two signals are similar, $\mathrm{PoR}(\mathrm{seg}, i)$ converges to 1. The length of the generated segment of FSS_i may be different from the original length, but it is robust to the sampling area by Fourier transform. The PoR value indicates the upper limit of the similarity of the synthetic benchmark generated by the SSC.

4. Performance Evaluation

To evaluate the performance of the proposed model, we used a HAR dataset provided by UCI Repository [1]. We evaluated the performance of the synthetic-benchmark generation of the proposed technique using acceleration data representing the characteristics of high variability and strong irregularity. The performance evaluation of the synthetic benchmark was determined based on the statistical significance of the time-series characteristics of the target empirical data.

PRIGM proposes automated synthetic benchmark generation techniques that can utilize various data types. Therefore, the comparison between other techniques is unfair in terms of the overhead and accuracy of data generation for their target restrictions and the requirement of pre-tuning process. In addition, PoR, a new indicator proposed in this paper, reflects the statistical similarity and time-series characteristics absolutely. For this reason, the discussion and comparison of our models are enough to show the effectiveness of the proposed method. The generation model was verified through time-series visualization and the distribution characteristics of individual values using descriptive statistics. Additionally, we verified the trend of the generated data according to the PoR value.

Figure 3 presents the generated synthetic benchmarks as time-series data and their descriptive statistics. Figure 3 (a) summarizes the characteristics of the synthetic benchmark when the proposed model has a PoR of 0.9. The generated synthetic data (top) are visualized in almost the same form as the empirical data so that time-series similarity can be confirmed. Additionally, each element of the



(a) PRIGM(0.9) (b) PRIGM(0.7) (c) ARIMA Model(POR<0.5) (d) Random Walk(POR<0.5)

Fig. 3 Visualization of time-series data (top) and the comparison of descriptive statistics (bottom). The x-axis represents the progress of time in a sampling section, and the y-axis represents an acceleration value measured.

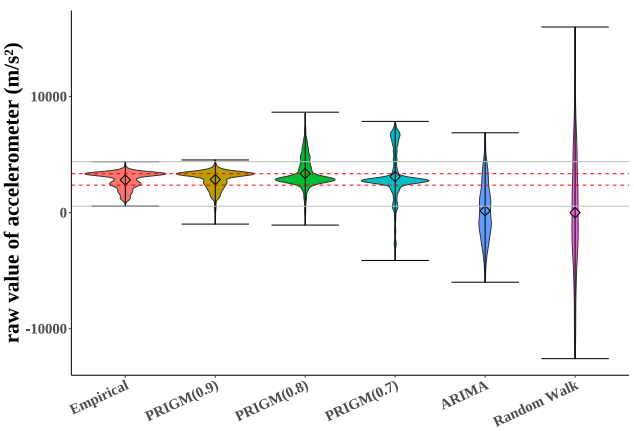


Fig. 4 Distribution of raw synthetic data generation.

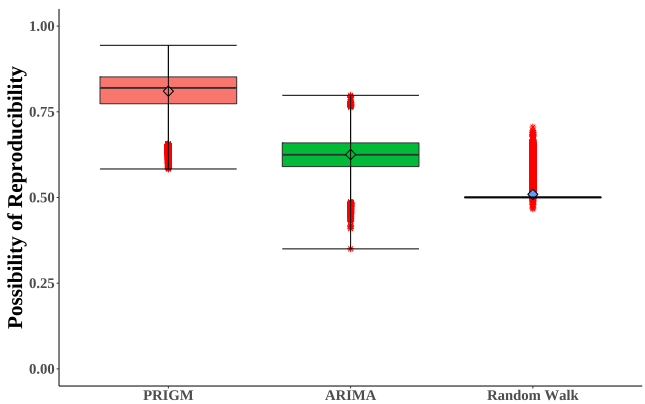


Fig. 5 Distribution of calculated PoR from 100,000 times.

descriptive statistics (bottom) shows a difference of 17% (Min), 0.8% (Median), 2% (Max), 0.1% (Mean), and 0.2% (Standard deviation), confirming the similarity of descriptive statistics between the synthetic data and the empirical data. Figure 3 (b) shows the characteristics of the synthetic benchmark with a PoR of 0.7. Data from ARIMA and Random Walk depicted in Fig. 3 (c) and Fig. 3 (d) do not reflect any fluctuations in the value of empirical data in the entire interval, and the difference increases as the ARIMA or Random Walk process proceeds.

