Dynamic Allocation of 5G Transport Network Slice Bandwidth Based on LSTM Traffic Prediction

Suchao Xiao and Wen Chen
School ofInformation Science&Technology
DongHua University
Shanghai 201620, China
suchao@maiLdhu.edu.cn, chenwen@dhu.edu.cn

Abstract-5G networks will be characterized by the extremely wide bandwidth that will be available to the user. A more flexible transmission network is necessary to support the demand generated by the access network to continuously increase network bandwidth. Transport network slicing will be a promising technology to address those challenges. In this paper, we focused on the dynamic resource allocation problem of bandwidth in transport network slices. We introduce a novel LSTM-based traffic-predict dynamic transport network slicing framework (LSTM-TPDTNS). Our approach consists of two phases: the traffic prediction phase and the bandwidth configuration phase. For the first phase, we use long and short memory models to predict traffic. For the second phase, we model our problem as a fractional knapsack problem, and we use greedy algorithms to find approximate solutions. Dynamic allocation of resources to services of different priorities can be realized, thereby improving the service quality and user experience of the entire system.

Keywords-Traffic prediction; Bandwidth allocation; Network slicing; Transport networks

I. INTRODUCTION

The next-generation mobile communications network is expected to support a large diversity of service types. The foreseeable services are mainly divided into three typical cases, enhanced mobile broadband (eMBB), ultra-reliability, low-latency connection (URLLC), and machine-type communications (mMTC)[I]. With their own specific requirements in terms of delay, capacity, and reliability, it is not realistic to build multiple different physical networks for different services. Therefore, the concept of network slicing has been proposed to address the diversified service requirements in 5G. Network slicing is an end-to-end logical network provisioned with a set of isolated virtual resources on the shared physical infrastructure. These logical networks are provided as different services to fulfill users' varying communication requirements[2].

The fifth generation (5G) of network technologies will be characterized by the extremely wide bandwidth that will be available to the user. In order to provide such a huge increase in available bandwidth, 5G radio access network (RAN) technologies will require fronthaul and backhaul solutions between the RAN and the packet core capable of dealing with this increased traffic load. In addition, the increased fluctuations in demand patterns due to network densification call for flexible transport architectures[3]. Transport network slicing will be a promising technology to address those challenges. In slice-based network architecture. the quality of slice partitioning may directly affect network performance. Therefore, how to dynamically allocate slice resources is essential to improve network performance and service quality.

In this paper, we introduce a novel LSTM-based traffic-predict dynamic transport network slicing framework (LSTM-TPDTNS). We focus on the dynamic resource allocation problem of bandwidth in the transport network slices for the three typical 5G network service mentioned above. Our approach consists of two stages: traffic forecasting followed by bandwidth provisioning. For the first stage, we use the LSTM model to forecast traffic. Experiment results indicate that the LSTM algorithm can achieve high accuracy in traffic prediction. For the second stage, we use a bandwidth provisioning scheme that allocates bandwidths depending on the traffic forecasting. It is divided into two parts. Firstly, we compare the packet loss ratios in the static and dynamic allocation schemes. The experiment results show that the total packet loss rate of dynamic resource allocation scheme is lower than static. Secondly, we set priorities for the three slice types according to service functions and emergencies, and model our problem as a Fractional Knapsack Problem for which we used a greedy algorithm in order to find an approximate solution. In this way, the dynamic allocation of resources to services of different priorities can be realized, thereby improving the service quality and user experience of the entire system.

II. RELA1ED WORK

Several studies have investigated network slicing and reallocating resources [3]-[8]. The 5G transport network architecture designed in the 5G-Crosshaul project is introduced in [3]. The proposed solution allows for flexible and efficient allocation of transport network resources to multiple tenants by leveraging on widespread architectural frameworks for NFV and SDN In [4], the authors proposed a bankruptcy game based algorithm to allocate resource for the Cloud-RAN slices. Cloud and slices are modeled to the bankrupt company and debtors in the game respectively. where Shapley value is adopted to obtain a stable solution. This algorithm significantly improves resource utilization and guarantees the fairness of allocation. The authors in [5] addressed the slicing of radio access network resources by multiple tenants. They considered a criterion for dynamic

