A Data-driven Base Station Sleeping Strategy Based on Traffic Prediction

Jiansheng Lin, Youjia Chen, *Member, IEEE*, Haifeng Zheng, *Member, IEEE*
Ming Ding, *Senior Member, IEEE*, Peng Cheng, *Senior Member, IEEE*
Lajos Hanzo, *Fellow, IEEE*

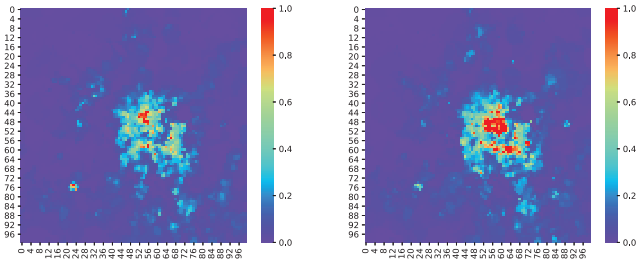
***Abstract*—Due to the rapidly increasing number of deployed base stations (BSs) in current cellular networks, energy consumption has emerged as a great challenge in network operation. In this paper, we propose a novel data-driven intelligent BS sleeping mechanism based on a wireless traffic prediction model that measures the BSs' capacity in different regions. Firstly, a spatial-temporal traffic prediction model is proposed, where a multi-graph convolutional network (MGCN) is developed to capture spatial features, and a multi-channel long short-term memory (LSTM) involving short-term, daily, and weekly periodic data is used to capture temporal features. Secondly, the capacities of macro-cell BSs (MBSs) and small-cell BSs (SBSs) with different environment characteristics are modeled, where both clustering and transfer learning algorithms are adopted to quantify the traffic supported by MBSs and SBSs. Finally, an optimal BS sleeping strategy is proposed to minimize the network power consumption. Experimental results show that the proposed MGCN-LSTM model outperforms the existing models in terms of traffic prediction, and the proposed BS sleeping strategy using an approximated non-linear model of capacity function achieves a near-optimal energy-saving performance with relatively low complexity.**

***Index Terms*—BS sleeping, traffic prediction, graph convolution network, transfer learning.**

I. Introduction

Network operators are deploying a large number of base stations (BSs) to meet the traffic requirements in peak hours. Hence, network densification has become an irresistible trend since 3G network. With the development of dense and ultra-dense small-cell networks, the excessive power consumption has become a major issue of network operation. For example, BSs account for two-thirds of the total energy consumption in a wireless network [1, 2]. Even if there is no or little traffic load, a BS can still consume more than 90% of their peak energy. The ultra-dense deployment of small-cell BSs (SBSs) leads to a low energy utilization in most time [3].

In this light, the BS sleeping strategy aiming at reducing the base operating power of BSs, becomes an attractive method to save energy consumption. Various strategies to enable dynamic on-off for BSs have been proposed in these years. In [4], the BS sleeping mode was designed according to a deterministic traffic changes over time. A similar idea was proposed in [5], where some BSs are shut down during the night time. All of these works utilize the periodic traffic fluctuation to save energy by switching some BSs to sleep mode. In Fig. 1, we plot the wireless traffic of Milan city at 10am and 10pm. From the figure, we can observe a significant traffic disparity between different regions for a given time instance. Meanwhile, comparing the traffic at 10am and that at 10pm, we can find that some regions tend to have traffic peaks in the day time, while some regions behave in a opposite way. Hence, in [6–9], the authors propose to selectively shut down BS resources in the network during off-peak hours based on the traffic monitor.



(a) 10pm. (b) 10am.

Fig. 1. The spatial distribution of cellular traffic at different time periods.

Taking a further step, it becomes desirable to develop BS sleeping strategies based on more specific and real-time traffic data. In [10], the authors design a deep Q-network (DQN) to obtain the effective BS sleeping policy by high-dimensional raw observations or un-quantized state vectors. It is proposed to enhance the original DQN algorithm with action-wise experience replay and adaptive reward scaling to deal with the challenges in non-stationary traffic. With the technique of deep learning, the traffic prediction enables the real-time BS sleeping strategy. Besides that, the accurate capacity model of BSs is another key factor of sleeping strategy. It is known that macro-cell BSs (MBSs) aim at providing an umbrella coverage while small-cell BSs (SBSs) target at throughput enhancement in traffic hotspots. The sleeping strategy should consider the different roles and capacities of BSs in the network operation. Moreover, the network environment greatly

