Method for Artificial KPI Generation With Realistic Time-Dependent Behaviour

Hao Qiang Luo-Chen, Carlos S. Alvarez-Merino, Emil J. Khatib, *Member, IEEE*, and Raquel Barco

***Abstract*—Machine Learning (ML) is the dominating solution for the implementation of Self-Organizing Networks (SON), which automate mobile network management. However, the data scarcity derived from the reluctance of operators complicates the necessary training phase ML algorithms. In this letter a method to generate artificial Key Performance Indicators (KPIs) time series is proposed considering their time-dependent behaviour. The data is modelled and categorised according to the time of the day and the data models are adapted with statistical copulas to create samples which present interrelation among different KPIs. Finally, results obtained from a real mobile network are presented.**

***Index Terms*—KPI modelling, time-awareness, statistical relationship, copula function, correlation.**

I. Introduction

SELF-ORGANIZING Networks (SON) [1] were conceived to fulfil the task of automatic management of mobile networks. The complexity and volume of data and devices make manual management economically unfeasible. SON consists of three main functionalities: Self-Configuration, Self-Optimization and Self-Healing (SH). This letter is centred on SH [2], that automates network troubleshooting. The current general tendency on SH systems, especially in troubleshooting, is the use of Artificial Intelligence (AI) algorithms that are trained using Machine Learning (ML) processes [1]. During the training phase, ML uses previously collected data to adapt the parameters of the AI algorithm to the application where it will be used. Once adapted, the AI algorithm can be exploited to perform SH functions such as problem detection or troubleshooting. In case of mobile networks, obtaining these data is difficult because of the reluctance among operators to share data with the research community and, especially, due to the lack of a structured way to save solved troubleshooting cases. As a consequence, the data available for researching and validating novel SH schemes is scarce and there is a need of mechanisms to make a better use of the few existing datasets.

In order to overcome this data deficit in ML training datasets, in the literature the use of testbenches [3], simulations [4] and modelling [2], [5], [6] are proposed. Firstly, in the testbenches [3], commercial equipment originate the data, but they are only partially realistic as a result of non-natural activity and traffic from the users. In case of real user monitoring, the cons derive from the high number of users and the extensive time-span requirements, which produce datasets with a very high dimensionality, apart from privacy concerns. Secondly, simulations [4] are versatile solutions, being able to cover different scenarios and adapt to their specifications. Nevertheless, the cumbersome calculations can only recreate up to a certain degree of realism, constrained by the assumptions in the simulation mechanism. Finally, the modelling approaches are significantly more efficient and keep the realism of the generated samples. Modelling can be done analytically, such as in [5] and [6], where queuing theory is used to model the access of users to the network and to estimate other aspects of network behaviour. In [2], a linear combination method is proposed to quantify the blocking rate. In [6], a method is presented to estimate the Quality of Service, comprehending the system performance, the failure and repairing time. Even so, both cases are restricted to very specific problems, and do not consider general degradations. In [2], a more general network problem modelling method based on statistical distributions is proposed. The approach is based on fitting different Probability Distribution Functions (PDFs) to datasets that are conditioned to the occurrence of a specific problem and scoring them with a Kolmogorov-Smirnoff test. However, the main shortcomings of [2] are the non-existence of time-dependent behaviour modelling nor the underlying correlation of KPIs.

The training data scarcity can also be tackled from the ML system setup perspective. For example, artificial samples can be generated via a GAN algorithm [7], with which the complex time series dynamics can be learned and recreated but at the expense of a huge number of training samples. Apart from that, different ML strategies are designed to perform better under few different situations. In this regard, the application of few-shot learning [8] has been proposed for handling the learning with LTE data with few samples for each data class. Likewise, incremental learning [9] can learn continuously and adapt to non-trained samples, starting from sparse data. Furthermore, transfer learning ideas [10] propose learning from certain datasets and extrapolating the trained system to new ones. These algorithms are highly promising but are usually restricted to certain ML architectures or imply the

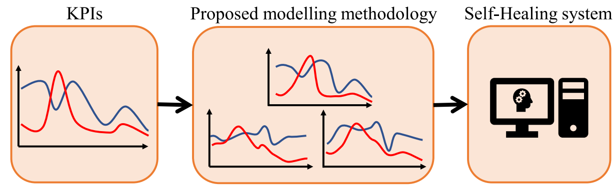

Fig. 1. Adaptation of KPIs to samples-demanding SH system.