The generation of the synthetic benchmark proceeds through the probability generation process, which generates different values for each process. Therefore, the validation and verification of the value distribution that appear in the iterative and optimized generation processes at a certain moment can be used as an indicator of the model stability. Figure 4 shows the characteristics of the data obtained through 100,000 synthetic-benchmark generations for an arbitrary acceleration dataset. As shown in Fig. 4, the empirical data has a fixed distribution, and the scope of values changes for each generation owing to the probabilistic process of the proposed method and the comparison method. In *PRIGM*, as the PoR value of the proposed technique increases, the distribution model and quantile index of the values become similar to those of the empirical data. The two red dotted lines represent the first and third quantiles of the empirical data, and the gray lines represent the minimum and maximum values. In contrast, in ARIMA and Random Walk, which are random probability processes, the distribution model of the overall value appears similar to the normal distribution, which is significantly different from the original data. We confirm that a high PoR value can generate synthetic data statistically similar to the empirical data.

Figure 5 shows the distribution of the PoR values according to *PRIGM*, ARIMA, and Random Walk. In 100,000 generation processes, the proposed PoR scheme shows a distribution of PoR values with a maximum of 0.95 and min-

imum of 0.63 for *PRIGM*. From the 25th percentile to the 75th percentile, the value is generally at the top side, and mild outliers are even in the main value distribution area of ARIMA and therefore we can expect that *PRIGM* will generate highly reproducible synthetic benchmarks. ARIMA shows a distribution of PoR values, with a maximum of 0.80 and minimum of 0.38. Mild outliers, which are sufficient to be classified as errors, appear in both the upper and lower sides. In addition, the maximum expected performance of ARIMA does not reach the main value distribution area of *PRIGM*. Random Walk is approximately 0.49, and the distribution of the PoR value of the synthetic benchmark appears narrow and invariant. This is attributed to the characteristics of the normal distribution, and most values other than 0.5 are treated as outliers. Therefore, the data generated with a random probability do not sufficiently reflect the characteristics of the distribution of values for data with time-series trends.

5. Conclusion

In this letter, we propose *PRIGM*, a synthetic-benchmark generation model that can be used for general-purpose time-series data. *PRIGM* divides sensor data based on singular points whose time-series characteristics differ, and it performs ARIMA operations for each division to attenuate errors in the autoregression analysis. To overcome the limitations of existing models for determining the similarity between synthetic and empirical benchmarks, *PRIGM* employs a new evaluation criterion called PoR, which assumes each divided section as the sum of frequencies and examines the similarity by considering the difference between a value obtained through Fourier transform and its adjacent value. The synthetic benchmark generated by *PRIGM* in the experimental evaluation mainly shows a PoR value of 0.8–0.9, and the descriptive statistics show a difference within 4.9% on average compared to the descriptive statistics of the empirical data. Therefore, *PRIGM* can generate synthetic benchmarks that statistically preserve the time-series characteristics of the empirical data.

References

[1] D. Dua and C. Graff, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], School of Information and Computer Science, University of California, Irvine, CA, 2019.

[2] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," Journal of Multiple-Valued Logic and Soft Computing vol.17, no.2-3, pp.255–287, 2011.

[3] M. Arlitt, M. Marwah, G. Bellala, A. Shah, J. Healey, and B. Vandiver, "IoTAbench: An internet of things analytics benchmark." Proc. 6th ACM/SPEC International Conference on Performance Engineering, pp.133–144. 2015.

[4] X. Liu, N. Iftikhar, H. Huo, R. Li, and P.S. Nielsen, "Two approaches for synthesizing scalable residential energy consumption data," Future Generation Computer Systems, vol.95, pp.586–600, 2019.