Processing ANN Traffic Predictions for RAN Energy Efficiency

Processing ANN Traffic Predictions for RAN Energy Efficiency

Greta Vallero
Politecnico di Torino
greta.vallero@polito.it

Daniela Renga
Politecnico di Torino
daniela.renga@polito.it

Michela Meo
Politecnico di Torino
michela.meo@polito.it

Marco Ajmone Marsan
Politecnico di Torino
marco.ajmone@polito.it

ABSTRACT

The field of networking, like many others, is experiencing a peak of interest in the use of Machine Learning (ML) algorithms. In this paper, we focus on the application of ML tools to resource management in a portion of a Radio Access Network (RAN) and, in particular, to Base Station (BS) activation and deactivation, aiming at reducing energy consumption while providing enough capacity to satisfy the variable traffic demand generated by end users. In order to properly decide on BS (de)activation, traffic predictions are needed, and Artificial Neural Networks (ANN) are used for this purpose. Since critical BS (de)activation decisions are not taken in proximity of minima and maxima of the traffic patterns, high accuracy in the traffic estimation is not required at those times, but only close to the times when a decision is taken. This calls for careful processing of the ANN traffic predictions to increase the probability of correct decision. Numerical performance results in terms of energy saving and traffic lost due to incorrect BS deactivations are obtained by simulating algorithms for traffic predictions processing, using real traffic as input. Results suggest that good performance trade-offs can be achieved even in presence of non-negligible traffic prediction errors, if these forecasts are properly processed.

KEYWORDS

Radio access network; base station; energy efficiency; traffic prediction; neural network

ACM Reference Format:
Greta Vallero, Daniela Renga, Michela Meo, and Marco Ajmone Marsan. 2020. Processing ANN Traffic Predictions for RAN Energy Efficiency. In *23rd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '20), November 16–20, 2020, Alicante, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3416010.3423222

1 INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) based solutions are being increasingly studied and applied in several domains of the Information and Communication Technology (ICT) sector, among which networking is not an exception. The growth of computational power, the availability of data, the improvement of learning algorithms are the boosts behind the pervasiveness of new AI- and ML-based mechanisms to respond to the new challenges of today's networks, which are often too complex to be properly understood, modeled, and managed with traditional approaches. This is the case of decision making for network management and configuration in presence of a huge set of parameters and fast changing scenarios, but also of catching the effect of complex interactions among multitudes of heterogeneous users and network elements (such as macro and small cells in heterogeneous networks), as well as understanding the hidden correlations among systems.

While these approaches are attractive because they are, by their nature, suited to handle problems with very large state spaces and complexity, in practice, effectively exploiting the potential of ML technologies is not easy. A deep domain knowledge is needed, as well as a careful processing of the outputs of ML tools. In this paper, we experiment this in Radio Access Network (RAN) management. We consider a portion of a RAN in which Base Stations (BSs) of macro and small cells can be activated and deactivated based on a variable traffic demand, so as to reduce energy consumption while guaranteeing that enough capacity is provided to satisfy the demand. BS activation and deactivation decisions are taken based on traffic predictions that are performed through Artificial Neural Networks (ANN). In order to properly operate the network so as not to deteriorate QoS, the outputs of the ANNs have to undergo a number of processing steps that, driven by a deep domain knowledge, are carefully tailored for the scope.

Solutions for RAN management and resource on demand provisioning have been formulated in several contexts and with a multitude of different objectives: the trade-off between the opposite needs to reduce energy consumption and provide Quality of Service (QoS) is a timely objective, motivated by concerns on sustainability, climate change, and network operational cost increase. The deployment of BS management mechanisms, in its turn, is easier today due to the flexibility of new network architectures and it is effective for energy consumption reduction due to the typical traffic demand profiles at the edge of the network. The demand profiles are characterised by (often short) peaks, followed by (often long - especially during night) valleys, and this makes the installed RAN equipment under-utilised for long periods of time. During these under-utilisation periods, some of the RAN equipment can be put in low consuming sleep modes. Moreover, in some areas and in some periods (typically right after technology upgrades), the RAN capacity is over-provisioned even with respect to traffic peaks, and this makes BS management even more attractive for energy saving purposes.

MSWiM '20, November 16–20, 2020, Alicante, Spain

ACM ISBN 978-1-4503-8117-8/20/11…$15.00
https://doi.org/10.1145/3416010.3423222

As mentioned above, BS activation and deactivation decisions are taken based on traffic predictions. When the traffic is predicted to be small enough, some small cell BSs can be deactivated, and traffic is carried by the small and macro BSs that remain active. Conversely, when traffic grows and additional capacity is needed, some BSs in sleep mode are re-activated. In [19], the performance of this BS management strategy was tested using several different ANNs for traffic predictions. The results showed a limited sensitivity to the type of ANN. Indeed, critical BS (de)activation decisions are taken in correspondence of specific traffic values, and high accuracy in the estimations is not required in general, but only close to the times when decisions are taken. Hence, to significantly improve performance, traffic predictions need to be carefully processed and the overall pattern understood. In this paper, we show that processing of ANN outputs is fundamental for improving performance.

The paper is organised as follows. After the related work discussed in Section 2, in Section 3 we present the scenario and the methodology of our study. The proposed approach is based on traffic predictions, obtained with the tools that are presented in Section 4, and on the prediction processing that is reported in Section 5. After presenting performance indicators in Section 6, results are discussed in Section 7. Conclusions are drawn in Section 8.

2 RELATED WORK

In literature, many RAN management solutions have been proposed, based on Resource on Demand (RoD) approaches. An overview of the RoD strategies to dynamically adapt the set of active radio resources to the current traffic demand is presented in [4, 5, 16]. With the purpose of reducing energy demand and limiting the RAN operational cost, the authors of [8] exploit RoD strategies to adapt the energy consumption to the actual traffic load. In [9], a framework is proposed to efficiently allocate spectrum resources to users, switching off unneeded BSs, in order to minimise power consumption. A time-varied probabilistic ON/OFF switching algorithm for cellular networks is presented in [13]. In [1, 14], RoD strategies are applied in a green mobile access network, with the objective of improving the interaction with the smart grid in a demand-response scenario, thus reducing the electricity bill and providing ancillary services.

Many of these works aim at dynamically allocating resources, under the assumption that the future traffic demand is exactly known. This means that predictions of the amount of traffic demand are necessary in order to make the proposed approaches viable. This aspect is very critical, since errors in the traffic estimation can significantly affect the performance of these strategies. If the traffic demand is overestimated, waste of energy occurs; in case of traffic underestimation, incorrect BS deactivations may deteriorate QoS. To overcome the issue related to QoS deterioration due to traffic underestimation, [3] uses a deep learning neural network structure based predictions, employing a customised loss function, to predict the needed network capacity. In particular, in case of underestimation, such function gives higher penalty than when the needed capacity is overestimated.

