Evaluation of synthetically generated traces towards a data-centre digital twin

Alejandro Fernández-Montes
Damián Fernández-Cerero,
Department of Computer
Languages and Systems
University of Seville
Av. Reina Mercedes s/n, 41012
Seville, Spain
{afdez,damiancerero}@us.es

Agnieszka Jakóbik,
Cracow University of
Technology
Warszawska st 24, 31-155
Cracow, Poland
akrok@pk.edu.pl

Belén Bermejo,
Carlos Juiz
Department of Computer
Science
University of the Balearic
Islands
07122 Palma, Spain
{cjuiz,belen.bermejo}@uib.es

KEYWORDS

time-series, synthetic data generation, data centre traces, generative adversarial networks, data augmentation.

ABSTRACT

Several approaches exist to generate synthetic data centre traces for various purposes: from augmenting operational traces for data centre simulators and digital twins to forecasting incoming workload to improve data centre behaviour. The evaluation of the quality of synthetically generated multivariate time-series datasets, such as those related to data-centre traces, is not a trivial task, since complex patterns and correlation between variables may be present.

This paper proposes a new multivariate time-series evaluation framework that computes a set of metrics and figures that can be used to measure the quality of synthetically generated data-centre traces. We then employ the proposed tool to compare two synthetic data centre traces with the original trace and assess their quality. These synthetic traces have been generated by means of Generative Adversarial Networks (GAN). In this work, we employ TimeGAN, a GAN model focused on the generation of multivariate time series traces.

We finally show how the proposed framework provides us with a set of metrics consistent with the observable behaviour and numerical insights on the quality of the generated data centre traces, which are hard to acquire otherwise.

I. Introduction

Some data centre trace datasets were published by various data centre operators, such as Alibaba [1] [6], Microsoft Azure [2] [4], and Google [8]. These datasets are useful for a better understanding of the operation and behaviour of real hyperscale data centres.

Data centre traces are considered multivariate time series datasets, as they include a series of time-defined events. These events are typically divided into at least two datasets:

- Dataset of job / task events, which includes information about arrival time, status change, and deployment information.
- Machine usage dataset, composed of periodic monitoring events that summarise the utilisation of every machine at a given time. The monitored parameters usually include CPU, memory, disk, and network usage.

The research community has been using these traces for various purposes, including the simulation of data centre operation to enhance several aspects, especially resource management and job scheduling. It is worth noting that job scheduling is critical for data-centre operating performance. In addition, some simulators can also apply various energy efficiency policies to reduce the energy consumption of the data centre.

Real data-centre traces have been proven to be crucial for the evaluation of such policies in realistic industry scenarios, but such traces fall short for many purposes, including machine learning models, which need very large datasets. Data augmentation is a technique that uses algorithms to artificially increase the size of a dataset by generating modified versions of existing data points. This can be useful in machine learning tasks where the amount of data available is limited, as it helps prevent overfitting and generalisation.

Note that when data augmentation techniques are applied to time series data, preserving the temporal relationship and patterns in the data set is important. The selection of the appropriate data-augmentation techniques, as well as ensuring that the resulting augmented data still represent the original data, can help optimise the performance of machine learning models on tasks related to problems where time series are present.

To ensure that the resulting data still represent the original data and preserve both the relationships between the properties and the time-related patterns, in this paper, we propose a multivariate time-series evaluation framework that computes a set of metrics and figures that can be used to measure the quality of synthetically generated data-centre traces. We then employ the proposed tool to compare two synthetic data centre traces with the original trace and assess their quality. These synthetic traces have been generated by

Communications of the ECMS, Volume 37, Issue 1,
Proceedings, ©ECMS Enrico Vicario, Romeo Bandinelli,
Virginia Fani, Michele Mastroianni (Editors) 2023
ISBN: 978-3-937436-80-7/978-3-937436-79-1 (CD) ISSN 2522-2414

means of Generative Adversarial Networks (GAN). In this work, we employ TimeGAN [10], a GAN model focused on the generation of multivariate time series traces, trained with the Alibaba cluster trace dataset [1].