Manuscript received April 5, 2021; revised May 26, 2021; accepted July 1, 2021. Date of publication July 7, 2021; date of current version September 10, 2021. This work was partially funded by Junta de Andalucía and ERDF (Programa Operativo FEDER Andalucía 2014-2020) within projects PENTA (PY18-4647), DAMA5G (UMA-CEIATECH-11) and IDADE-5G (UMA18-FEDERJA-201). The associate editor coordinating the review of this letter and approving it for publication was H. S. Ghadikolaei. *(Corresponding author: Hao Qiang Luo-Chen.)*

The authors are with Instituto Universitario de Investigación en Telecomunicación (TELMA), Universidad de Málaga, CEI Andalucía TECH E.T.S.I. Telecomunicación, 29010 Málaga, Spain (e-mail: hao@ic.uma.es; cam@ic.uma.es; emil@uma.es; rbm@ic.uma.es).

Digital Object Identifier 10.1109/LCOMM.2021.3095372

implementations of changes, hindering the direct application over existing SH solutions.

In this letter a modelling method which overcomes all these limitations is presented. Initially, the method classifies the data into categories based on the time of the day and, then, it fits models to each category. Afterwards, the method applies a statistical copula [11] to modify the models into new ones maintaining time-dependent relationships among different variables. Copulas are generalisations of distributions which allow techniques such as Pearson or Kendall to measure and establish the statistical dependencies among them. Finally, the method is used for modelling a small dataset and generating a larger synthetic dataset that can be used in ML (with any setup and without modifying it) to train AI algorithms for SH; closing the gap between available and required training samples, as summarised in Fig. 1. The net contribution of this method is that it helps to overcome the problem of lack of training data present in many ML based solutions, which ultimately hinders the development of SON systems, maintaining their implementations. The generated data retain time-dependent effects and are based on real observed data.

The rest of this letter is organised as follows. Section II presents the modelling problem. Section III describes the proposed methodology, starting with a time modelling of the KPIs and describing a method to maintain certain statistical relationship among the KPIs. In Section IV some results obtained from real network data are presented. Finally, the conclusion and some future lines are discussed in Section V.

II. Formulation

Troubleshooting is a challenging task due to the vast diversity of factors such as extensive interactions of elements, network configurations or surrounding conditions and the dynamic nature of its behaviour. Under these circumstances, the network performance and its status are inferred from indirect parameters allowing operations such as troubleshooting or SH. Historically, engineer experts provide diagnosis through manual observation and analysis of those parameters [12]. The growing use of ML and AI techniques has eased this exhaustive task. These techniques used for SON demand a previous training phase which requires an elevated number of samples. In [2], a modelling method based on statistical distributions is proposed to solve the shortness of available training data. Nonetheless, it does not consider the time-dependent characteristics and assumes that different variables are independent from each other, and only depend on the root cause of a problem. So, we propose a methodology which can generate synthetic data in order to enrich the database for the training stage, of ML techniques used in SON. As an improvement over the technique proposed in [2], these synthetic data consider the time dynamism and data dependencies from the original network information for each root cause, gaining realism with respect to the previous method.

The network status is represented by a set of hierarchical variables, such as counters, Key Performance Indicators (KPIs) and Key Quality Indicators (KQIs). These variables are categorised according to their aggregation level (e.g. user elements (UE), cells, areas). In this letter, KPIs are selected as the main performance indicator in the datasets. Examples of KPIs are: *number of active user*, *number of successful inter radio access technologies (IRAT) handover*, *traffic volume* or *the average throughput received for each user*. KPIs are generally reported in cycles of 1h, referred as Report Output Period (ROP); although it also may have another frequency (e.g. 15 min or 1 day). The concatenation of the same KPI for several cycles generates a time series, as represented in Fig. 2. The progression of the time series may evidence the statistical relationship among the KPIs which they are generated from. For example, a linear relation can be traced between the number of users and the traffic volume. Nevertheless, the relationships of KPIs are more complex than this direct case, making them difficult to be determined intuitively. As example of this latter case, the number of users has influence in the number of IRAT handovers; albeit in a non-linear sense, with influence of phenomenon such as the SINR of the communication channel or some handover configuration parameters.