978-1-5386-6565-7/18/$31.00

resource allocation amongst tenants, based on a weighted proportionally fair objective, which achieves desirable fairness/protection across the network slices of the different tenants and their associated users. A centralized joint power and resource allocation scheme for prioritized multi-tier cellular networks has been designed in [6]. The scheme has been developed to admit users wifh higher priority level in order to maximize the number of users. Dynamic slicing approach for multitenant 5G transport networks is introduced in[7]. The authors presented a solution for the dynamic slicing problem in terms of both mixed integer linear programming formulations and heuristic algorithms. However, they didnt consider fhe differences between service levels for different slice types. In [8], the authors introduced a novel machine learning-based traffic-aware dynamic slicing framework, which can dynamically allocate network resources and realize flexible resources sharing. However, the case where multiple slice types coexist in the network was overlooked.

Based on the above research, we introduce a dynamic bandwidfh resource allocation method based on traffic prediction using LSTM model, which maximizes resource utilization while taking the service level requirements of different slice types into account, and ultimately Improves network service quality and user experience.

III. SYSTEM MODEL

The designed framework of LSTM-based traffic-predict dynamic transport network slicing (LSTM-TPDTNS) is illustrated in Fig. I. The data plane is responsible for the actual forwarding of packets. The radio access network (RAN) is aggregated into a wavelength division multiplexing (WDM) optical metropolitan area network and connected to the core network through an optical transport network The control plane consists of RAN controller, optical transport controller, and cloud controller. It is responsible for routing information distribution and network data collection. In addition, orchestration includes data analytics module and orchestration module. Data analytics module takes charge of data collection and data analysis. The orchestration module performs resource allocation according to the predicted resource requirements.

We focus on the reconfiguration of bandwidth resources in the transport network using fhe LSTM traffic prediction mefhod and the dynamic resource allocation algorithm. Our approach is using forecasted traffic demands to calculate in advance the best resource allocation to reduce its drop probability. The workflow is shown in Fig. 2. The traffic requests and network resources occupation infonnation from the network clients are received and stored in fhe database by SDN controllers. The LSTM prediction algorithm module in orchestration application uses the data to predict the next time period traffic. Then, traffic demands are used to achieve the configuration of the optimized bandwidfh resources in the partial knapsack algorithm. Hence, the allocation of resources can adapt to the actual traffic volume dynamically. Finally, the orchestration application sends control messages to each controller and realizes the collaborative control.

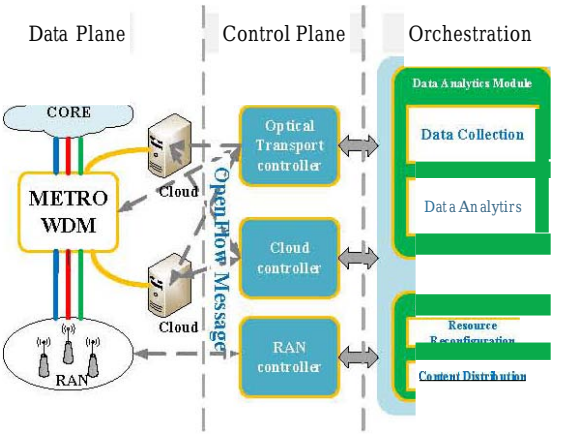

Figure. I. Framework oftraffic-aware dynamic slicing

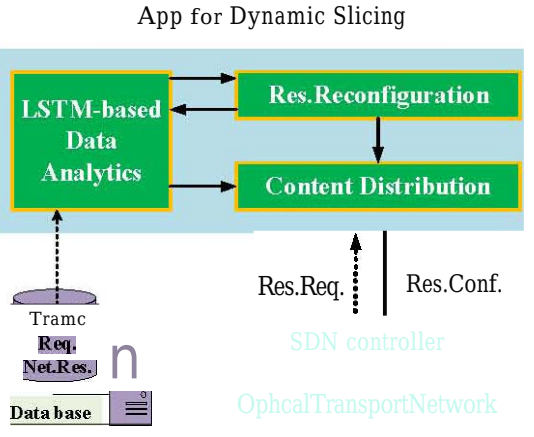

Figure. 2. workflow ofLSTM-TPDTNS

IV DESCRIPTION OF THE PROPOSED METHOD

The Knapsack Problem has been widely used to model many resource sharing problems[9]-[II]. In the standard Knapsack Problem, given a set of items, each with a weight and a value, determine the number of each item to include in a collection so that the total weight is less than or equal to a given limit and the total value is as large as possible. In our model, the weight and the value of the items can be regarded as the traffic amount of each class and traffic priority, respectively. The capacity of knapsack corresponds to the total bandwidth. Detailed parameters mapping is shown in Table1 ,

TABLE I. PARAMETERS MAPPING FOR FRACTIONAL KNAPSACK PROBLEM.