This article is organised as follows. Section II of this paper covers the metrics and measurements specifically orientated to the evaluation of time series. In Section III we present some figures that support visual analysis for the comparison of multivariate time series. This is followed by Section IV, which includes: a) dataset description; b) a summary of Generative Adversarial Networks and TimeGAN; c) and which experiments were performed, including their parameterisation. Finally, in Section V we present and discuss the results for the use case, and conclusions are drawn in Section VI

II. Metrics and measurements for the evaluation of multivariate synthetic time series

In this section, we present the most popular metrics and measurements found in the literature that are used to evaluate time-series datasets.

Kullback–Leibler divergence

The Kullback-Leibler divergence (also known as KL divergence or relative entropy) is a measure of the difference between two probability distributions. Often used in machine learning and statistics to compare the similarity of two distributions or to compare the model fit to datasets.

The KL divergence can be a useful tool for evaluating the similarity of two time series. However, it is important to remember that it is sensitive to the specific probability distributions that are used to represent them.

Jensen–Shannon divergence

The Jensen-Shannon divergence (JS) is a measure of similarity between two probability distributions. It is a symmetric version of the Kullback-Leibler divergence (KL divergence) and is defined as the average of KL divergences between the distribution a and the distribution b, and between the distribution b and the distribution a.

The JS divergence is a useful tool to compare the similarity of two time series. As it is symmetric, the order of the two distributions under evaluation does not have any influence on the results.

Like the KL divergence, the JS divergence is sensitive to the specific probability distributions that are used to represent time series.

Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (KS) test is a statistical test used to compare cumulative distribution functions (CDF) between two samples. It is often used to test whether two samples come from the same distribution or to compare the fit of a theoretical distribution to a sample.

In the context of time series analysis, the KS test can be used to compare the similarity of two time series. Both time series must be transformed into a set of feature vectors that represent the distributions. Then, we computed the KS test over the extracted featured vectors of each time series to compute the maximum difference between the CDFs of the two samples.

The KS test is sensitive to specific feature functions that are used to represent time series.

In summary, the KL divergence, the KS test, and the JS divergence can be used to compare the similarity of two time series, but they are sensitive to the specific methods used to represent the time series as probability distributions or feature vectors. Choosing the appropriate methods to represent the time series can be important for an accurate comparison of their similarity.

Maximum Mean Discrepancy

Maximum mean discrepancy (MMD) is a kernel-based statistical test used to determine whether two given distributions are the same, which is proposed in [5].

In the context of time series analysis, MMD can be used to compare the similarity of two time series. The MMD is calculated as the maximum difference between the mean of the feature vectors in one distribution and the mean of the feature vectors in the other distribution, taken over all possible feature functions.

Like the KS test, MMD is sensitive to the specific feature functions that are used to represent the time series.

Dynamic Time Warping

Dynamic Time Warping (DTW) is an algorithm that is used to compare time series by aligning them in a way that minimises the distance between them. It is often used in speech recognition and pattern recognition tasks because it enables the comparison of time series that may have different lengths or that may be shifted in time.

DTW is not a metric in the strict mathematical sense of the term. A metric is a function that satisfies certain properties, such as being non-negative, symmetric, and satisfying the triangle inequality. DTW does not satisfy these properties, so it is not a metric in the traditional sense. However, it is often referred to as a ”distance measure” or a ”similarity measure” because it quantifies the distance between two time series.

In this work, we use the proposal for multidimensional DTW of [7]

Difference of covariances

The difference of covariances can be used to assess how different the relationships between variables are between two time series. It is computed as the average of the row-wise Frobenius norm for the covariance difference matrix.

Difference of correlations

In the same way, the difference in Pearson's correlation can be used to assess how similar the relationships between variables are between two time series. It is computed as the average of the row-wise Frobenius norm for the Pearson correlation difference matrix.

Difference of histograms

Finally, the difference of histograms can determine how far the ranges of values are between two time series.

Synthesis

Whether each row is new or matches an original row of the real data is checked and calculated from 0.0 to 1.0 (all rows are new) using [3].

Coverage

Whether the synthetic data cover the full range of values of the real data is checked and calculated from 0.0 to 1.0 (full coverage) [3].