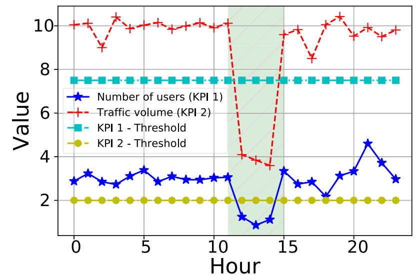

Fig. 2. Degradation in time series.

The evolution of these data also displays when a network suffers from degradation. A degradation is a temporal interval when the network service is altered and the cell performance is below the desirable. This problem is reflected in the KPIs and, subsequently, in the time series. A general method for its detection is based on defining thresholds, as limits for the time series values. When it is crossed, as observed in the hatched interval in Fig. 2, it is considered that a degradation has occurred. This work models the behaviour of KPIs in the degraded intervals, in order to generate degraded samples for the ML training phase in SH.

III. Proposed Method

The main issues for precise network modelling, as highlighted earlier, are two: time dynamism and parameter relationship. The method proposed here takes into account these issues, and only needs data from degraded KPIs as inputs. Each issue is solved separately in two stages. In the first one or *time modelling* stage, the method creates statistical distributions that reflect the KPI dynamism. Afterwards, the second or *copula application* stage modifies the earlier distributions to force statistical relationships among different KPIs. Fig. 3 illustrates a general scheme of the method.

A. Time Modelling - Consideration of KPI Dynamism

User habits are related to a specific day moment; e.g. working or sleeping hours, generating specific time-dependent

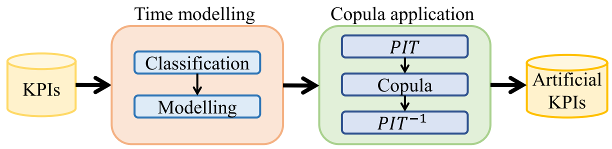

Fig. 3. Scheme of the proposed method.

patterns in the behaviour of the network. Operators deploy policies according to the traffic; for example, shutting down some base stations when the user traffic is more scarce at night hours, even relying on legacy Radio Access technologies in these hours. These time-dependent effects shape the KPI behaviour. Hence, in order to maintain time-dependent behaviour, the proposed method first classifies the input data in: weekdays or weekend and each group is sliced in hours. Altogether, there are 48 categories that organise data samples.

Thereafter, proceeding as in [2], a set of statistical distributions is fitted to the samples of each category of each KPI. The adjusted distributions vary depending on the nature of the variable: 1) *Bounded variables*: proportions that are defined between 0 and 1. The normal, Laplace, Gumbel-R and Gumbel-L distributions are fitted. 2) *Semi-bounded variables*: those defined between 0 and ∞, such as counters. The gamma, exponential and log-normal distributions are used in this case. 3) *Unbounded variables*: those that can theoretically take any value. In this case beta, arcsine and Johnson SB are used.

The goodness of fit of each distributions is tested with the Kolmogorov-Smirnov test. This test measures the difference from each fitted distribution to the actual KPI. The distribution that minimises the difference is the one which fits best with the data. Consequently, it is chosen as the model for the category.

B. Copula Application - Establishment of KPI Relationships

Generally, in order to simply the modelling process, the KPIs are assumed to be independent, in which case the joint distribution is defined as

$$F(x_1,\ldots,x_n) = \prod_{i=1}^{n} F_i(x_i); \tag{1}$$

where x_i is a random variable (i.e. a KPI), n is the number of considered KPIs and $F_i(x)$ a distribution fitted to a specific KPI. Nevertheless, KPIs are shaped by underlying factors which produce statistical dependencies among them. This section will show the use of copulas [11] as the key to include these phenomenons in the modelling.

A copula is a multivariate distribution function whose elements are traditional uniform distributions. It can create new samples of each distribution although they are modified to add a certain degree of relationship. Specifically, this function defines the joint distributions of the variables, which no longer is simply a multiplication as the independent case in Eq. (1). To overcome the limitation to uniform distributions, the Probability Integral Transformation (PIT) method may be used, allowing the transformation of any distribution into an uniform distribution. After applying the copula, the inverse transformation (PIT^{-1}) allows to return to the original distributions. Hence, the copula can relate any distribution function.