Knapsack canacltv	Total bandwidth
Items	Each service class(i.e., eMBB, URLLC, mMTC)
Weight	Expected traffic amount of each class (applying one-step ahead prediction) $\bar{X}_U$ ' $\bar{X}_R$ ' $\bar{X}_N$
Value	Relative priority of each class(e.g. mMTC:eMBB:URLLC =1 :2:3)'

Given a collection of traffic classes G = {gig 2,.. g n,}, where each traffic class gi = ⟨v_i,w i) worth v_i, and weight w, bps, our goal is to fill a backpack wifh max-capacity of Wbps

with traffic classes from G, so that the total value of items in the network bandwidth is maximized. It is allowed to take a fraction of any item. Formally, given a filling pattern $F = \{f1f\ 2,\cdots f\ n\}$ where $O \leq f_i \leq 1$, denotes the fraction of the ith item we take, the value of this pattern is $V_F = \sum_{i=1}^{n} f_j v_j$ and the weight of this pattern is $WF = \sum_{i=1}^{n} f_i w_i$. We wish to find a filling pattern $F^* = \{f_1^*, f_2^*, \ldots, f_n^*\}$ such that $W_{F^*} \leq W$, and V_{F^*} is maximized. Let $Pi = V.r\ll.$ be the value per bps (VPB) for item-i, and sort the set of items by their VPB values. Assume that $G = \{glg\ 2,\cdots g\ n,\}$ is such a sorted list. The Fractional Knapsack Problem can be solved by a greedy strategy as shown in the following algorithm.

TABLE II. FRACTIONAL KNAPSACK ALOORITIIM

Algorithm 1. Fractional Knapsack (G, n, W)

```
/* G is already sorted by the VPB values of items. */
F is an array of size n, initialized to all zeros;
A_w ← W; i ← 1;
While A_w > 0 and i ≤ n do
  If A_w ≥ w, then
    f_i ← 1;
  else
    f_i ← Aw/wj;
  end if
  A_w ← A_w ← fiwj;
  i+ +;
end while
return F;
```

The output F of the above algorithm is an optimal solution. The proposed dynamic bandwidth allocation scheme reallocates bandwidth for each traffic class based on forecasted bandwidth values and on the priority of each class. The dynamic bandwidth provisioning algorithm is as follows:

TABLE III. F DYNAMIC BANDWIDIH REpROVISIONINO ALGORITHM

Algorithm 2. Dynamic Bandwidth Reprovisioning

```
/* Note that Σ_{i=1}^{n} w_i = X̂_{Ut+1} + X̂_{Rt+1} + X̂_{Nt+1} + X̂_{Bt+1} */
If Σ_{i=1}^{n} w_i < W then
  BWeMBB ← X̂_{Ut+1} * W/Σ_{i=1}^{n} w_i;
  BWURLLC ← X̂_{Rt+1} * W / Σ_{i=1}^{n} wi;
  BWmMTC ← X̂_{Nt+1} * W/Σ_{i=1}^{n} wi;
else
  Fractional_Knapsack (G, n, W) /* return F",'
  /* i E G such that mMTC,URLLC,eMBB */
  while i ≤ n do
    BW_i ← f_1 * X̂_{it+1r}.
  end while
end if
```