Jiansheng Lin, Youjia Chen and Haifeng Zheng are with College of Physics and Information Engineering, Fuzhou University, P. R. China. E-mail: {N191120071,youjia.chen,zhenghf}@fzu.edu.cn. Ming Ding is with Data61, CSIRO, Australia. E-mail: Ming.Ding@data61.csiro.au. Peng Cheng is with La Trobe University, Australia. E-mail: p.cheng@latrobe.edu.au. Lajos Hanzo is with University of Southampton, UK. E-mail: lh@ecs.soton.ac.uk.

L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/P034284/1 and EP/P003990/1 (COALESCE) as well as of the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028)

affects the transmission of wireless signal. For instance, city CBD is full of tall buildings while the rural area is usually scattered with low-density houses. Therefore, an accurate capacity model of different types of BSs considering various region characteristics should be constructed.

The main contributions of this paper are three-fold:

- We propose a novel cellular traffic prediction method, where a multi-graph convolutional network (MGCN) is used to extract the spatial features, and a multi-channel long short-term memory (LSTM) network is adopted to efficiently capture the major frequency components in the traffic data. In addition, an attention scheme is designed to distinguish the significance of traffic sequences at different time.
- The capacity models of MBSs and SBSs are established where the performance impacts of environment characteristics are considered. Since the original dataset does not differentiate MBSs and SBSs, a clustering algorithm is utilized to distinguish the areas deployed with MBSs only, and transfer learning is then adopted to identify the portions of traffic supported by MBSs and SBSs in an area.
- Based on the predicted traffic demand and the learned MBS and SBS capacity models, the optimal numbers of active MBSs and SBSs are obtained to minimize the power consumption in a given area. Due to the intractable expression of the capability functions obtained by the neural network, we have used both linear and non-linear fitting models to derive the optimal strategies without the high complexity of exhaustive searching.

II. RELATED WORK

In this section, we investigate the related works on traffic prediction and BS sleeping control.

A. Traffic Prediction

Many works on traffic forecasting are based on neural networks. In [11], bilinear recurrent neural networks were designed in the aspects of training and structure for the real-time Ethernet traffic prediction. In [12], the differential evolution algorithm was combined with the back propagation algorithm to optimize the fuzzy neural network forecasting network traffic. A deep belief network (DBN) with Gaussian models was proposed in [13] to forecast the traffic demand in wireless mesh networks. In [14], the multitask learning (ML) was incorporated with DBN, where the DBN is employed for feature learning and the multitask regression is used at the top for supervised prediction. These models mainly focused on capturing temporal features of the training data, while not utilizing the potential features in other domains.

In contrast, autoencoders were utilized to extract the spatial features in [15], then LSTM was adopted after to perform the traffic prediction.In [16–18], convolution neural network (CNN) and LSTM were combined to construct a spatial-temporal neural network for traffic prediction, where the CNN concentrates on the spatial features. In [19], the attention mechanism was introduced into the Conv-LSTM network, which automatically exploited the different importance of the traffic data at different time periods. Moreover, in [20] the authors proposed a spatial-temporal cross-domain neural network, where multi-domain data were integrated to obtain the complex patterns hidden in wireless cellular traffic. However, cellular traffic are time-series data distributed over the wireless network, whose structure is non-Euclidean. Thus, spatial features learned in CNN are not optimal for representing the traffic network structure.

Graph convolutional network (GCN) has also been widely used in traffic prediction, which generalizes the convolution operation to non-Euclidean domains based on the spectral graph theory [21]. In [22], T-GCN was proposed, where a GCN was utilized to learn the traffic network topology. In [23], the authors adopted the GCN in the spatial domain and the gated CNN in the temporal domain. MGCN has been used in road traffic prediction recently [24, 25]. In [24], the connectivity between regions, i.e., the road information, constructs an important graph, together with the distance between regions and the region functionalities. In [25], three graphs were involved: the distance between any two bike stations, the number of rides between them, and correlation of ride record between them. In contrast to the road traffic, where the road connectivity and the start-end stations are key factors, for cellular network traffic, the geographical environment and the city functionality are more important information. Although MGCN can explore more pair-wise relationships in different domains, it may introduce higher computational complexity and some extra noise due to overestimated relationships.