Boundaries

Whether the synthetic data respect the boundaries of the real data is checked and calculated from 0.0 to 1.0 (all data respect the boundaries) [3].

III. **Figures**

We also present other visual measurements that can be helpful in comparing and visually representing the similarity between two time series. Notice that the figures presented are some examples that were generated from some experiments to show the usefulness of these techniques to make a visual comparison between multivariate time series.

T-distributed Stochastic Neighbour Embedding (t-SNE)

T-SNE is a tool for visualising high-dimensional data sets in a 2D or 3D graphical representation proposed by [9], allowing the creation of a single map that reveals the structure of the data at many different scales. T-SNE is a non-linear technique that aims to preserve the local structure of the data.

An example of a generated t-SNE representation is shown in 1.

Figure 1 shows noticeable differences between the synthetic and original data.

Principal component analysis (PCA)

PCA is a linear dimensionality reduction technique that aims to find the principal components of a data set by computing the linear combinations of the original characteristics that explain the most variance in the data.

Therefore, PCA is better suited for datasets with linear structure, whereas t-SNE is better suited for datasets with nonlinear structure.

An example of a generated PCA representation is shown in 2.

Figure 2 shows noticeable differences between the synthetic and original data.

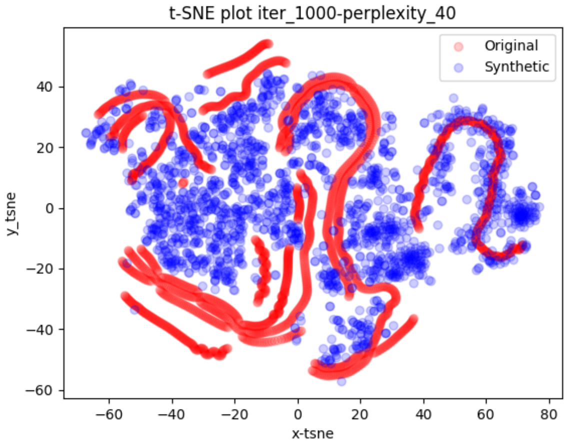


Fig. 1: Example t-SNE between synthetic and real data

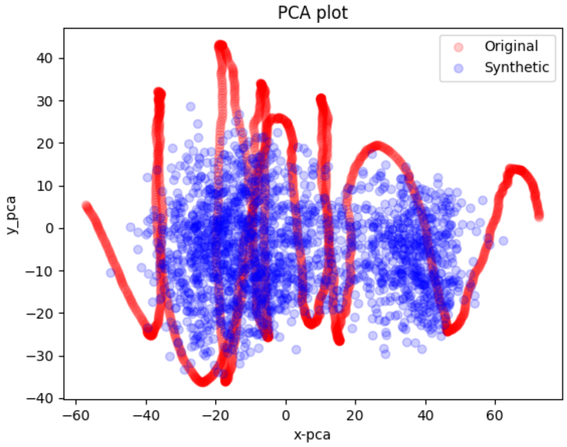


Fig. 2: Example PCA computed by the framework

Dynamic Time Warping path

In addition to the numerical similarity measure, the graphical representation of the DTW path of each column can be useful to better analyse the similarities or differences between the original and synthetic columns. Notice that there is no multivariate representation of DTW paths, only column representations, as shown in Figure 3.

Figure 3 shows that the patterns found in the synthetic data (lower half) are quite similar to those presented in the original data (upper half).

Time Series plotting

We can use the proposed framework to directly plot the ordinary graphical representation of the time series in a 2D figure with the time represented on the x axis and the data values on the y-axis for a) the complete multivariate time series; and b) a per column plot.

Each generated figure plots both the original and the synthetically generated data to easily obtain key insights into the similarities or differences between them.