When the bandwidth resources are underutilized, we allocate the bandwidth for each slice type proportional to the current traffic amount. When the bandwidth has no room to accept data packets anymore, the bandwidth for each class is reallocated according to Algorithm 1 to maximize the total value of items. It can maximize bandwidth utilization and realize the dynamic allocation of resources for different priority services, thereby improving the service quality of the system.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, the performance of the proposed scheme is evaluated in tenns of the packet drop ratio. 5G wireless systems will support three generic services, which, according to ITU-R, are classified as enhanced mobile broadband (e:tvIBB), massive machine-type communications (mMIC), and ultra-reliable and low latency communications (URLLC) (also referred to as mission-critical communications)[1],[2]. A succinct characterization of these services can be put forward as follows: (a) eMBB supports stable connections with very high peak data rates, as well as moderate rates for cell-edge users; (b) mMIC supports a massive number of Internet of Things (loT) devices, which are only sporadically active and send small data payloads; (c) URLLC supports low-latency transmissions of small payloads with very high reliability from a limited set of terminals [12]

TABLE IV. THREE TYPICAL SERVICES IN 50 NETWORK

5G Use Cases	Examples	Requirements
eMBB	4K ultra high definition (UH Drvideo	High capacity, video cache
mMTC	Sensor Networks	Massive connection coverina a larae area
URLLC	Smart-grid	Low latency and high reliability

Since the 5G network is researching, there is no appropriate mature dataset can be used. Based on the characteristics of the services and the supported service examples in the table, we simulated data set that satisfies the three service characteristics. The dataset consists of 48-h data sampled at a 5-min interval.

A. Trqffi c Prediction Experiment And Result Analysis

1) Pre-Processing Data

a) Normalization: The data set is first standardized with Min-Max scaling method so that attributes in greater numeric ranges do not dominate attributes in a smaller range. The range of features is in [-1, 1].

b) Split training and test data set: We used the first 24 hours as a training set and the last 24 hours as a test set.

c) Sliding window: Change to the mode of sliding window. The time step is the variation. Pile the training data and test data in the fonn of sliding windows to achieve better fit results.

2) Establish LSTM Model

This article is based on Keras framework of Python build the LSTM model[13]. It uses the sequence to sequence LSTM network to establish a unidirectional RNN network. When establishing LSTM model, three to four hidden layers can make it faster to converge, effectively reduce the number of cell units

in the hidden layer, and shorten the training time. This study uses a four-layer LSTM network as training model.

3) Model evaluation

We use R-squared[14] to evaluate regression loss of our prediction model. R^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An R^2 of 1 indicates that the regression predictions perfectly fit the data. The corresponding equation is given as follows:

$$R^2=1-\frac{SS_{res}}{SS_{rot}} \quad (1)$$

$$SSres = \sum_i (y_i - f_i)^2 = \sum_i e_i^2 \quad (2)$$

$$SS_{tot}=\sum_i (y_i-\bar{y})^2 \quad (3)$$

SSres is the regression sum of squares. SStot is the total sum of squares. $\bar{y}$ is the mean of the observed data. f is predicted value. Yi is the real value.

4) Experimental Results

Fig.3, 4 and 5 show the result of the traffic forecast for three service data. The blue dotted lines represent actual traffic values. The corresponding predicted values are solid lines. The experiment shows that the predicted values can capture the actual traffic values well for the three different traffic classes. The value of R^2 in the three service traffic predicting is approximately 0.96.