In the recent literature, many works focus on traffic estimation. In [10], an Auto-Regressive Integrated Moving Average (ARIMA) is used for the prediction of mobile data traffic and a Seasonal ARIMA (SA) model is used in [7]. These works demonstrate that these two methods provide high accuracy, but require slow training and forecasting, which make them impractical for on-line forecasting. The work presented in [15], uses Markovian models, while [7], [6], [17], [18], [21], [20], [12] employ ML approaches. According to [7] and [6], ANNs provide promising results in forecasting the hourly amount of traffic in TCP/IP networks. In [17], very good performance is reached in the forecast of the mobile traffic of an LTE BS, using a Recurrent Neural Network (RNN) and 1 ms resolution data. High accuracy in traffic predictions is achieved with the same approach, in [18]. An hybrid scheme, structured in an ANN and a RNN is discussed in [21]. Moreover, [20] and [12] predict traffic demand with Least Squares Support Vector Machines and Linear Regression based approach, respectively.

Table 1: Values of the parameters of the consumption model for macro and small cell BSs.

BS type	N_{trx}	P_{max} (W)	P_0 (W)	Δ_p
Macro	6	20	84	2.8
Small	2	6.3	56	2.6

3 SCENARIO

A portion of an heterogeneous LTE RAN is considered, comprising one macro cell BS, and a few small cell BSs, whose coverage overlaps with the macro cell. Small cell BSs are deployed to provide additional capacity during high traffic demand periods. A centralised Management and Orchestration System (MANO) decides the activation of resources (i.e., small cell BSs), according to predictions of the future traffic demand. These predictions are performed on a temporal horizon of 15 minutes. Small cell BSs can be switched on and off in order to reduce the RAN energy consumption, without compromising QoS. This means that, when not all the capacity is needed to satisfy the predicted traffic demand, some small cell BSs are put in sleep mode. On the contrary, all BSs are activated in those periods when all the capacity is required for the traffic demand satisfaction. The activation/deactivation of a BS cannot occur at intervals shorter than one hour, in order to avoid too frequent switches, which can induce equipment failures and impair device lifetime [11].

The input power required for the operation of a BS, denoted as P_{in}, is derived according to the linear model proposed in [2]:

$$P_{in} = N_{trx} \cdot (P_0 + \Delta_p P_{max} \rho), \qquad 0 \leq \rho \leq 1 \tag{1}$$

where N_{trx} is the number of transceivers, P_0 represents the power consumption when the radio frequency output power is null, Δ_p is the slope of the load dependent power consumption, ρ is the traffic load and P_{max} is the maximum radio frequency output power at maximum load. Table 1 summarises the value of the parameters for macro and small cell BSs [2]. The consumption of the BS in sleep mode is considered negligible.

Traffic predictions are performed through ANNs, and are processed before reaching a decision about BS activation and deactivation. The BS management works in two phases.

(1) **Training phase**. This phase is performed only once, as a preliminary step of our online management system. In this

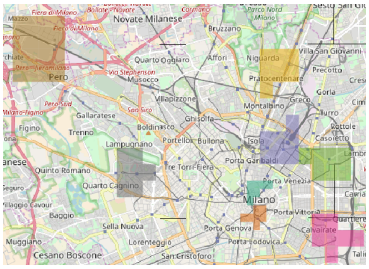


Figure 1: Considered traffic areas: train station (purple), Rho Fiere (brown), Duomo di Milano (orange), Politecnico di Milano (light green), San Siro (grey), a business area (dark green), a residential area (yellow), an industrial area (magenta).

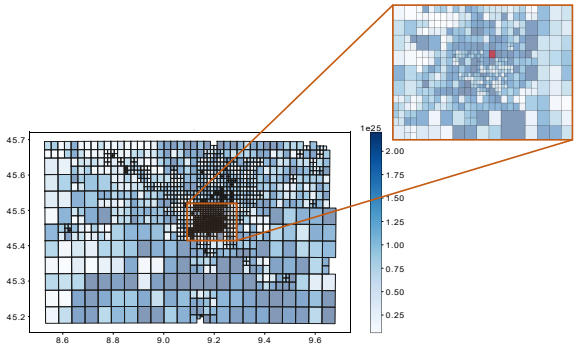


Figure 2: Spatial cross-correlation among cells with lag= 15 minutes.

step, the ML algorithm used to predict the traffic demand is trained using historical data.

(2) **Run-time phase**. The traffic is predicted using the previously trained ANN, and BS activations or deactivations are decided. During this phase, the following two steps are performed at every time slot, i.e., every 15 minutes:
   (a) **Prediction.** The traffic demand is forecast for the following 4 time slots.
   (b) **Prediction processing and decision.** The four predictions are processed by the MANO, which decides which small cell BS must be active in the next hour.

4 TRAFFIC PREDICTIONS

In this section, we present the traffic prediction tools used during run-time as a preliminary step to the BS management decision. An ANN-based approach is used in this work. This is because, as mentioned, in the literature, the potentiality of ANN has been widely demonstrated [7] [6]. Moreover, [19] shows that the performance of the BS activation/deactivation are not significantly affected by the ML approach used for the traffic predictions. Thus, the usage of a simple ANN represents a good trade-off between performance and complexity.

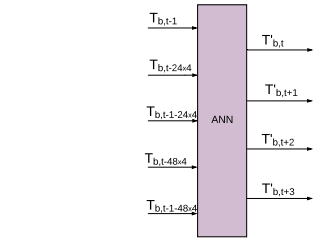
4.1 Input Data

Data provided by a large Italian mobile network operator are used in this study. They report the traffic demand volume, in bit, of 1420 BSs located in the city of Milan (Italy) and in a wide area around it, for two months in 2015, with granularity of 15 minutes. The traffic traces are normalised; hence, the peak of each traffic pattern is equal to the maximum capacity of each BS. Note that this is a pessimistic assumption with respect to energy saving possibilities, since the capacity of the network is usually overdimensioned. For our work, eight portions of the city are selected, which are shown in Fig. 1. These areas were selected as samples of quite different scenarios, and, hence, traffic patterns. All together, the selected areas are representative of the various zones that coexist in a urban environment. The train station area (purple square in Fig. 1) is characterised by intense activity levels, especially at the beginning and at the end of the working hours. The Rho Fiere district (brown in the figure) is an area that hosts big events, fairs and exhibitions, that last for a few days. The Duomo di Milano area (orange square) is a touristic area, with high activity during several hours of the day. The Politecnico di Milano area (light green) hosts a large campus with many students. The San Siro neighbourhood includes a large soccer stadium (grey), and the activity here is quite bursty and variable depending on the scheduled matches and concerts. A part of a business neighbourhood (dark green) and some residential streets (yellow) are also considered: the traffic in these areas follows the typical behaviour of people in their daily life. In the business area, traffic peaks are observed during the central hours of the day, whereas in the residential area a traffic rise is observed in the evening. Finally, the industrial zone (magenta) is a particular case of a business area. In each of these portions of the RAN, we assume that one macro BS and 6 small cell BSs are present, so that the service area is covered by one macro cell which overlaps with 6 small cells. To do this, for each area, we selected 7 traffic patterns recorded in that area. The trace which presents the highest traffic demand is chosen as the macro cell BS, while the remaining six as micro cell BSs.