Figure 4 shows an example of a graphical represen-

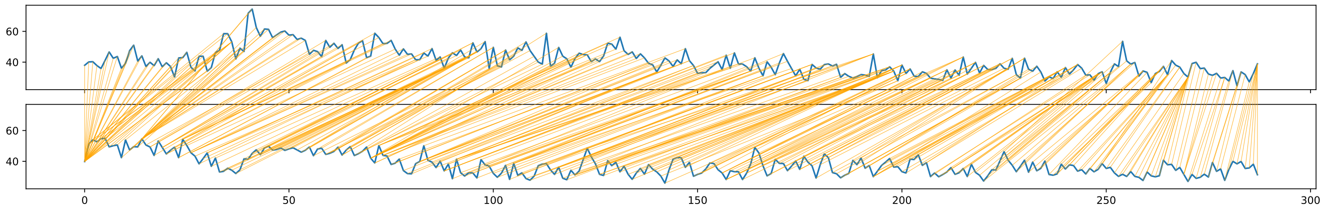


Fig. 3: Example DTW path between synthetic and real data

tation for the comparison of various columns between the synthetic and original data (dotted). To this end, the original and synthetic data are overlapped in the same plot.

Figure 4 helps show that there is a very good fit of $net_i nand$

Finally, we can compute and plot the differences between the values of each column grouped by periods of time. For instance, the differences between the cpu usage every 5 minutes or every 30 minutes. These deltas can be used as a means of comparison between synthetic dataset samples and real data samples.

Figure 5 shows an example of the graphical representation of deltas of a synthetically generated sample and five real data samples.

IV. Experiment design

A. Dataset

In this article we use the Alibaba 2018 machine usage trace, which contains records produced by a server monitoring system of a production data centre for an operation period of 8 days. Each monitoring record contains the following information: a) timestamp; b) machine id; c) percentage of CPU usage for that machine; d) percentage of memory usage for that machine; e) percentage of input network flow; f) percentage of output network flow; and g) percentage of disk usage for that machine. The monitoring system generates a record every 10 seconds. As a pre-processing stage, we grouped and averaged all the values by machine id. This dataset can be obtained in [1].

B. Generative Adversarial Networks

Generative adversarial networks (GAN) can be used to augment the data by training a generator that creates new data similar to the original. These samples can be added to the original data to increase the size of the data set, so that we can reduce overfitting and increase the diversity of the original data.

TimeGAN

In this article, we use TimeGAN [10], a GAN model focused on the generation of time series datasets. TimeGAN combines the unsupervised paradigm with the control afforded by supervised training. The novelties of TimeGAN include the proposal of an embedded space in which learning, generation, and discrimination are performed. Thus, the original data is transformed to the embedded space, and the generated data is transformed (recovered) to the original space.

To this end, TimeGAN performs three steps: 1. Embedder training, 2. Supervised training, 3. Joint GAN training.

It is important to note that the effectiveness of GAN models, such as TimeGAN, depends largely on the parameterisation of the model. As in many unsupervised generative models, empirically evaluating the impact of such parameters on complex multivariate time series is not trivial. This use case is a perfect fit for the proposed evaluation framework.

TABLE I: Parameterisation of the two TimeGAN models employed.

Model	**RNN**	**L**	**H**
Good	GRU	3	8
Bad	LSTM	4	16

C. Parameterisation

In this work, we propose a multivariate time series evaluation framework as a tool to determine the goodness of synthetically generated datasets. We use TimeGAN trained with the Alibaba 2018 machine usage dataset to generate the synthetic datasets. In order to illustrate how the proposed evaluation framework helps in the analysis and comparison of the ability of unsupervised ML models to produce realistic results, in this Section we will compare two TimeGAN models as an example. The aim is to check whether the proposed evaluation framework can provide us with significant information to discriminate between good and bad results. The hyperparameters of TimeGAN include:

- *Batch size, fixed as 100 in both models;*
- *Training iterations, fixed as 1500 in both models;*
- *Sequence length, fixed as 8640 for both models. This means that we will produce synthetic traces that represent one day of operating time;*
- *Type of RNN model employed RNN;*
- *Number of layers L;*
- *Hidden dimensions H;*

The Adam optimiser with a learning rate of 0.001 is employed for both models. The parameterisation of each model is shown in Table I