B. Base Station Sleeping Control

Various approaches have been proposed to reduce the energy consumption of a mobile cellular network, which can generally be divided into the two categories: *1)* improving the energy efficiency of the network components, and *2)* turning off some of the network resources selectively. The BS sleeping strategy is a typical approach in the second category. Deactivating some deployed BSs during off-peak hours, not only the energy consumption but also the inter-cell interference in the network can be reduced.

In [26], the state of a BS was adjusted according to its cell load. An N-policy was proposed in [27] to achieve the best energy-delay trade-off, where a BS remains in the sleep mode until N users are accumulated in its coverage. Such kind of methods controlled the BS mode based on the real-time detection of the traffic load or user requirements. In the scenarios of dense networks, the real-time detection in each BS and frequent hand-over operation among BSs results too much cost in the control signalling.

Furthermore, the rapid development of big data technology and deep learning offers more possible approaches for network control. However, the works on data-driven intelligent BS sleeping control are still rare. In [28], a self-organizing ultra-dense small cell network was proposed, where the necessary active SBSs are determined by traffic loads, and high-rank SBSs transmit at the maximum power while the transmission power of low-rank SBSs will be adjusted. In [29], the authors

partitioned the network into multiple communities, and a heuristic switch-off strategy was adopted independently in each community to maximize its energy saving while guaranteeing the minimal service requirement. In this work, the traffic requirement is the main concern while the characteristics of different communities are not considered, which are highly related with the network performance as mentioned in [20].

III. Problem Statement and the Proposed Framework

In this section, we formulate the BS sleeping optimization problem and propose our solution.

A. Problem Statement

As mentioned before, in this paper we focus on a data-driven strategy of BS-sleeping control to save energy consumption, which relies on the traffic prediction and the capacity modeling. The current cellular networks consist of two kinds of BSs: 1) MBSs with a high transmission power, providing an umbrella wide coverage and support user mobility. 2) SBSs with a relative low transmission power, enhancing network capacity in traffic hot-spots. Due to these different properties, different power and capacity models should be adopted for MBSs and SBSs. Moreover, for an area, its physical environment and its regional function play important roles in the wireless traffic patterns [30]. Hence, in this work, the area characteristics are also considered.

Given an area with a characteristic vector $\mathbf{r}$, and a traffic requirement μ, the investigated BS-sleeping control problem can be formulated as

$$\min_{n_m, n_s} \; n_m P_m + n_s P_s, \tag{1a}$$

$$s.t. \quad \mathcal{C}_m(\mathbf{r}, n_m) + \mathcal{C}_s(\mathbf{r}, n_s) \geq \mu + \Delta, \tag{1b}$$

$$\mathcal{A}(\boldsymbol{r}) \leq n_m \leq N_m, \tag{1c}$$

$$0 \leq n_s \leq N_s, \tag{1d}$$

$$n_m, n_s \in \mathbb{N}, \tag{1e}$$

where n_m, n_s represent the numbers of active MBSs and SBSs in this area, respectively. N_m and N_s denote the numbers of deployed MBSs and SBSs, i.e., the maximum MBSs and SBSs can be activated. The power consumptions of an active MBS and SBS are denoted by P_m and P_s, respectively. In addition, to satisfy the basic network coverage of this area with the characteristic vector $\boldsymbol{r}$, a minimum number of active MBSs are needed, denoted by $\mathcal{A}(\boldsymbol{r})$. Moreover, $\mathcal{C}_m(\cdot)$ and $\mathcal{C}_s(\cdot)$ denote the capacities that MBSs and SBSs could provide in this area, and Δ denotes the traffic surplus that can be provided by the network.

In more detail, Eq. (1a) means that the objective is to minimize the power consumption by adjusting the number of active MBSs and SBSs. Eq. (1b) indicates that the active BSs should provide enough capacity to support the traffic requirement with a desirable surplus. Eq. (1c) and (1d) regulate that the activated number of MBSs and SBSs are upper bounded by the overall number of deployed BSs, and a specific number of MBSs are needed to ensure the basic network coverage.