Similarly to distributions, copulas are made up of different families such as elliptics (Gaussian, t-Student, etc.) or archimedeans (Clayton, Frank, Gumbel, etc.). Each copula family confers a specific form to the samples. Additionally, in the case of elliptics, the correlation matrix Σ_ρ defines the degree of relationships among distributions. In the case of archimedeans, Σ_ρ is replaced by an equivalent value θ. Eq. (2) presents the mathematical expression for copulas.

$$F(x_1,\ldots,x_n;\Sigma_\rho) = C\{F_1(x_1),\ldots,F_n(x_n);\Sigma_\rho\}; \tag{2}$$

where C defines a specific copula function. This corresponds to a general case. Archimedean copulas of 3 or more dimensions (variables) cannot model negative relationships, while in elliptics this limitation is absence.

Eq. (3) exposes a correlation matrix that follows an unstructured form as dispersion structure. Each element $\rho_{i,j}$ represents the correlation between distributions i and j. Any correlation technique, such as Pearson or Kendall, is a potential candidate for the construction of Σ_ρ, as long as it ensures the matrix is positive-definite. The usage of different correlation techniques captures different kinds of dependency.

$$\Sigma_\rho = \begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,n-1} & \rho_{1,n} \\ \rho_{2,1} & 1 & \cdots & \rho_{1,n-2} & \rho_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{n-1,1} & \rho_{n-1,2} & \cdots & 1 & \rho_{n-1,n} \\ \rho_{n,1} & \rho_{n,2} & \cdots & \rho_{n,n-1} & 1 \end{pmatrix} \tag{3}$$

To construct related KPIs, a copula is selected from a family. KPI data from the same hour are used to calculate a correlation coefficient and, then, a matrix Σ_ρ is formed. Afterwards, copulas are constructed for each hour, allowing the generation of samples with relationship within each hour. Hence, the creation of time series is trivial, only concatenating samples of each hour from each KPI.

IV. Results

KPI data collected from a real mobile network have been used to test the proposed method. They encompass several sectors and sites with KPIs at 1h periodicity and from all possible days. In addition, the data contain normal and degraded KPI samples. Considering this, they have been analysed with a degradation detection tool. This tool extracts separately normal and degraded intervals of KPI samples, which have then been manually diagnosed to label them with the root cause. For the sake of simplicity and comparison, the rest of this section is only focused on three KPIs. The KPIs are *traffic volume* for downlink (KPI_1), the *average number of active user elements (UE)* in downlink (KPI_2) and the *number of IRAT handovers* (KPI_3). KPI_1 and KPI_2 have a linear relationship, while KPI_1 and KPI_3 are non-linear. In view of confidentiality issues, the KPIs are not directly shown except for the results calculated from them.

Table I lists the weekday models of the KPIs as examples of KPI time-aware modelling. The model contains two columns, the *Normal* status and the *Congestion* degradation. T stands for the hour, N for the number of available samples, *dist* and *params* for the distributions and their parameters, respectively.