[1] https://github.com/alejandrofdez-us/DataCenter-Traces-Datasets

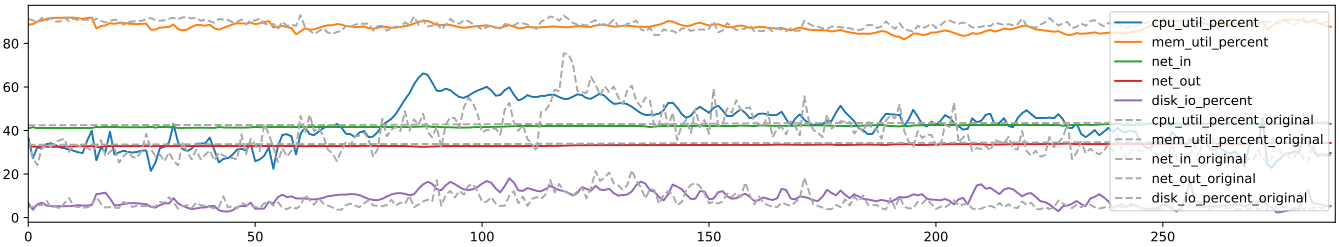


Fig. 4: A time series plot that overlaps synthetic data with real data.

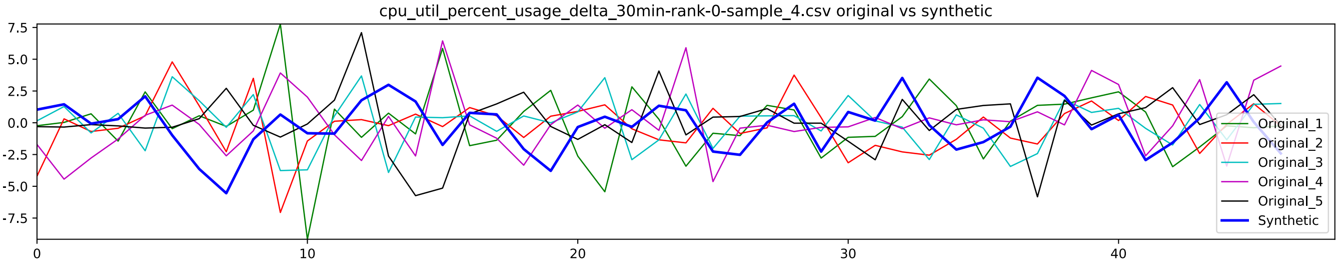


Fig. 5: Figure that overlaps the 30-minute period deltas for synthetic data with five different real data samples to check their similarity.

V. Results

We used the proposed framework to compare the datasets produced by the models presented in Section IV-C with the original Alibaba dataset and gather quality metrics.

The proposed tool provides the following set of results for the most significant metrics presented in Section II:

a) one set of results for the multivariate analysis of the time-series dataset; and b) one set of results for the univariate analysis for each column in the dataset.

For the seek of clarity, in this section we present an analysis of the results provided for the multivariate analysis and only two of the columns of the dataset, even though the results are available for all of the columns. The selected results of this analysis are presented in Table II.

Regarding the multivariate analysis, the framework provides insights about which model performed better: the maximum mean discrepancy, Jensen-Shannon, Kullback-Leiber and the Kolmogorov-Smirnov metrics are much lower for the Good model.

According to the single-variable results for memory usage, the results are clear: the Good model strongly outperforms the Bad model. We can check that the results are consistent with the behaviour presented in Figure 6.

However, it can be noticed that other metrics, such as the difference of covariances and Pearson correlation, as well as the Dynamic Time Warping do not show good results. Due to this, a key insight can be obtained: even though the Good model outperforms the Bad model, the patterns of the traces produced by the Good model don't fit well to the original dataset, not in a single-variable analysis, but as a whole, so it may indicate the model is not able to reproduce correlation between the different variables in the trace.

Short- and long-term behaviour can also be analysed using the period deltas provided by the evaluation framework. Figure 7 shows the short-term (five minutes) deltas for memory usage, as well as the long-term deltas for disk usage. In this figure, it becomes evident that the results provided by the metrics presented in Table II are consistent with the patterns produced for both good and bad models.