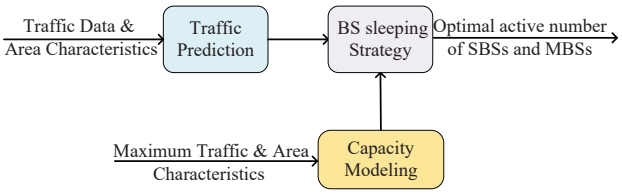


Fig. 2. Three components of the proposed data-driven BS sleeping control.

B. Framework

To solve this optimization problem, we propose a big-data-driven framework, which consists of three components, as shown in Fig. 2. The first component realizes the traffic prediction function to quantify the real-time traffic requirement, i.e., μ in Eq. (1b). The second component aims at modeling the capacities of MBSs and SBSs in different areas, i.e., the capacity functions $\mathcal{C}_m(\cdot)$ and $\mathcal{C}_s(\cdot)$. The third component implements the optimal BS sleeping strategy. The detail steps of the three components are described as follows:

- **Traffic Prediction:** In this module, the historical cellular traffic data is used to perform the traffic prediction for the current time, together with multi-domain data which are highly related to the wireless traffic, such as the number of restaurants, schools etc., named POIs. Based on the analysis of the time-domain and spatial-domain characteristics, a spatial-temporal hybrid deep learning framework is proposed to perform the traffic prediction, which includes an MGCN network, a multi-channel LSTM network, an attention mechanism and a fully-connected layer.
- **Capacity Modeling:** In this module, the maximum traffic that a number of MBSs and SBSs can provide in different area environments are analyzed. Firstly, a clustering algorithm is adopted to classify an area into one of the interested regions. Then the data of the areas classified as rural regions with MBS-only deployments are utilized as the source domain, and transfer learning is used to identify the numbers of MBSs and SBSs in the rest of the areas. After that, based on the maximum traffic and deployed number of MBSs and SBSs, a multi-layer perceptron (MLP) neural network is used to construct the capacity models.
- **BS Sleeping Strategy:** In this module, the optimization problem is addressed based on the predicted traffic demand and the model-based BS capacities. The optimal numbers of active MBSs and SBSs are derived to minimize the power consumption, while a number of MBSs are firstly activated to ensure the basic network coverage requirement, and then a number of MBSs and SBSs are selected to meet the traffic requirement.

C. Cellular Network Datasets

1) Cellular Traffic Dataset: The cellular traffic involved in this paper was provided by a large telephone service provider in European, as a part of the "Big Data Challenge" [31]. The raw dataset was collected from 11/01/2013 to 01/01/2014 with a temporal interval of 10 minutes over the entire city of Milan.

TABLE I
DATA OF WIRELESS TRAFFIC IN MILAN CITY.

Square ID	Country code	Time-stamp	SMS-in	SMS-out	Call-in	Call-out	Internet
1	39	1383260400000	0.14186	0.15679	0.16094	0.05227	11.02837
1	0	1383261000000	0.13658	*	0.02730	*	*
...	...	...	...	...	...	...	...
6359	39	1383306000000	7.18483	3.21435	6.72216	4.98282	142.80549
...	...	...	...	...	...	...	...
10000	39	1383432600000	*	0.638219	0.145000	*	24.14275

The city of Milan is divided into 100×100 square grids, and the size of each square is about $0.235 \times 0.235\text{km}^2$. In each square, three kinds of cellular traffic are recorded: short message service (SMS), call service (Call) and Internet service. The original dataset consists of Square ID, Time stamp, SMS-in, SMS-out, Call-in, Call-out and Internet. The details of the records in this dataset. In this work, the time interval is adjusted to 1 hour for traffic aggregation.

The cellular traffic can be represented by a spatial-temporal sequence of data elements, $X_{(i,j)}^{(s,t)}$, which denotes the cellular traffic at t-th interval of the square grid (i, j) for the service type, $s \in \{\text{SMS, Call, Internet}\}$. Omitting s, $\boldsymbol{X}^t$ denotes the traffic of one service type at the time interval t in the entire city of Milan given by:

$$\boldsymbol{X}^t = \begin{bmatrix} X_{(1,1)}^t & X_{(1,2)}^t & \cdots & X_{(1,J)}^t \\ X_{(2,1)}^t & X_{(2,2)}^t & \cdots & X_{(2,J)}^t \\ \vdots & \vdots & \ddots & \vdots \\ X_{(I,1)}^t & X_{(I,2)}^t & \cdots & X_{(I,J)}^t \end{bmatrix}. \tag{2}$$

Hence, $\boldsymbol{X} = \{\boldsymbol{X}^t | t = 1, 2, 3, \ldots, T\} \in \mathbb{R}^{I \times J \times T}$ represents all the traffic data recorded. Note that, the traffic in the datasets is measured with a unit: number of events, which is proportional to the number of call detail records (CDRs). Moreover, a CDR is a record created when a connection lasts for more than 15 mins or generates more than 5 MB traffic [31].