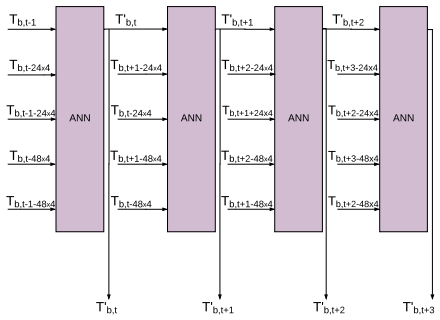
4.2 Selection of the ANN input features

In order to predict traffic demand, the ANN must be fed with carefully selected input features. The investigation of the best choice for the ANN input features was made accounting for the temporal and spatial correlations of traffic. In particular, we exploit the traffic temporal periodicity (which we observed to be present in most traffic patterns), due to the periodicity of human activities, and we investigated the possibility of also using the spatial correlation which is expected to be present among adjacent cells. In Fig. 2, the cross-correlation obtained between the traffic at one BS in the city centre, indicated in red, and all others is plotted, choosing as time lag 15 minutes. We can see that correlation only mildly depends on the spatial closeness to the considered BS (darker colours correspond to higher correlation values). Indeed, high correlation values are present even among cells that are very far from each other. For this reason, in this paper we focus only on input features based on the temporal periodicity of traffic patterns.

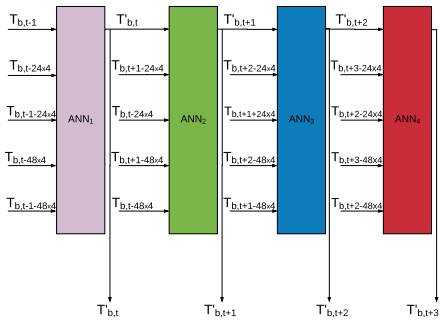
Let us define by $T_{b,i}$ the traffic demand at BS b and time slot i. For simplicity of notation, in what follows we drop the index b



(a)



(b)



(c)

Figure 3: Scheme of the three proposed prediction techniques: (a) 1 ANN-4 outputs, (b) 1 ANN-1 output, (c) 4 ANNs-1 output

if there is no ambiguity. At the beginning of each time slot t, the traffic demand at time t (the time slot that is just beginning), *t+1* (the following time slot), *t+2* and *t+3* are predicted; predictions are denoted by T'_t, T'_{t+1}, T'_{t+2}, T'_{t+3}, respectively.

The prediction tool receives as inputs:

- T_{t-1}: the traffic at the time slot just past, i.e., $t-1$;
- $T_{t-(24\cdot4)}$: the traffic one day before the current time slot (the factor 4 comes from our time slots being 15 minutes long);
- $T_{t-1-(24\cdot4)}$: the traffic one day before the time slot just past;
- $T_{t-(48\cdot4)}$: the traffic two days before the current time slot;
- $T_{t-1-(48\cdot4)}$: the traffic two days before the time slot just past.

4.3 Traffic Forecast Approaches

Different ANN-based prediction approaches are tested.

4.3.1 1 ANN-4 outputs. One ANN for each BS is used. At time t, the ANN outputs the traffic demand samples at time slots *t*, *t+1*, *t+2* and *t+3* (see Fig. 3a).

4.3.2 1 ANN-1 output. One ANN for each BS is used. The ANN is trained to predict the traffic demand at the current time slot, e.g. at time t, and it is used in cascade to predict also the three future traffic samples, e.g. at time $t+1$, $t+2$, $t+3$. This means that the ANN produces the prediction of the traffic demand at time t, namely T'_t, using in input the traffic at previous time slot, T_{t-1}, as well as the traffic of previous days. Once T'_t is computed, for predicting the traffic at time $t+1$, the same ANN is used but it receives as input the predicted traffic T'_t instead of T_t that is unknown. Similarly, for the prediction of traffic at times $t+2$ and $t+3$, predictions are used instead of traffic samples for the unknown values of the input. The logical schema is reported in Fig. 3b.

4.3.3 4 ANNs-1 output. Four ANNs are used for each BS. Each ANN is dedicated to the prediction of the traffic demand at a given time lag. This means that the 4 future traffic samples are separately predicted, using 4 different ANNs, but the inputs are as in the previous case: predictions are used instead of missing samples whenever needed. The schema is reported in Fig. 3c.

As in [19], each ANN mentioned above is structured in 3 layers: the input layer which has 8 nodes, one hidden layer with 17 nodes, and the output layer with one node, if *1 ANN-1 output* and *4 ANNs-1 output* are employed, or 4 nodes in case of *1 ANN-4 outputs* usage. The number of layers, as well the number of nodes for each layer are among the hyper-parameters which need to be selected. These have been chosen in order to achieve a good trade off between the accuracy and the time needed to train the network. Each ANN is trained minimising the Mean Squared Error (MSE).

5 PROCESSING TRAFFIC PREDICTIONS

After the ANN has generated traffic predictions, they must be processed by the MANO to decide about small cell BS activation, with the objective to save energy, without compromising QoS. In this section, we propose some strategies for processing the predictions and deciding resource activation and deactivation.

5.1 Resource Allocation

Different algorithms can be used to combine traffic predictions in a BS management strategy, based on the approach in [8], which states that a small cell BS is switched off if its traffic demand is lower than a threshold ρ, provided that such amount of traffic can be carried by the macro cell BS. The threshold depends on the energy consumption per carried bit: when the traffic is below ρ, the energy needed to carry a unit of traffic in the small cell is larger than in the macro, so that it is more convenient to switch off the small cell BS, if this is possible in terms of total capacity. As demonstrated in [8], the optimal value of the threshold is 37% of the maximum load of the BS.

5.1.1 Max2Max. In this case, resources can be allocated only at the beginning of each hour: at 00:00, 01:00, etc. At the beginning of each hour, $T'_{b,t}$, $T'_{b,t+1}$, $T'_{b,t+2}$, $T'_{b,t+3}$, the 4 traffic demands corresponding to that hour are predicted for each small cell BS b, as well as for the macro cell B; predictions in the macro are denoted by $T'_{B,t}$, $T'_{B,t+1}$, $T'_{B,t+2}$, $T'_{B,t+3}$. Among these 4 samples, the maximum, M'_b, for each small cell BS b and the maximum, M'_B, for the macro cell are computed:

$$M'_b = \max\left(T'_{b,t}, T'_{b,t+1}, T'_{b,t+2}, T'_{b,t+3}\right) \tag{2}$$

$$M'_B = \max\left(T'_{B,t}, T'_{B,t+1}, T'_{B,t+2}, T'_{B,t+3}\right) \tag{3}$$

A small cell BS b is switched off if M'_b is lower than the threshold, and its traffic can be carried by the macro, given that the macro is expected to be carrying an amount of traffic M'_B:

$$\text{if } \left(M'_b < \rho\right) \wedge \left(M'_B + M'_b < C\right) \rightarrow \text{switch off } b \tag{4}$$

where C is the capacity of the macro cell B. Basically, the decision is taken based on the maximum of the predicted traffic samples. As the decision is taken, M'_B is updated accordingly, to account for the traffic load that will be transferred from the considered BS.