TABLE I

MODELS FOR THE KPIS: (LEFT) TRAFFIC VOLUME, (CENTER) AVERAGE NUMBER OF ACTIVE UES AND (RIGHT) INTER RAT HANDOVER

	Normal			Congestion		
T	**N**	**dist**	**params**	**N**	**dist**	**params**
0	41	gamma	[1.54 0. 0.4]	0		
1	40	gamma	[1.35 0. 0.34]	0		
2	39	lognorm	[0.84 0. 0.21]	0		
3	40	gamma	[1.29 0. 0.15]	0		
4	41	expon	[0. 0.18]	0		
5	39	expon	[0. 0.29]	0		
6	38	gamma	[1.1 0. 0.6]	0		
7	37	gamma	[2.23 0. 0.59]	0		
8	35	gamma	[6.63 0. 0.3]	0		
9	35	gamma	[8.09 0. 0.24]	2	lognorm	[0.26 0. 1.92]
10	33	gamma	[6.01 0. 0.38]	5	gamma	[12.93 0. 0.15]
11	30	gamma	[4.27 0. 0.48]	9	gamma	[16.36 0. 0.12]
12	28	gamma	[6.66 0. 0.35]	9	gamma	[6.87 0. 0.29]
13	31	lognorm	[0.41 0. 2.08]	8	gamma	[6.53 0. 0.3]
14	35	gamma	[4.71 0. 0.47]	6	gamma	[2. 0. 0.84]
15	36	lognorm	[0.37 0. 2.17]	5	gamma	[3.51 0. 0.53]
16	35	gamma	[10.66 0. 0.2]	2	lognorm	[0.09 0. 2.68]
17	37	gamma	[9.67 0. 0.24]	2	lognorm	[0.1 0. 2.74]
18	38	gamma	[2.27 0. 0.92]	0		
19	34	gamma	[1.92 0. 0.94]	0		
20	32	lognorm	[0.33 0. 1.63]	0		
21	34	gamma	[1.12 0. 1.59]	0		
22	39	gamma	[1.54 0. 0.91]	0		
23	38	gamma	[1.36 0. 0.74]	0		

	Normal			Congestion		
T	**N**	**dist**	**params**	**N**	**dist**	**params**
0	41	gamma	[1.54 0. 0.11]	0		
1	40	gamma	[1.76 0. 0.06]	0		
2	39	lognorm	[0.74 0. 0.06]	0		
3	40	gamma	[1.6 0. 0.04]	0		
4	41	gamma	[1.08 0. 0.05]	0		
5	39	gamma	[0.85 0. 0.13]	0		
6	38	expon	[0. 0.24]	0		
7	37	gamma	[1.13 0. 0.63]	0		
8	35	lognorm	[0.96 0. 1.01]	0		
9	35	lognorm	[0.95 0. 1.18]	2	expon	[0. 4.09]
10	33	lognorm	[1.07 0. 1.43]	5	expon	[0. 5.89]
11	30	lognorm	[1.06 0. 1.48]	9	expon	[0. 4.94]
12	28	lognorm	[1.19 0. 1.87]	9	lognorm	[0.83 0. 3.45]
13	31	lognorm	[1.16 0. 1.78]	8	expon	[0. 5.59]
14	35	lognorm	[1.12 0. 1.55]	6	lognorm	[1.57 0. 1.83]
15	36	lognorm	[1.1 0. 1.58]	5	gamma	[1.27 0. 2.09]
16	35	lognorm	[1.1 0. 1.51]	2	lognorm	[0.25 0. 2.62]
17	37	lognorm	[1.13 0. 1.49]	2	lognorm	[0.41 0. 3.1]
18	38	gamma	[0.69 0. 3.73]	0		
19	34	gamma	[0.78 0. 1.9]	0		
20	32	lognorm	[0.73 0. 0.7]	0		
21	34	gamma	[0.54 0. 3.01]	0		
22	39	lognorm	[1.68 0. 0.45]	0		
23	38	gamma	[1.26 0. 0.27]	0		

	Normal			Congestion		
T	**N**	**dist**	**params**	**N**	**dist**	**params**
0	41	johnsonsb	[6.18 1.22 0. 100.]	0		
1	40	beta	[1.53 205.03 0. 100.]	0		
2	39	johnsonsb	[6.38 1.17 0. 100.]	0		
3	40	beta	[1.3 186.72 0. 100.]	0		
4	41	johnsonsb	[6.19 1.16 0. 100.]	0		
5	39	johnsonsb	[6.48 1.33 0. 100.]	0		
6	38	beta	[2.75 186.32 0. 100.]	0		
7	37	johnsonsb	[6.64 1.64 0. 100.]	0		
8	35	johnsonsb	[6.63 1.63 0. 100.]	0		
9	35	beta	[3.08 159.23 0. 100.]	2	johnsonsb	[74.16 17.86 0. 100.]
10	33	beta	[3.08 162.15 0. 100.]	5	johnsonsb	[22.21 4.98 0. 100.]
11	30	johnsonsb	[6.09 1.47 0. 100.]	9	johnsonsb	[15.57 3.57 0. 100.]
12	28	beta	[4. 269.66 0. 100.]	9	beta	[15.63 1061.3 0. 100.]
13	31	johnsonsb	[6.67 1.57 0. 100.]	8	johnsonsb	[18.23 4.29 0. 100.]
14	35	beta	[2.95 166.6 0. 100.]	6	beta	[11.71 947.51 0. 100.]
15	36	johnsonsb	[7.67 1.81 0. 100.]	5	beta	[17.38 1482.21 0. 100.]
16	35	johnsonsb	[7.24 1.7 0. 100.]	2	johnsonsb	[27.84 6.29 0. 100.]
17	37	beta	[2.62 149.68 0. 100.]	2	johnsonsb	[18.12 4.36 0. 100.]
18	38	johnsonsb	[6.5 1.57 0. 100.]	0		
19	34	johnsonsb	[7.61 1.88 0. 100.]	0		
20	32	johnsonsb	[7.68 1.83 0. 100.]	0		
21	34	johnsonsb	[7.25 1.67 0. 100.]	0		
22	39	beta	[2.65 185.32 0. 100.]	0		
23	38	johnsonsb	[4.38 0.96 0. 100.]	0		