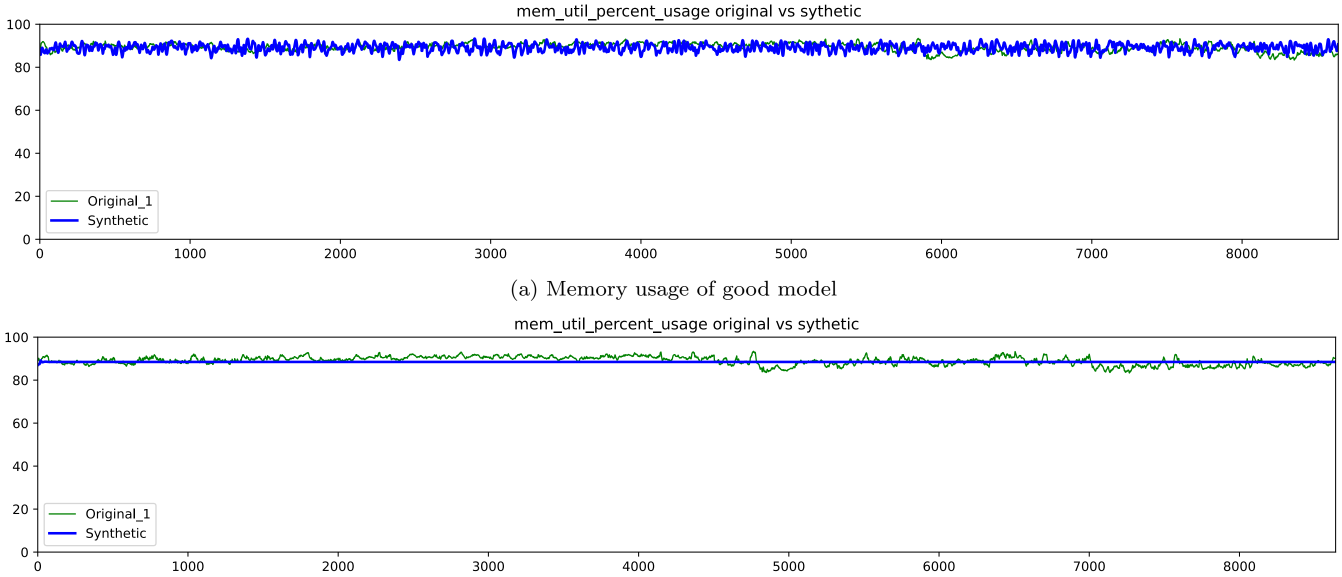
TABLE II: Results of the proposed time series evaluation framework for the comparison of two synthetic traces with the original trace.

Trace	**MMD**	**DTW**	**JS**	**KL**	**KS**	**CC**	**CP**	**HI**
Multivariate analysis results								
Good	0.06	549.27	8.29	10.47	0.23	7866	1.16	1119
Bad	0.97	392.91	nan	15.32	0.64	7524	0.79	8139
Single variable analysis results: Memory usage								
Good	0.01	237.92	0.08	0.05	0.09	N/A	N/A	912
Bad	0.68	183.82	6.04	7.53	0.63	N/A	N/A	7432
Single variable analysis results: Disk usage								
Good	0.02	463.85	0.45	0.13	0.17	N/A	N/A	1832
Bad	0.90	342.73	5.84	6.59	0.63	N/A	N/A	8532

VI. Conclusions

In this paper, we presented a framework for the evaluation of multivariate time series which allowed us to compare synthetically generated data centre traces with the original traces and get valuable insights about the behaviour of unsupervised ML models, such as TimeGAN.

The analysis of the behaviour of complex multivariate time series and the related patterns is not trivial, and there is no single metric that can show the fitness of the generated traces.



(a) Memory usage of good model

(b) Memory usage of bad model

Fig. 6: Comparison of memory usage between good and bad models.

To overcome such a limitation, the evaluation framework provides a set of metrics that represent different facets of the behaviour of the time series patterns. In most of the cases, we propose that the tuple composed of Dynamic Time Warping and Jensen-Shannon should be used as the main metrics for the evaluation of patterns and distances between time-series traces.