5.1.2 Max2Max Continuous. This strategy is very similar to *Max2Max*, but, in this case, it is applied at the beginning of each time slot and not only at the beginning of an hour, as in the previous case. The decision to switch off a cell for 4 consecutive time slots (1 hour) can be taken in any time slot.

5.1.3 I2I. When this strategy is used, the switch on/off is possible only at the beginning of each hour. Given the four predicted traffic demands belonging to the considered hour, for each small cell BS b and for the macro cell BS B, a small cell BS b is switched off when, for every slot $t+i$ with $i = 0, 1, 2, 3$, the estimated traffic $T'_{b,t+i}$ is lower than the threshold ρ and there is enough available capacity on the macro BS:

$$\begin{aligned}\text{if } \forall i = 0, \cdots, 3 \quad & \left(T'_{b,t+i} < \rho\right) \wedge \left(T'_{b,t+i} + T'_{B,t+i} < C\right) \\ & \rightarrow \text{switch off } b\end{aligned} \tag{5}$$

In this case, the decision to switch off is taken if the requested conditions are verified slot by slot.

5.1.4 I2I Continuous. When this strategy is used, *I2I* is applied at the beginning of each time slot. As in *Max2Max Continuous*, each small cell BS remains active or in sleep mode for at least 1 hour (4 consecutive time slots), but a change of state can happen in any time slot.

5.1.5 I2I Flexible. This is a further variation of *I2I Continuous*. As before, at the beginning of each time slot, *I2I* is applied. Nevertheless, when a small cell BS has been put in sleep mode for at least one hour, it remains sleeping if the necessary conditions are verified for one more time slot. This means that when we are at time t, given that the small cell BS has been deactivated since at least t-4, that small cell BS remains in sleep mode, if $T'_{b,t}$ is lower than ρ, provided that $T'_{b,t}$ can be carried by the macro BS during the t-th time interval.

Each of the considered micro cell BSs is analysed for its possible deactivation as described above, starting from the least loaded to the most loaded, in the following hour. Given that the load of a micro BS is lower than ρ, its energy consumption per bit is larger, if its load is smaller. Thus, giving larger priority to micro BSs in the deactivation procedure leads to minimum network energy consumption [8]. The load during the following hour on each micro BS is given by summing the traffic demand during the 4 time slots belonging to that hour.

5.2 Descending front detection

The presence of noise in traffic patterns may result in incorrect deactivation of the small cell BSs, thus deteriorating QoS. For this reason, the concept of *Descending Front Detection* (DFD) is introduced in the processing of traffic predictions. In particular, the switching from active to sleep mode of a small cell BS is permitted only if a descending front is detected: if an active small cell BS is detected to be in a descending phase, the necessary conditions for the small cell switchoff are checked. Because of the noise inherent in traffic patterns, a negative first derivative is not a sufficiently good indicator of a descending front. Therefore, a moving average filter is used for this purpose. It smooths data by replacing each traffic sample with the average of the neighbouring samples. This operation practically acts as a low-pass filter on traffic patterns. In our case, a triangular smoothing is applied twice. In particular, at time t, the following expression is computed, for z = t-4, t-5, t-6:

$$S'_{b,z} = \frac{1}{81} \sum_{j=-2}^{2} (3 - |j|) \sum_{i=-2}^{2} (3 - |i|) T_{b,z+j+i} \tag{6}$$

where $T_{b,z+j+i}$ is the real traffic demand on BS b at time z+j+i. However, notice that for $z = t-4$ and for $j = 2$ and $i = 2$, $T_{b,z+j+i}$, is $T_{b,t}$, which is not known. Thus, its prediction, $T'_{b,t}$ is used in this case. The maximum z is chosen equal to t-4, in order to avoid using other predicted samples.

If $S'_{b,t-4} < S'_{b,t-5} < S'_{b,t-6}$, we conclude that a descending front is detected. If this is the case, the necessary small cell BS switchoff conditions are checked. If they are verified, as described in section 5.1, the considered small cell BS can be deactivated.

6 KEY PERFORMANCE INDICATORS

6.0.1 Average Relative Error. The ARE (Average Relative Error) measures the average relative error between the real and predicted traffic samples. It is computed as:

$$ARE = \frac{1}{N_{BS}} \sum_{b=1}^{N_{BS}} RE_b \tag{7}$$

where N_{BS} is the number of the considered BSs and RE_b is the RE (Relative Error) on BS b, derived as:

$$RE_b = \frac{1}{H} \sum_{t=1}^{H} \frac{|T_{b,t} - T'_{b,t}|}{T_{b,t}} \tag{8}$$

where $T_{b,t}$ is the real traffic demand at time t on BS b, $T'_{b,t}$ is the forecast traffic demand at time t on BS b, H is the duration of the testing period.

6.0.2 Energy Consumption Reduction. When the resource allocation strategies presented in Section 5.1 are used, in each time slot some BSs are active and consume energy, while some others may

Table 2: Average relative error, ARE, with the different approaches at different time lags.

ARE	1 ANN-4 outputs	1 ANN-1 output	4 ANNs-1 output	4 ANNs-1 output (spatial)
ARE_t	0.33	0.33	0.33	0.37
ARE_{t+1}	0.43	0.44	0.43	0.47
ARE_{t+2}	0.52	0.52	0.48	0.52
ARE_{t+3}	0.61	0.57	0.52	0.54

be in sleep mode and thus consume no (or very little) energy. The energy consumption of each BS is given by (1). In order to measure the effectiveness of these strategies, the energy consumption saving is computed. It is calculated with respect to the *always ON* scenario: this is the case in which all the BSs are always active regardless the amount of traffic demand.

6.0.3 Lost Traffic. This metric evaluates the QoS deterioration due to traffic prediction errors. It is defined as the percentage of the traffic demand that cannot be carried by the network, accounting for the fact that in each time slot some BSs are active and can handle their traffic demand, while some others may be off and thus cannot provide any service. Let us define the traffic that overflows from the small cell BS b to the macro cell as:

$$O_{b,t} = \begin{cases} T_{b,t} & \text{if } b \text{ is in sleep mode} \\ 0 & \text{if } b \text{ is active} \end{cases} \tag{9}$$

the lost traffic is given by:

$$L = \frac{\sum_{t=1}^{H} \max(0, T_{B,t} + \sum_{b=1}^{N_{BS}} O_{b,t} - C)}{\sum_{t=1}^{H} \left(T_{B,t} + \sum_{b=1}^{N_{BS}} T_{b,t} \right)} \cdot 100 \tag{10}$$

where C is the capacity of a BS. The lost traffic is the percentage of traffic that cannot be carried by the macro cell BS when traffic overflows from deactivated small cell BSs.