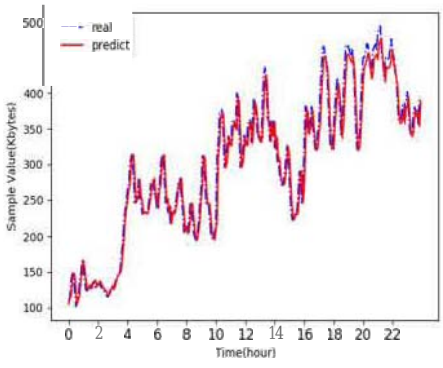

Figure. 3 URLLC traffic prediction

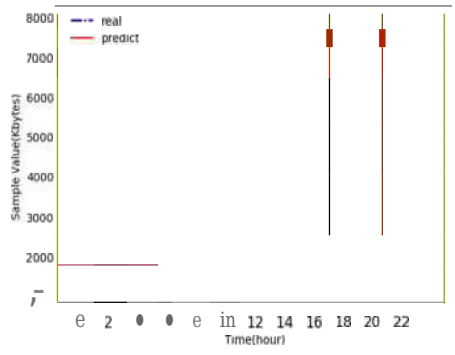

Figure. 4 eMBB traffic prediction

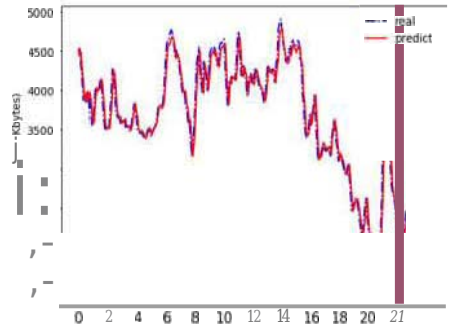

Figure. 5 mMTC traffic prediction

B. Bandwidth reallocation

In the static allocation, bandwidth for each traffic class is allocated statically based on the average traffic amount for each class. Initially we allocated *3342Kb/s* for mMfC, *3606Kb/s* for eMBB, *303Kb/s* for URLLC. The bandwidth for each of the three traffic classes is reallocated according to the proposed scheme, which is based on the predicted bandwidth amount of each class and Fractional Knapsack Problem.

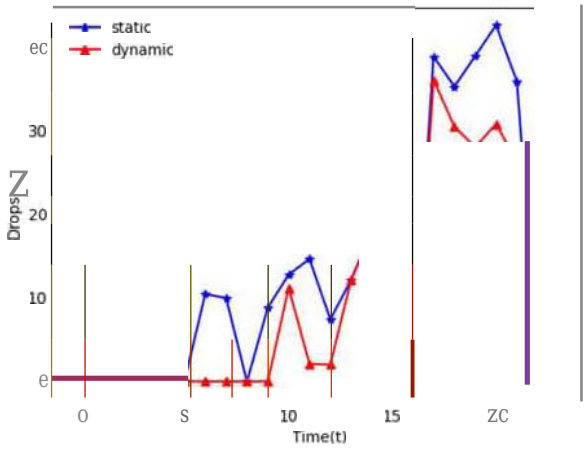

Figure. 6 The total packet drcp probabilities

Fig. 6 shows the total packet drop probabilities for the proposed scheme. For all the schemes, the number of packet drops depends on the offered traffic load. In static allocation for data set, the data loss of 99.792 Gbytes in average was monitored between t = 0 and t =24. When the proposed reprovisioning scheme was used for data set, the data loss of 69 Gbytes in average was observed between t =0 and t = 24. In most cases, the packet drop probability of the proposed scheme is smaller than the static provisioning.

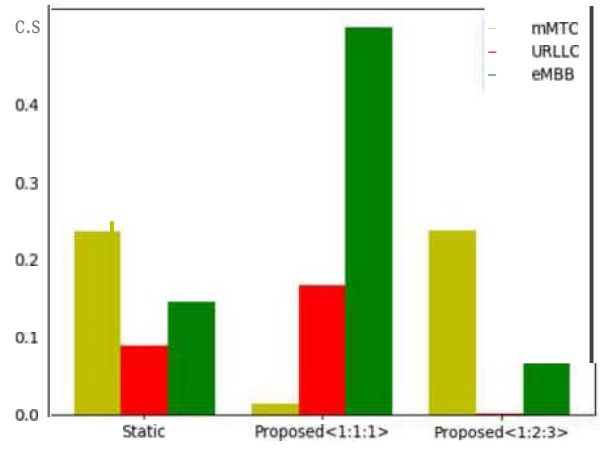

Figure. 7 Packet drop probabilities per class

In order to observe the different treatments on different traffic classes more directly, we show the individual packet drop probabilities of each traffic class. Fig. 7 presents the packet drop probabilities for the proposed scheme and static allocation. We observe an average loss of 23.63 % for mMfC, 8.91% for URLLC, and 14.57% for eMBB, in static provisioning scheme. On the other hand, we compared the drop probability between

the static provisioning scheme and the dynamic provisioning scheme in that data service type has the same priority. We find that the drop probability of mMIC (5.52%) is better compared to static provisioning, while the probability of eMBB and URLLC are worse. Because mMTC is actruelly a large amount in terms of a number of packets, and Fractional Knapsack uses the greedy algorithm as an optimization Ihe eMBB and URLLC cannot be processed, even they have the same priority withmMTC