Acknowledgments

Project PID2021-122208OB-I00, PROYEXCEL_00286 and TED2021-132695B-I00 project, funded by MCIN / AEI / 10.13039 / 501100011033, by Andalusian Regional Government, and by the European Union - NextGenerationEU.

References

[1] Alibaba production cluster data v2018. `https://github.com/alibaba/clusterdata/tree/v2018`, 2018. Version 2018.

[2] Azure public dataset v2. `https://github.com/Azure/AzurePublicDataset/blob/master/AzurePublicDatasetV2.md`, 2019. Version 2.

[3] Synthetic data metrics. `https://docs.sdv.dev/sdmetrics/`, 10 2022. Version 0.8.0.

[4] Eli Cortez, Anand Bonde, Alexandre Muzio, Mark Russinovich, Marcus Fontoura, and Ricardo Bianchini. Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pages 153–167, 2017.

[5] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[6] Jing Guo, Zihao Chang, Sa Wang, Haiyang Ding, Yihui Feng, Liang Mao, and Yungang Bao. Who limits the resource efficiency of my datacenter: An analysis of alibaba datacenter traces. In *Proceedings of the International Symposium on Quality of Service*, IWQoS '19, New York, NY, USA, 2019. Association for Computing Machinery.

[7] Wannes Meert, Kilian Hendrickx, Toon Van Craenendonck, and Pieter Robberechts. Dtaidistance. `https://doi.org/10.5281/zenodo.7158824`, August 2020.

[8] Muhammad Tirmazi, Adam Barker, Nan Deng, Md E. Haque, Zhijing Gene Qin, Steven Hand, Mor Harchol-Balter, and John Wilkes. Borg: the Next Generation. In *Proceedings of the Fifteenth European Conference on Computer Systems (EuroSys'20)*, Heraklion, Greece, 2020. ACM.

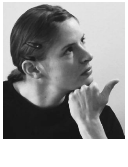
[9] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[10] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