7 PERFORMANCE EVALUATION

In this section, we discuss numerical results obtained by experimenting the different prediction, processing and decision algorithms presented in the previous sections on the considered RAN portions. Out of the 61 days for which we have real traffic data, the first 47 are used for the ANN training phase, while the remaining 14 days are used for the run-time phase.

7.1 Choice of the ANN

As a first step, we analyse the effectiveness of the different ANN configurations for traffic predictions, using the previously defined ARE (average relative error) as a performance metric. The results provided by the considered ANN configurations, namely *1 ANN-4 outputs*, *1 ANN-1 output* and *4 ANNs-1 output*, for each time lag, are reported in Table 2, averaged over the eight considered geographical areas. Observe that numerical results confirm what is intuitively expected, and was quantitatively shown in [17]: the error increases with the time horizon of the predictions. Moreover, typically, the *1 ANN-4 outputs* provides the largest ARE. This is because, when the other 2 approaches are used, the sample corresponding to the most recent traffic demand, even if only predicted, is provided as an input feature. In Fig. 4 the percentage of the reduction of ARE gained with *1 ANN-1 output* and *4 ANNs-1 output*, with respect to *1 ANN-4 outputs*, are shown in blue and orange, respectively. The reduction of the estimation error is largest for *4 ANNs-1 output*, especially when the time horizon of the predictions is longer. This is because this ANN configuration uses 4 ANNs: during the training phase, each ANN learns how to forecast the desired output, managing the error which affects the input traffic sample derived from a prediction. For these reasons, in the rest of this study we will mostly use *4 ANNs-1 output* for traffic forecast. In order to confirm the mild correlation among adjacent cells, we also report the ARE which is obtained when we provide to *1 ANN-4 outputs* an additional input feature. In order to select this additional input feature, the cross-correlation between the traffic demand of the current BS and each of its adjacent ones, is performed. Then, the argmax function is performed, in order to select the BS bs_{MAX} and the time lag l_{MAX} which provide the largest value of cross-correlation. Thus, when we are predicting the traffic demand at time t, the additional input feature is the traffic demand on bs_{MAX} at time t-l_{MAX}. Similarly, for the prediction of the traffic demand at *t+1*, *t+2* and *t+3*, the traffic demand on bs_{MAX} at time *t+1*-l_{MAX}, *t+2*-l_{MAX}, *t+3*-l_{MAX}, respectively, are given as additional input feature. From Table 2, it is possible to notice that the presence of this feature deteriorates the precision of the forecast.

7.2 Dynamic resource allocation performance

We now investigate the performance of the resource allocation strategies presented in Section 5.1. Our solutions are compared against 3 benchmarks: (*i*) the *TNSM19* approach presented in [19], which allocates the resources of a RAN according to the hourly traffic predictions obtained using an ANN, with no processing of the ANN outputs; (*ii*) the *PIMRC18* approach: in this case the traffic is predicted using the LSTM (Long Short Term Memory) network proposed in [17] and resources are allocated based on *I2I*; (*iii*) the *15 min* approach, similar to the *TNSM19* case, but operated over 15 minutes time slots. Each small cell BS can be switched to/from sleep mode as soon as needed, with no constraint on the frequency of switching.

For each strategy and zone, Fig. 5a reports the energy consumption reduction computed with respect to the always ON scenario, and Fig. 5b reports the percentage of lost traffic.

First, it is possible to confirm that, as expected, the percentage energy saving directly depends on the shape of the traffic pattern (peak/off-peak ratio, duration of peaks, ...), which is characteristic of the considered area. If the traffic demand is low for many hours, the BS management approach can be very effective: up to 40% of the energy consumed with respect to the always ON approach can be saved, as we see in the San Siro and Rho Fiere areas. When the traffic demand is larger than ρ for longer periods, the small cell BSs cannot be switched off for shorter periods, and a lower amount of energy is saved. This is the case of the PoliMi and Train Station areas, where the energy saving is lower than 15%.

In addition, notice that the reduction of the energy consumption obtained with our proposals is slightly lower than with the chosen

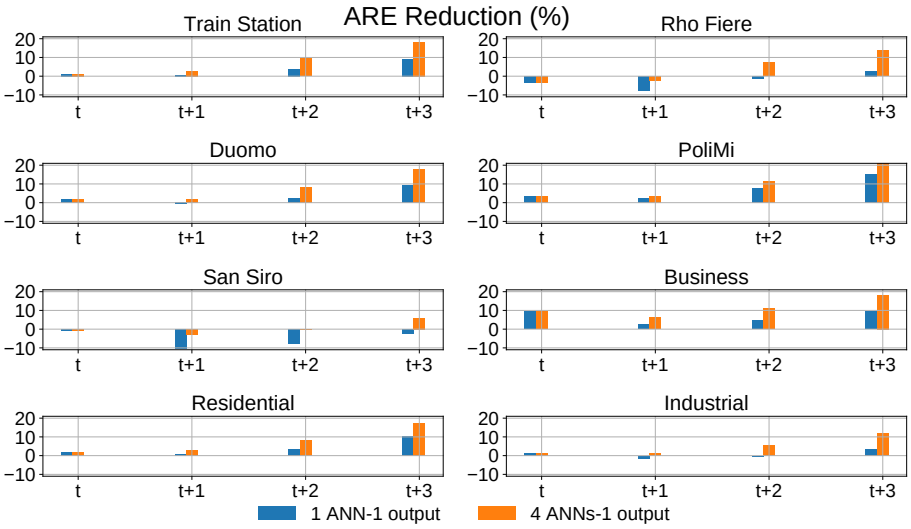
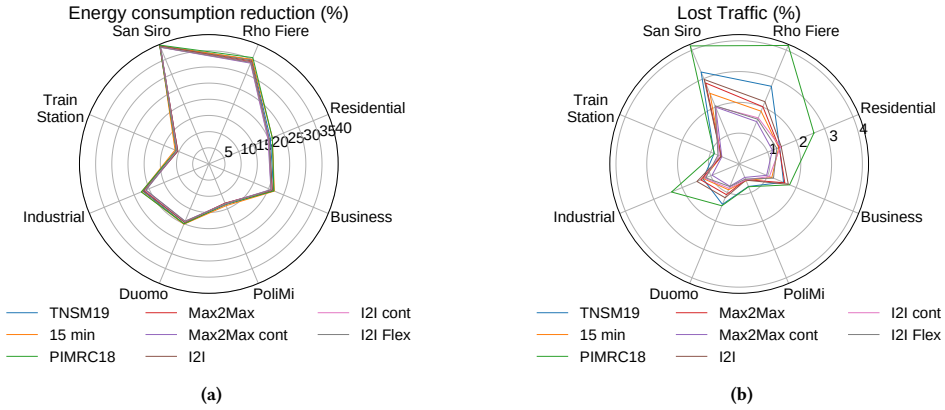


Figure 4: Percentage of the reduction of absolute relative error, ARE, obtained by *1 ANN-1 output* or *4 ANNs-1 output* with respect to the *1 ANN-4 outputs*, for different time lags.