By selecting vi- <I, 2, 3> to set the priority for the three service types, we observe that the packet loss probability of URLLC decreases to 0.21%, the packet loss probability of eMBB decreases to 0.67%, and the packet loss probability of mMIC increases to 23.76%. It is due to the different filling pattemF, which is derived by Algorithm I. eMBB and URLLC service are prioritized according to different requirements. Ihen the remaining network bandwidth is allocated to the mMIC service. By treating traffic of different service types differently, the user service experience is improved while maximizing network bandwidth utilization

VI. CONCLUSIONS

It is important to releasing unutilized bandwidth for other services for mcreasmg resource utilization. We have investigated how to increase the available bandwidth for high priority class traffic while guaranteeing adequate service quality for low priority class as welL We use LSIM model to predict the bandwidth of each traffic class, model the bandwidth reallocation problem as a Fractional Knapsack Problem, and use a greedy algorithm to find an approximate solution Ihe experiments showed that the proposed reallocation scheme can reduce the total number of packet drops. While our approach works quite well most of the time, it is not perfect When facing an extremely unexpected turn in traffic, there will be some difficulties.

VII. ACKNOWLEDGMENT

Ihis work is supported by the National Natural Science Foundation of China under grant Nos.61501108.

REFERENCES

[1] IEEE-5G, "IEEE 5G and Beyond Technology Roadmap White Paper," 2017.

[2] D. C. Mur, P. Flegkas, D. Syrivelis, Q. Wei, and J. B. T.-I. C. on U. C. and C. and 2016 I. S. on C. and S. Gutierrez, "5G-XHaul: Enabling Scalable Virtualization for Future 5G Transport Networks," 2017, pp. 173-180.

[3] A. De Oliva, X. C. P, A. Azcorra, A. Di Giglio, F. Cavaliere, 1. Lessmann, T. Haustein, A. Mourad, and P. Iovanna, "5G-Crosshaul: The 5G Integrated Fronthaul / Backhaul," vol. i, pp. 1-3.

[4] Y. Jia, H. Tian, S. Fan, P. Zhao, andK. Zhao, "Bankruptcy game based resource allocation algorithm for 5G Cloud-RAN slicing," in *Wireless Communications and Networking Conference (WCNC), 2018 IEEE, 2018*, pp.I-6.

[5] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Perez, "Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads," *IEEE/ACMTrans. Netw.*, vol. 25, no. 5, pp. 3044-3058, 2017.

[6] S. Muppala, G. Chen, and X. Zhou, "Multi-tier service differentiation by coordinated learning-based resource provisioning and admission control," *J. Parallel Dis/rib. Comput.*, vol. 74, no. 5, pp. 2351-2364,2014.

[7] M. R. Raza, M. Fiorani, A. Rostami, P. Ohlen, L. Wosinska, andP. Monti, "Dynamic slicing approach for multi-tenant 5G transport networks [invited]," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 10, no. 1, pp. A77-A90,2018.

[8] C. Song, M. Zhang, X. Huang, Y. Zhan, D. Wang, M. Liu, and Y. Rong, "Machine Learning Enabling Traffic-Aware Dynamic Slicing for 5G Optical Transport Networks," *2018 Con! Lasers Electro-Optics,* no. c, pp. 1-2,2018.

[9] D. C. Vanderster, R. Parra-Hernandez, R. Parra-Hernandez, and R. 1. Sobie, "Resource allocation on computational grids using a utility model and the knapsack problem," *Futur. Gener. Comput. Syst.*, vol. 25, no. 1, pp. 35-50, 2009.

[10] P. Jacko, "Resource capacity allocation to stochastic dynamic competitors: knapsack problem for perishable items and index-knapsack heuristic," *Ann. Oper. Res.*, vol. 241, no. 1-2, pp. 1-25, 2016.

[11] R. Min, G. Branch, and C. T. Co, "Resource Allocation in Cognitive OFDM System Based on Knapsack Problem," *Video Eng.*, 2013.

[12] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View," 2018.

[13] N. Navarin, B. Vincenzi, M. Polato, and A. Sperduti, "LSTM Networks for Data-Aware Remaining Time Prediction of Business Process Instances," 2017.

[14] M. S. Lewis-Beck, "R-squared," 2004.