2) Other Related Datasets: As mentioned before, the wireless traffic and the deployment of MBSs and SBSs in an area highly depend on the the spatial/region information, the propagation environment and the population distribution. Hence, to analyze the characteristics of an area, several related datasets are introduced in this work: POIs, social activities, and the number of BSs deployed [20]. All these dataset are transformed to represent the 10000 square grids like the cellular traffic dataset.

- **POI**: In this dataset, the numbers of 12 kinds of POIs [1] in each square grid are collected, including banks, bars, etc., which are listed in Table II. Similar to the traffic data $\boldsymbol{X}$, a matrix $\boldsymbol{P}$ is defined to represent the POI data in the entire Milan city, where the element $\boldsymbol{p}_{(i,j)} = \left[p^{\text{Bank}}, \cdots, p^{\text{Lodging}}\right]$ denotes the numbers of POIs in the square grid (i, j).
- **BS number**: Meanwhile, we have the number of deployed BSs [2] in each square grid. Denoted by a matrix $\boldsymbol{B}$ the BS number deployed in the entire city, the element $b_{(i,j)}$ represents the BS number in the grid (i, j).
- **Social activity**: Social activity [3] is a measure of how active users are and how dependent they are on network services. For the Milan city, the data about Twitter usage from 11/01/2013 to 01/01/2014 is collected [20]. Similarly, dividing them into 10000 square grid, a matrix $\boldsymbol{S}$ is defined where its element $s_{(i,j)}$ denotes the number of social activity generated in the square grid (i, j).

[1]POIs information be crawled using Google Places API. Available: https://developers.google.com/maps

[2]BSs information is obtained from OpenCellID. Available: https://opencellid.org/

[3]Social activity is collected through Dandelion API. Available: https://dandelion.eu

IV. TRAFFIC PREDICTION BY MGCN-LSTM

In this section, a deep learning-based prediction model is proposed, which consists of an MGCN, an LSTM network, and an attention mechanism, referred to as MGCN-LSTM.

As shown in Fig. 3, in the spatial domain, we use the correlation between square grids to construct three graphs, where each vertex denotes a square grid and the edge between them represents their correlations. Then GCN is used to capture the spatial characteristics in each graph. In the time domain, the LSTM network is used to extract the temporal features in short term, daily period and weekly period. Then the attention mechanism exploits different levels of importance for the feature extracted from at different time intervals.

A. Spatial Features and Multi-Graph Convolution Network

The undirected graph $G = (V, \boldsymbol{A})$ is constructed to presents the correlation between different square grids, where V denotes the set of nodes and $\boldsymbol{A} \in \mathbb{R}^{|V| \times |V|}$ denotes the adjacency matrix between the nodes.

1) Multi-Graph Construction: Different graphs are constructed to encode the different types of correlation among nodes: 1) Neighborhood graph, $G_N = (V, \boldsymbol{A}^N)$, represents the spatial distance between the square grids. 2) Functional similarity, $G_F = (V, \boldsymbol{A}^F)$, denotes the similarity of regional functions in each square grid. 3) Spatial traffic correlation, $G_S = (V, \boldsymbol{A}^S)$, represents the spatial correlation of traffic flow among the square grids.

- **Neighborhood graph,** $G_N = (V, \boldsymbol{A}^N)$: During a period of time, the user's activities are physically limited in a