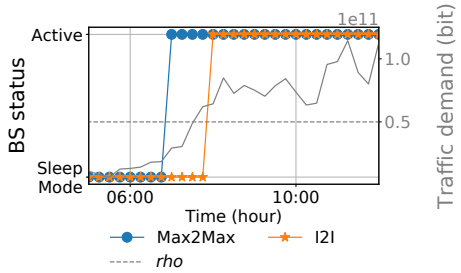
(a) (b)

Figure 5: Comparison of dynamic resources allocation strategy in the various areas: (a) Energy consumption reduction and (b) and lost traffic.

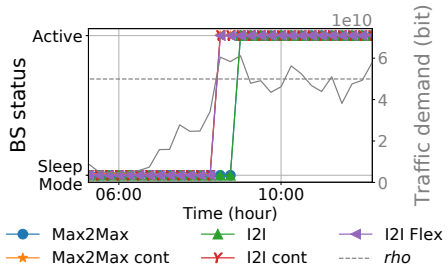
benchmarks. Indeed, with our proposals the energy consumption increases, at most, by 1.7%, 1.9% and 2.9%, with respect to the *TNSM19*, *15 min* and *PIMRC18* benchmarks. This is because our approaches are slightly more conservative in energy saving, but better preserve QoS, measured as the percentage of lost traffic. When *TNSM19* is used, resources are allocated under the unrealistic assumption that the traffic demand is uniformly distributed within a whole hour. For this reason, the lost traffic results higher than with the other approaches. With *PIMRC18*, up to 4% of traffic is lost. Even if it provides traffic predictions affected by lower *ARE* (0.29, 0.37, 0.42, 0.47, for the forecast at time *t*, *t+1*, *t+2* and *t+3*, respectively), it usually generates more underestimated traffic samples that contribute to QoS deterioration.

The comparison with the *15 min* case is also interesting. The *15 min* case is based only on traffic predictions performed over a time horizon of 15 minutes for which the error is lowest (see table 2). Nonetheless, in this case no processing of the ANN outputs is performed; hence, despite the small error in predictions, the lost traffic is quite large. This is a clear indication of the importance of the processing of ANN outputs.

Let us now focus on the proposed approaches. The lost traffic is lower in the *Max2Max* case than in the *I2I* one, since its switching condition is stricter. Fig. 6a shows the status of a micro cell BS of the Train Station area in orange and blue, when *I2I* and *Max2Max* are used, respectively, its traffic demand in grey and ρ with the dashed grey curve. Even if these two approaches use the same prediction samples for the resources allocation, *Max2Max* makes the BS active sooner than *I2I*. At 7.00 a.m., predictions of the traffic demand during the following hour are erroneously smaller that



(a)



(b)

Figure 6: Comparison of dynamic resources allocation strategies in the various areas: (a) *Max2Max* and *I2I* and (b) *Cont* version.

the threshold. Nevertheless, *Max2Max* switches BSs if the maximum, among the traffic demand samples belonging to that hour, is smaller than the ρ and can be carried by the macro supposing that it is managing an amount of traffic which is the maximum traffic demand among the 4 traffic demand samples of that hour. Thus, the micro BS is activated, since its traffic demand can not be carried by the macro BS because of the capacity constraint. The use of the *cont* variation provides benefits in terms of QoS in both strategies, *Max2Max* and *I2I*. When *cont* is used, the effect of higher errors, which characterises traffic predictions over longer time lags, is further mitigated, so that a more accurate resource allocation can be performed. As a result, the achieved lost traffic is always less than 2%. Specifically, in the areas where the resource allocation is more difficult due to the unpredictability of traffic demand, i.e., Rho Fiere and San Siro, 1.6% and 2% of the traffic is lost. In areas where patterns are more regular, values are always lower than 1.4%. Similar results are given by *I2I Flex*: the lost traffic is lower than the chosen benchmarks because the BS switching can react to the traffic demand every 15 minutes, provided that the last switching has occurred since at least 1 hour. This can be observed in Fig. 6b, where each curve corresponds to the status of a micro cell BS of the Duomo area, obtained with each of the considered allocation approaches. Also its traffic demand (in grey) and the ρ (grey dashed curve) are reported. The *cont* variation and *I2I Flex* react as soon as the traffic demand increases (at 8.30). When *I2I* and *Max2Max* are used, resources are allocated at 8.00 a.m. and the prediction with lag equals to 3 is used for that time slot. Because of the large error which affects this forecast, this sample results lower than the threshold and micro cell BS is not activated. Thus, from Figs. 6a and 6b, it is possible to notice that strategies behave similar if the traffic demand is far from the threshold. Indeed, in this case, the large error, which affects typically deeper forecasts used by *I2I* and *Max2Max*, does not impact the resources allocation, correctly detecting the value of the traffic demand with respect to the threshold. As soon as the traffic demand moves closer the threshold, even if based on the same predictions, the resources allocation is different. In case of *max*, conditions for the deactivation are stricter; with *cont* based approaches, more accurate predictions can be used. This results in more likely activation of micro BSs and, consequently, in lower lost traffic.

7.3 Impact of descending front detection

We now investigate the impact of the use of DFD in resource allocation. When DFD is used, the deactivation of a small cell BS is possible only if a descending front is detected, according to the conditions described in Sec. 5.1. In Fig. 7a, blue triangles mark the detection of a descending front during one day of the run-time phase of 2 small cell BSs, belonging to the PoliMi and Rho Fiere areas. As can be observed, descending fronts are mostly correctly identified. Since the current predicted traffic demand has lower impact on DFD than past samples, see equation (6), it is possible that DFD is activated after a local minimum.

Fig. 8 reports the energy consumption reduction, in bars, and the lost traffic, indicated by the blue and red lines with circle markers. Blue markers refer to no DFD, while red markers refer to DFD. The results of the chosen benchmarks are reported in grey. The figure reveals that the usage of DFD generates a systematic drop in both energy efficiency, for a small amount, and lost traffic, for more significant values. The energy consumption reduction remains between 10% and 39%, similar to the case of no DFD, when *TNSM19*, *15 min* and *PIMRC18* are used, but QoS improves significantly: lost traffic is usually below 1%. In the San Siro and Rho Fiere areas, because of the critical characteristics of traffic patterns, this value is between 1% and 1.5%. The reductions of lost traffic are due to the stricter conditions to switch off the small cell BSs. This can be seen in Fig. 7b, which illustrates an example of the traffic demand, in black, of the 2 small cells BSs of Fig. 7a. In Fig. 7b, the orange and blue points indicate the time slot during which the considered small cell BS is in sleep mode when *I2I cont* and *I2I cont with DFD* are used, respectively. During periods of almost constant but noisy traffic demand, if traffic values are close to the threshold ρ, incorrect small cell BSs deactivations may occur. Indeed, for those traffic values, even a small error in the traffic predictions can determine a wrong allocation of resources. This is the case reported in the figure: with DFD, incorrect deactivation of the considered small cell BSs is avoided since a descending front is not detected. Without DFD, with the *I2I cont* alone, the estimation error (even if small) makes the predictions lower than ρ, and a wrong switch off decision is taken. With DFD, the small cell BS is not switched off because the descending front is not detected. This behaviour explains the slight increase of energy consumption when DFD is applied. However, in spite of a very limited raise in energy consumption, the traffic loss can be reduced by up to 74% with respect to the benchmarks.