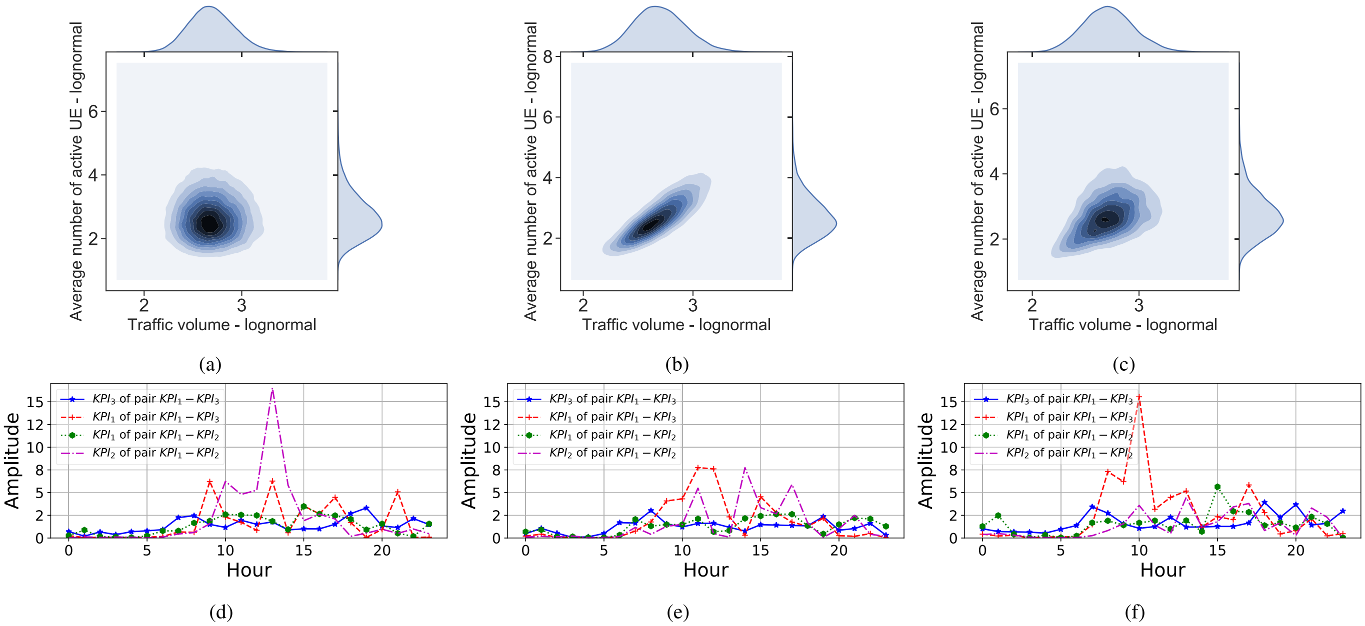

Fig. 4. Modelling results of KPI$_1$-KPI$_2$: (a) independent KPIs distributions (only time modelling), (b) Gaussian and (c) Clayton copulas related distributions for the hour 17. Generated time-series using: (d) independent models (only time modelling), (e) Gaussian and (f) Clayton copulas.