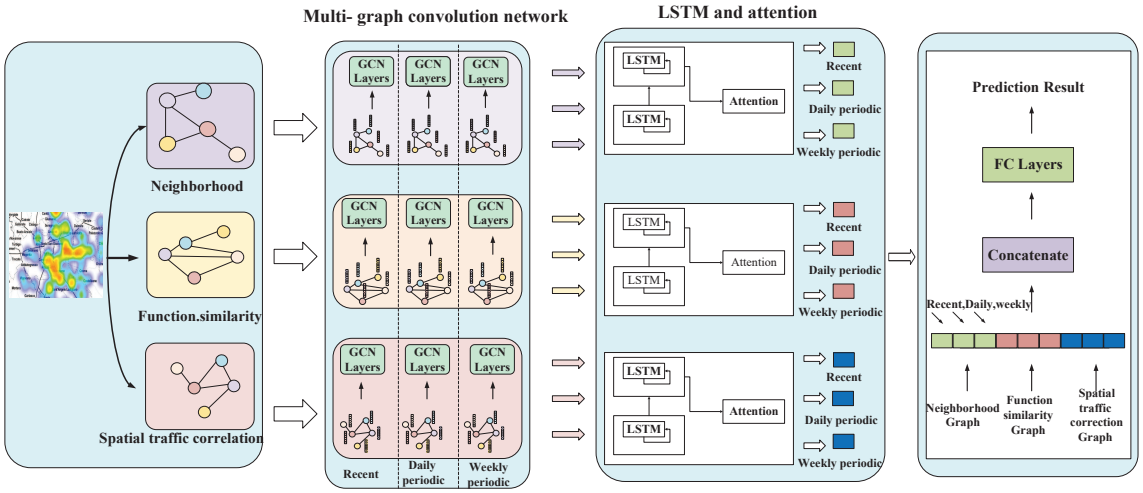


Fig. 3. An illustration of the proposed MGCN-LSTM network.

TABLE II
DATASETS OF SOCIAL ACTIVITY, POI AND BS NUMBER.

Square ID	Social activity	BS number	Point of interest											
			Bank	Bar	Cafe	Church	Park	Parking	Restaurant	School	Store	Subway	Library	Lodging
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5970	122	24	3	1	1	0	0	1	2	0	13	100	0	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10000	0	1	0	0	0	0	0	0	0	0	0	0	0	0

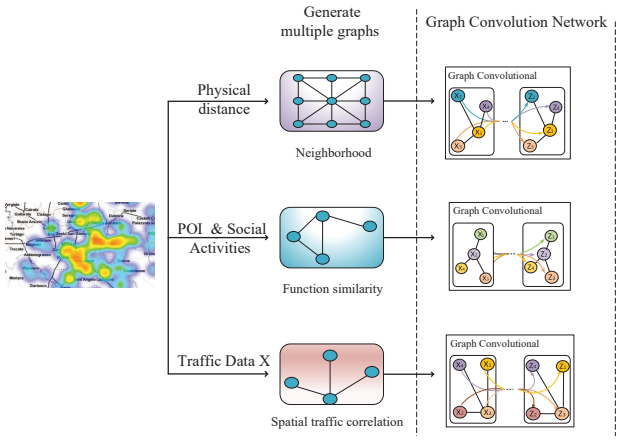


Fig. 4. An illustration of multi-graph convolution networks adopted in the spatial domain.

certain range. The neighborhood graph aggregates information from adjacent nodes and quantifies their correlationship through physical distance. Intuitively, the closer distance of two nodes, the higher correlation between them. Hence, the element in adjacency matrix, $A^N_{h,w}$, is defined using the Euclidean distance between nodes, that is

$$A^N_{h,w} = \exp\left(-\frac{||v_h - v_w||^2}{2\sigma^2}\right), \forall h, w, \tag{3}$$

where v_h denotes the location of node h and $||v_h - v_w||$ representing the Euclidean distance between node h and w[4], σ is a free parameter. Note that, in the used Milan dataset, the distance between two nodes (i.e., square grids) can be calculated from the distance between their grid centers.

- **Functional similarity graph,** $G_F = (V, \boldsymbol{A}^F)$: Square grids with similar regional functions, for instance, residential area, office area and transportation area, have commonality in their traffic pattern [32]. Here we use the POI matrix $\boldsymbol{P}$, social activity matrix, $\boldsymbol{S}$, and BS number matrix, $\boldsymbol{B}$, to represent the area characteristic vector $\boldsymbol{r}_{i,j} = [p_{i,j}, s_{i,j}, b_{i,j}]$. Cosine similarity function is introduced to measure the functional similarity, that is, the element in the adjacency matrix is given by

$$A^F_{h,w} = \frac{\langle \boldsymbol{r}_h, \boldsymbol{r}_w \rangle}{||\boldsymbol{r}_h|| \cdot ||\boldsymbol{r}_w||}, \tag{4}$$

where $\langle \cdot \rangle$ represents inner product operation and $|| \cdot ||$ represents 2-norm.