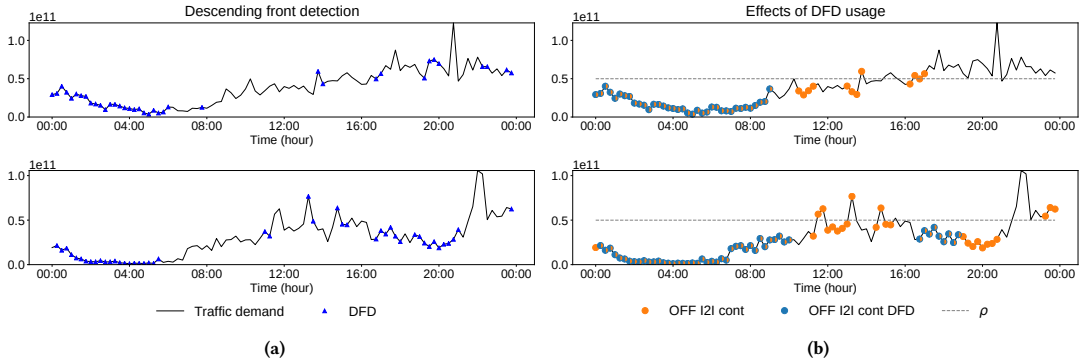


Figure 7: Impact of descending front detection for two areas: (a) detection of fronts and (b) switch off decisions with and without DFD.

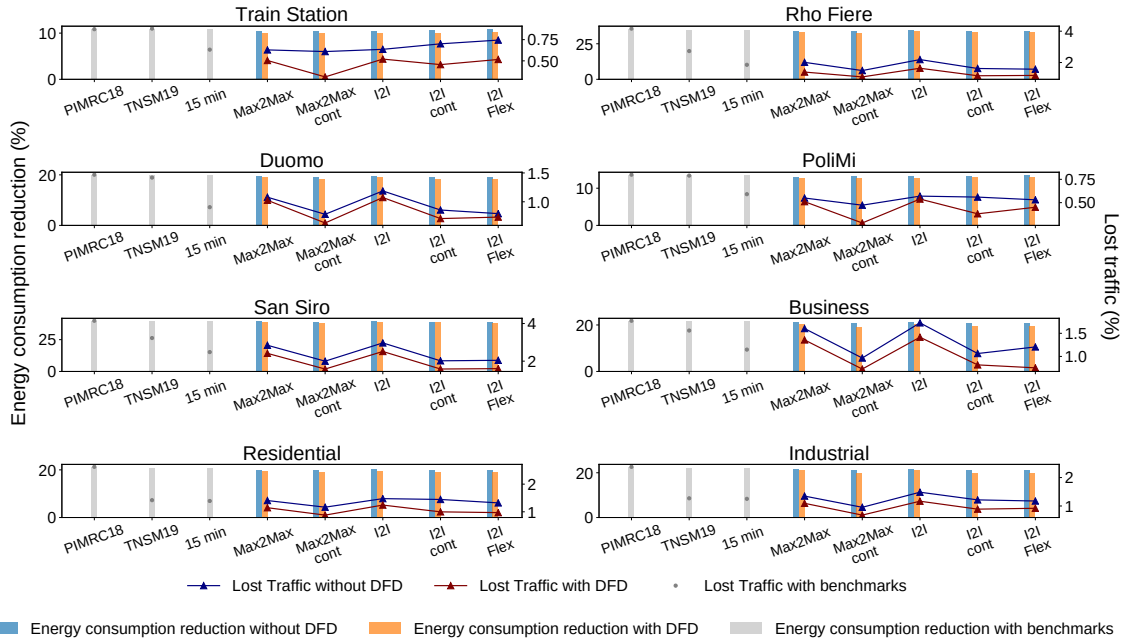


Figure 8: Energy consumption reduction and lost traffic in each area, with each dynamic resource allocation with and without descending front detection, DFD.

7.4 Impact of the traffic prediction technique

Finally, let us consider the impact of the traffic prediction technique. Fig. 9 reports with the blue and the orange bars, the energy consumption reduction achieved with and without DFD, if the traffic demand is forecast with *4 ANNs-1 output*, in San Siro and Business areas. The resulting lost traffic is shown with the red and blue lines with triangle markers, if the DFD is used or not, respectively. Similarly, the green and red bars in Fig. 9 indicate the energy consumption reduction obtained with and without DFD, when the traffic demand is forecast with *1 ANN-4 outputs*, which is the ANN that we identified as the one performing worst in predicting traffic. The obtained lost traffic is reported, respectively, with the red and blue lines with circle markers. In spite of the larger estimation error with respect to *4 ANNs-1 output* (see Table 2), performance is very similar: the values of lost traffic and energy consumption are almost equal to the previous case. Indeed, lost traffic drops up to 1%, while energy consumption is reduced between 9% and 40%. Similar results are achieved in the other areas. This means that the choice of an effective processing algorithm can have more impact on performance than the choice of the ANN. Only with a careful processing, the ANN prediction errors are mitigated, and a good trade-off between energy consumption reduction and QoS is achieved.

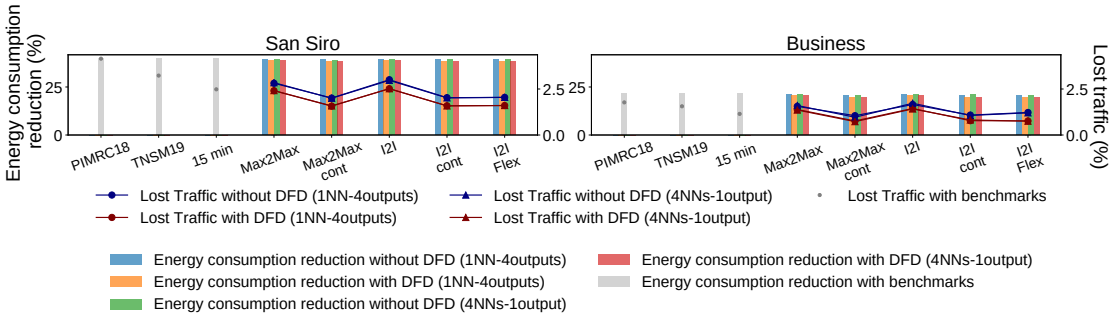


Figure 9: Energy consumption reduction and lost traffic in each area, with each dynamic resource allocation with and without descending front detection, DFD, in San Siro and Business areas, using *1 ANN-4 outputs* and *4 ANNs-1 output*.