As seen, the hypothesis of network dynamism is confirmed by the different hourly models. The degradation data are limited only to a few hours, as a consequence of the behaviour of this specific problem. Since the majority of traffic occurs in peak hours, there are no samples for congestion in non-peak hours; therefore, a cell that has this root cause present will behave just like a normal one. Consequently, the model for congestion in non-peak hours is the same as the model for normal behaviour. Fig. 4(d) depicts KPI time series generated with the models shown in Table I. The presence of several local maximums in the interval 9-17h aligns with the degradation interval.

For each hour, the matrix Σ_ρ (or θ) is constructed using Pearson correlation for KPI$_1$-KPI$_2$, indicating their direct dependencies, and SSIM [13] for KPI$_1$-KPI$_3$, calculating a non-linear human based similarity. Then, a copula is created to generate new time series. Fig. 5 shows the correlation coefficients between the samples of the original selected KPIs, where the value for the hour 17 is highlighted as an example. Fig. 4(a) shows the generated KPI samples (a total of 10000 samples for each one) using their own hour 17 model, the extracted models from Table I. This figure has been created before using the copula; therefore, the variables are independent from one another, only taking into account the time modelling. Figs. 4(b) and 4(c) exhibit new distributions originated by different copula families, from the same hour and Pearson correlation. The figures show the distributions attained from an elliptic (Gaussian) and an archimedean (Clayton) copulas, respectively. Each copula family member adapts the shape of the distributions and presents the inter-dependency by how it narrows the sample dispersion. In consequence, the new built time series evolve in a particular form according to each copula, as illustrated in Figs. 4(e) and 4(f). It is important to remember that hours with no degradation data are filled with data from normal status, which are equally related with the described procedure using copula function.

Finally, in order to support the visual results, 10000 pairs of time-series have been created for each pair (KPI$_1$-KPI$_2$ and

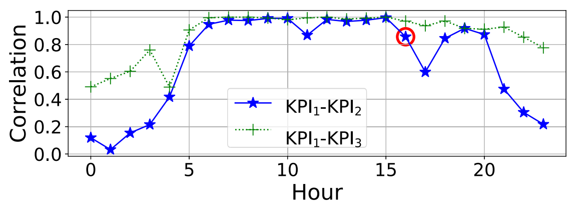

Fig. 5. Correlation values for the samples of KPI_1-KPI_2 pair and KPI_1-KPI_3 pair. The value for hour 17 is 0.86 (red circle).

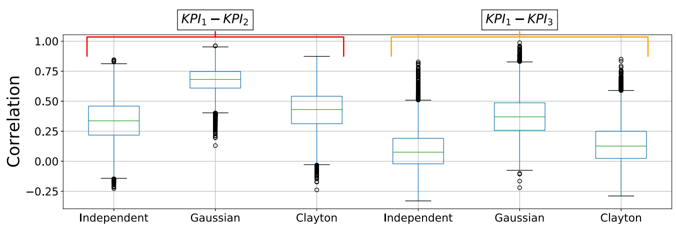

Fig. 6. Summary of Pearson correlation results of the generated time series.

TABLE II

CORRELATION MEAN AND STD DEVIATION FOR THE PEARSON CORRELATION OF 10000 GENERATED TIME-SERIES

KPI_1-KPI_2	**Independent**	**Gaussian**	**Clayton**
Mean	0.337239	0.673903	0.423458
Std deviation	0.173260	0.105509	0.167281
Mean KPI_1-original	0.542422	0.543420	0.525086
Mean KPI_2-original	0.623695	0.622807	0.502105
KPI_1-KPI_3	**Independent**	**Gaussian**	**Clayton**
Mean	0.338198	0.681243	0.402514
Std deviation	0.174194	0.109630	0.171653
Mean KPI_1-original	0.543218	0.542884	0.542937
Mean KPI_3-original	0.403222	0.404100	0.301255

KPI_1-KPI_3). Thereafter, the Pearson correlation is calculated between each pair. The summary of the results is presented in the Fig. 6. The mean and standard deviation have been disposed in the Table II to provide numbers for comparison. In the same table, comparison among the generated time series with regard to representative time series of the original data (achieved averaging samples from each hour of days with degradations) has been provided to demonstrate the general significant existing similarity. Finally, despite of the mentioned similarity, the results from pairs of time series of copulas expose higher correlations among them, specially in the case of Gaussian copula. The mean of the independent case is slightly higher than expected, which it is attributed to the similarity of some hour models obtained for each KPI.