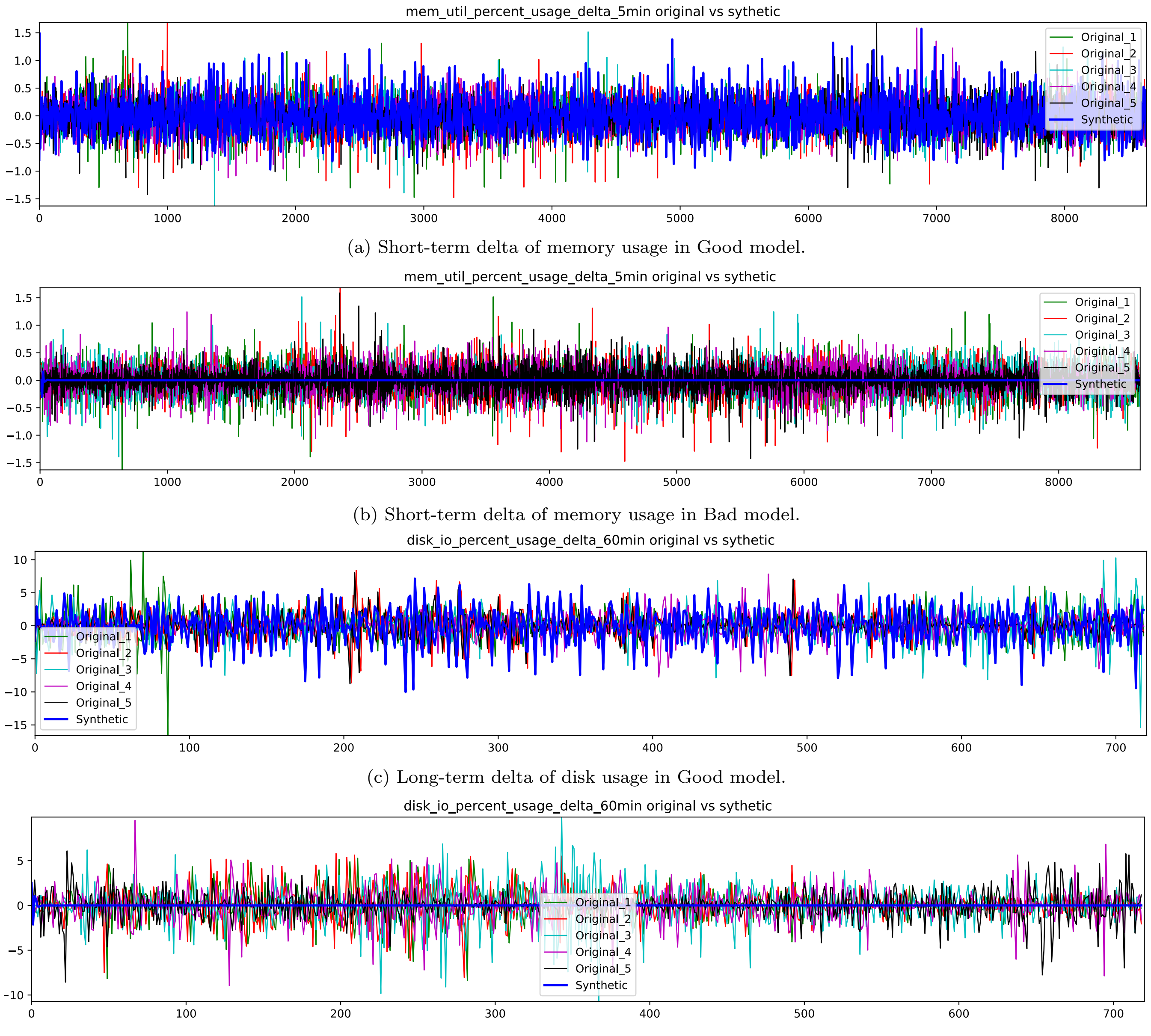
AUTHOR BIOGRAPHIES

ALEJANDRO FERNÁNDEZ-MONTES *received the B.E. degree, M. Tech. and International Ph.D. degrees in Software Engineering from the University of Sevilla, Spain. In 2006, he joined the Department of Computer Languages and Systems, University of Sevilla, and in 2013 became Assistant Professor. His research interests include energy efficiency in distributed computing, applying prediction models to balance load, and applying on-off policies to data centres.*

DAMIÁN FERNÁNDEZ-CERERO *received the B.E. degree and the M.Tech. degrees in Software Engineering from the University of Sevilla. In 2014, he joined the Department of Computer Languages and Systems, University of Seville, as a Ph.D. Student. Currently, he teaches and conducts research at the University of Sevilla. He has worked on several research projects supported by the Spanish government and the European Union. His research interests include energy efficiency and resource scheduling in data centres.*



AGNIESZKA JAKÓBIK *(KROK) received her M.Sc. in the field of stochastic processes at Jagiellonian*



(a) Short-term delta of memory usage in Good model.

(b) Short-term delta of memory usage in Bad model.

(c) Long-term delta of disk usage in Good model.

(d) Long-term delta of disk usage in Bad model.

Fig. 7: Comparison of short-term and long-term memory and disk usage between good and bad models.

University, Cracow, Poland, and her Ph.D. degree in the field of neural networks at Tadeusz Kosciuszko Cracow University of Technology, Poland, in 2003 and 2007. She is an Assistant Professor.

BELÉN BERMEJO *obtained her Ph.D. in 2020 from the University of the Balearic Islands. Since 2016 she has been an assistant lecturer in the computer science department of the UIB and a member of the ACSIC group. She is chair of the WiE section of IEEE Spain. Her research is mainly focused on improving energy efficiency and performance in data centres and servers, especially virtualised data centres.*

CARLOS JUIZ *received the PhD degree in Informatics from the University of the Balearic Islands, Spain. From 1990 he was Systems Analyst at Xerox, leaving this position as Senior Analyst in 1999. He was a visiting researcher at the Department of Computer Science and Business Informatics, University of Vienna, in 2003, and a Visiting Associate Professor at Biomedical Informatics Research, in 2011, at Stanford University. Carlos Juiz' research interests focus on performance engineering, Green IT, and IT governance. He has been involved in several regional, national, European, and international research projects. Carlos Juiz is a senior member of the IEEE and the ACM.*