8 CONCLUSIONS

In this paper, the traffic demands of BSs of a portion of a RAN are forecast with the objective of enabling BS management strategies that aim at reducing the RAN energy consumption. Results show that, in order to avoid QoS deterioration, the traffic predictions need to be carefully processed. Prediction processing requires both the understanding of traffic patterns over long time scales, so as to detect the overall trend of increasing or decreasing traffic, as well as strategies to combine predictions at different time lags.

A general lesson learnt in this work is that machine learning approaches are powerful enough to enable network management mechanisms that adapt to traffic variability. This is interesting in perspective, for the promising possibilities offered toward the deployment of new networks that are easily and automatically reconfigurable. However, machine learning approaches become particularly effective only if their outputs are integrated into decision processes that are driven by a deep domain knowledge, which cannot be eliminated if the desired objectives are to be achieved. In other words, artificial intelligence can be most effective only when driven by human wisdom.

REFERENCES

[1] Muhammad Ali, Michela Meo, and Daniela Renga. 2019. Cost saving and ancillary service provisioning in green Mobile Networks. In *The Internet of Things for Smart Urban Ecosystems*. Springer, 201–224.

[2] Gunther Auer, Oliver Blume, Vito Giannini, Istvan Godor, M Imran, Ylva Jading, Efstathios Katranaras, Magnus Olsson, Dario Sabella, Per Skillermark, et al. 2010. D2. 3: Energy efficiency analysis of the reference systems, areas of improvements and target breakdown. *Earth* 20, 10 (2010).

[3] Dario Bega, Marco Gramaglia, Marco Fiore, Albert Banchs, and Xavier Costa-Perez. 2019. DeepCog: Cognitive Network Management in Sliced 5G Networks with Deep Learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 280–288.

[4] Ł. Budzisz, F. Ganji, G. Rizzo, M. Ajmone Marsan, M. Meo, Y. Zhang, G. Koutitas, L. Tassiulas, S. Lambert, B. Lannoo, M. Pickavet, A. Conte, I. Haratcherev, and A. Wolisz. 2014. Dynamic Resource Provisioning for Energy Efficiency in Wireless Access Networks: A Survey and an Outlook. *IEEE Communications Surveys Tutorials* 16, 4 (Fourthquarter 2014), 2259–2285. https://doi.org/10.1109/COMST.2014.2329505

[5] Stefano Buzzi, I Chih-Lin, Thierry E Klein, H Vincent Poor, Chenyang Yang, and Alessio Zappone. 2016. A survey of energy-efficient techniques for 5G networks and challenges ahead. *IEEE Journal on Selected Areas in Communications* 34, 4 (2016), 697–709.

[6] Paulo Cortez, Miguel Rio, Miguel Rocha, and Pedro Sousa. 2006. Internet traffic forecasting using neural networks. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. IEEE, 2635–2642.

[7] Paulo Cortez, Miguel Rio, Miguel Rocha, and Pedro Sousa. 2012. Multi-scale Internet traffic forecasting using neural networks and time series methods. *Expert Systems* 29, 2 (2012), 143–155.

[8] Mattia Dalmasso, Michela Meo, and Daniela Renga. 2016. Radio resource management for improving energy self-sufficiency of green mobile networks. *ACM SIGMETRICS Performance Evaluation Review* 44, 2 (2016), 82–87.

[9] Hakim Ghazzai, Muhammad Junaid Farooq, Ahmad Alsharoa, Elias Yaacoub, Abdullah Kadri, and Mohamed-Slim Alouini. 2017. Green networking in cellular hetnets: A unified radio resource management framework with base station on/off switching. *IEEE Transactions on Vehicular Technology* 66, 7 (2017), 5879–5893.

[10] Jia Guo, Yu Peng, Xiyuan Peng, Qiang Chen, Jiang Yu, and Yufeng Dai. 2009. Traffic forecasting for mobile networks with multiplicative seasonal arima models. In *2009 9th International Conference on Electronic Measurement & Instruments*. IEEE, 3–377.

[11] S. Krishnasamy, P. T. Akhil, A. Arapostathis, S. Shakkottai, and R. Sundaresan. 2017. Augmenting max-weight with explicit learning for wireless scheduling with switching costs. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*. 1–9.

[12] Huimin Pan, Jingchu Liu, Sheng Zhou, and Zhisheng Niu. 2015. A block regression model for short-term mobile traffic forecasting. In *2015 IEEE/CIC International Conference on Communications in China (ICCC)*. IEEE, 1–5.

[13] Nadhir Ben Rached, Hakim Ghazzai, Abdullah Kadri, and Mohamed-Slim Alouini. 2018. A Time-Varied Probabilistic ON/OFF Switching Algorithm for Cellular Networks. *IEEE Communications Letters* 22, 3 (2018), 634–637.

[14] Daniela Renga, Hussein Al Haj Hassan, Michela Meo, and Loutfi Nuaymi. 2018. Energy management and base station on/off switching in green mobile networks for offering ancillary services. *IEEE Transactions on Green Communications and Networking* 2, 3 (2018), 868–880.

[15] M Zubair Shafiq, Lusheng Ji, Alex X Liu, and Jia Wang. 2011. Characterizing and modeling internet traffic dynamics of cellular devices. *ACM SIGMETRICS Performance Evaluation Review* 39, 1 (2011), 265–276.

[16] T Shankar et al. 2016. A survey on techniques related to base station sleeping in green communication and CoMP analysis. In *2016 IEEE International Conference on Engineering and Technology (ICETECH)*. IEEE, 1059–1067.

[17] Hoang Duy Trinh, Lorenza Giupponi, and Paolo Dini. 2018. Mobile traffic prediction from raw data using LSTM networks. In *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 1827–1832.

[18] Sebastian Troia, Rodolfo Alvizu, Youduo Zhou, Guido Maier, and Achille Pattavina. 2018. Deep Learning-based Traffic Prediction for Network Optimization. In *2018 20th International Conference on Transparent Optical Networks (ICTON)*. IEEE, 1–4.

[19] G. Vallero, D. Renga, M. Meo, and M. A. Marsan. 2019. Greener RAN Operation Through Machine Learning. *IEEE Transactions on Network and Service Management* 16, 3 (2019), 896–908.

[20] Shaojun Wang, Jia Guo, Qi Liu, and Xiyuan Peng. 2010. On-line traffic forecasting of mobile communication system. In *2010 First International Conference on Pervasive Computing, Signal Processing and Applications*. IEEE, 97–100.

[21] Leether Yao and Teng-Shih Tsai. 2016. Novel Hybrid Scheme of Solar Energy Forecasting for Home Energy Management System. In *2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. IEEE, 150–155.