- **Spatial traffic correlation graph,** $G_S = (V, \boldsymbol{A}^{\boldsymbol{S}})$: The Pearson coefficient is adopted to analyze the spatial correlation between the wireless traffics in different square grids [20]. Hence, this coefficient is used to define the adjacent matrix as

$$A^S_{h,w} = \frac{cov(\boldsymbol{X}_h, \boldsymbol{X}_w)}{\sigma_{\boldsymbol{X}_h} \sigma_{\boldsymbol{X}_w}}, \tag{5}$$

[4]A node w of graph can be expressed as $w = 100 \times i + j$, where (i, j) is the coordinate of square, this conclusion also applies to other nodes.

where $cov(\cdot)$ is the covariance operator, and σ is the standard deviation, $\boldsymbol{X}_h$ and $\boldsymbol{X}_w$ denote the cellular traffic of node h and w, respectively.

Denoted by $\boldsymbol{A}$ the adjacent matrix for each graph constructed before, with the traffic data $\boldsymbol{X}$ as the other input parameter, a graph convolutional network is utilized to capture the spatial characteristics. The detail process is explicated as follows.

2) Graph Convolutional Network : With the adjacent matrix $\boldsymbol{A} \in \mathbb{R}^{V\times V}$ of a constructed graph, a degree matrix $\mathbf{D}$ can be calculated with elements $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$. Since $\boldsymbol{A}$ is a symmetric matrix, $\mathbf{D}$ is a diagonal matrix. Hence, the normalized Laplacian matrix can be obtained by $\boldsymbol{L} = \boldsymbol{I}_V - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, where $\mathbf{I}_V$ denotes the identity matrix. With eigenvalue decomposition $\boldsymbol{L} = \boldsymbol{U}\Lambda\boldsymbol{U}^{\boldsymbol{T}}$, we can obtain the eigenvector matrix $\boldsymbol{U}$ and a diagonal matrix $\boldsymbol{\Lambda}$.

Since $\boldsymbol{L}$ is also a symmetric matrix with V linearly independent eigenvectors, the corresponding eigenvector matrix $\boldsymbol{U}$ is an orthogonal matrix, which can be used as a basis set of Fourier transform. Hence, with $\boldsymbol{U}$, we perform Fourier transform on the input feature $\boldsymbol{X}$: $\hat{\boldsymbol{X}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}$, where $(\cdot)^{\mathrm{T}}$ denotes the transpose.

Following Eq. (4) in [21], a filter $g_\theta(\boldsymbol{\Lambda})$ can be constructed by the Chebyshev polynomial $T_k(\cdot)$ [33]: $g_\theta(\boldsymbol{\Lambda}) = \sum_{k=0}^{K-1}\theta_k T_k(\tilde{\boldsymbol{\Lambda}})$, where $\tilde{\boldsymbol{\Lambda}} = \frac{2}{\lambda_{\max}}(\boldsymbol{\Lambda}) - \boldsymbol{I}_V$, and $\lambda_{\max} = \max\{\boldsymbol{\Lambda}_{ii}\}$ is the largest eigenvalue. Then, a graph convolution operation, $*_{\mathcal{G}}$, is performed between the constructed filter $g_\theta(\boldsymbol{\Lambda})$ and input feature $\boldsymbol{X}$. That is,

$$g_\theta(\boldsymbol{\Lambda}) *_{\mathcal{G}} \boldsymbol{X} = \boldsymbol{U} g_\theta(\boldsymbol{\Lambda})\boldsymbol{U}^T\boldsymbol{X}.$$

B. Time Domain Properties and Multi-channel LSTM

1) Characteristics of Traffic Data: Based on Parseval theorem, the major frequency components play the key roles in signal recovering. In order to analyze the periodic characteristics in wireless traffic data, the discrete fourier transform (DFT) is used to investigate its frequency components, that is, $\hat{X}[k] = \sum_{n=1}^{N} x[n]\exp^{-2\pi ikn/N}$, where N is the number of traffic cellular samples.