V. CONCLUSION

In this work a new method for the generation of synthetic KPI samples is proposed, to solve the lack of samples for ML-based SH systems in mobile networks. The literature methods do not cover the KPI time dynamism, considering only the general KPI characteristics for the modelling and samples generation. Moreover, the underlying relationship among KPIs has been widely overlooked. Here, both characteristics are modelled with the use of time dependent models and the use of copulas, respectively. The relationship is modelled by the copula family and the technique selected for the correlation matrix construction, allowing flexibility in the type of searched relationship. Both aspects have been tested applying the proposed method on data from a real mobile network scenario.

A first path for new research is the study of t-Student copula; which does not have the positive relationship restriction for higher dimensions, as achimedean copulas, and weights the distribution tails more than a Gaussian copula. A second path can be directed to the simulation field. The proposed method solves a problem that occurs in the input of ML algorithms (specifically, in the training sets). However, it can also be applied over the output of simulations; modifying simulator results to exhibit a more realistic relationship, using a correlation matrix obtained from real measurements. For example, to relate SNR and throughput in scenarios where no analytical expressions exist. Finally, new approaches of feature extraction can be designed. Since the models of the KPIs for each hour are analytical statistical distributions, they can be fused to create samples of a new fused KPI with mixture characteristics, using methods mentioned in [14].

REFERENCES

[1] D. Mulvey, C. H. Foh, M. A. Imran, and R. Tafazolli, "Cell fault management using machine learning techniques," *IEEE Access*, vol. 7, pp. 124514–124539, 2019.

[2] E. J. Khatib, A. Gómez-Andrades, I. Serrano, and R. Barco, "Modelling LTE solved troubleshooting cases," *J. Netw. Syst. Manage.*, vol. 26, no. 1, pp. 23–50, Jan. 2018.

[3] A. Herrera-Garcia, S. Fortes, E. Baena, J. Mendoza, C. Baena, and R. Barco, "Modeling of key quality indicators for end-to-end network management: Preparing for 5G," *IEEE Veh. Technol. Mag.*, vol. 14, no. 4, pp. 76–84, Dec. 2019.

[4] S. Hahn, M. Schweins, and T. Kurner, "Impact of SON function combinations on the KPI behaviour in realistic mobile network scenarios," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Apr. 2018, pp. 1–6.

[5] I. Aldmour and K. Al-Begain, "Degradation of service modelling and investigation in WCDMA mobile communications," in *Proc. 10th Int. Conf. Comput. Model. Simul. (uksim)*, Apr. 2008, pp. 28–33.

[6] Y. Kiral and Y. K. Ever, "Analytical modelling and performability evaluation of LTE as dominant connectivity technology for Internet of Things," in *Proc. 30th Int. Conf. Adv. Inf. Netw. Appl. Workshops (WAINA)*, Mar. 2016, pp. 300–305.

[7] J. Yoon, D. Jarrett, and M. van der Schaar, "Time-series generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 1, 2019, pp. 5509–5519.

[8] S. Aoki, K. Shiomoto, C. L. Eng, and S. Backstad, "Few-shot learning for eNodeB performance metric analysis for service level assurance in LTE networks," in *Proc. IEEE/IFIP Netw. Oper. Manage. Symp. (NOMS)*, Apr. 2020, pp. 1–4.

[9] F. M. Castro, M. J. Marin-Jimenez, N. Guil, C. Schmid, and K. Alahari, "End-to-end incremental learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 233–248.

[10] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[11] J. Yan, "Enjoy the joy of copulas: With a package copula," *J. Stat. Softw., Articles*, vol. 21, no. 4, pp. 1–21, 2007.

[12] R. Barco, L. Nielsen, R. Guerrero, G. Hylander, and S. Patel, "Automated troubleshooting of a mobile communication network using Bayesian networks," in *Proc. 4th Int. Workshop Mobile Wireless Commun. Netw.*, Sep. 2002, pp. 606–610.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[14] M. Gunay, U. Orguner, and M. Demirekler, "Chernoff fusion of Gaussian mixtures based on sigma-point approximation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 6, pp. 2732–2746, Dec. 2016.