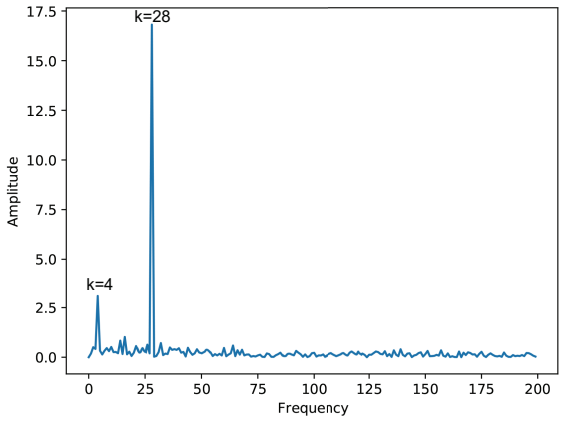


Fig. 5. Frequency Analysis of Traffic Data.

Fig. 5 shows DFT results of square grid $(10, 0)$, from which it can be seen that there are two main frequency components: $k = 4$ and $k = 28$. Since the data duration is 4 weeks, the frequency component related with k can be derived from the formula $\frac{4\text{weeks}}{k}$: 1) The first frequency component, $k = 4$, represents **1** week; 2) the second frequency component, $k = 28$, means **1** day. A similar conclusion about cellular traffic period was obtained in [32].

2) Multi-channel Data: To aggregate these periodic characteristic, multi-channel traffic data are used as inputs in time-domain prediction. That is, to predict the traffic from $\boldsymbol{X}_{t+1}$ to $\boldsymbol{X}_{t+T_p}$, three traffic sequences are fed into the prediction model as training data:

- **Recent segment**: Due to the continuity of the cellular traffic, the traffic data in previous T_r time intervals acts as an important role in prediction. Hence, we define a recent segment, $\boldsymbol{\mathcal{X}}_r \in \mathbb{R}^{I\times J\times T_r}$, with a length T_r, as the first input:

$$\boldsymbol{\mathcal{X}}_r \triangleq (\boldsymbol{X}_{(t-T_r+1)}, \boldsymbol{X}_{(t-T_r+2)}, \ldots, \boldsymbol{X}_t). \tag{6}$$

- **Daily periodic segment**: As analyzed, the wireless traffic shows an apparent daily period, hence the traffic data at 24 hours before, 48 hours before, and so on, are used as the second input:

$$\begin{aligned}\boldsymbol{\mathcal{X}}_d \triangleq \big(&\boldsymbol{X}_{t-24T_d+1}, \cdots, \boldsymbol{X}_{t-24T_d+T_p},\\ &\boldsymbol{X}_{t-24(T_d-1)+1}, \cdots, \boldsymbol{X}_{t-24(T_d-1)+T_p},\\ &\cdots \boldsymbol{X}_{t-24+1}, \cdots, \boldsymbol{X}_{t-24+T_p}\big),\end{aligned} \tag{7}$$

  where T_d is the number of days taking into consideration.

- **Weekly periodic segment**: Similarly, the weekly periodic segment is adopted as the third input:

$$\begin{aligned}\boldsymbol{\mathcal{X}}_w \triangleq \big(&\boldsymbol{X}_{t-24*7*T_w+1}, \cdots, \boldsymbol{X}_{t-24*7*T_w+T_p},\\ &\boldsymbol{X}_{t-24*7*(T_w-1)+1}, \cdots, \boldsymbol{X}_{t-24*7*(T_w-1)+T_p},\\ &\cdots \boldsymbol{X}_{t-24*7+1}, \cdots, \boldsymbol{X}_{t-24*7+T_p}\big),\end{aligned} \tag{8}$$

  where T_w is the number of weeks taking into consideration.

The proposed multi-channel LSTM is illustrated in Fig. 6. In each channel, we adopt two-layer LSTM, because such temporally concatenated model always achieves a good performance in sequence modeling problems. Since LSTM is widely used model, we omit its introduction here and only briefly introduce attention network in the following.

3) Attention Mechanism: As an optimization model algorithm, the attention mechanism has been proposed to explore the inherent features of data and improve the efficiency of information processing. The main feature lies in its ability to distinguish the importance of different data. In our prediction model, the information provided by the data flow at different times may be not equally important for prediction performance. However, the standard LSTM cannot detect which is the important part for a traffic data sequence. Hence, we embed the model of attention mechanism into the LSTM system, which allows the model to have direct dependence between the states at different times. At each time slot t, given the traffic

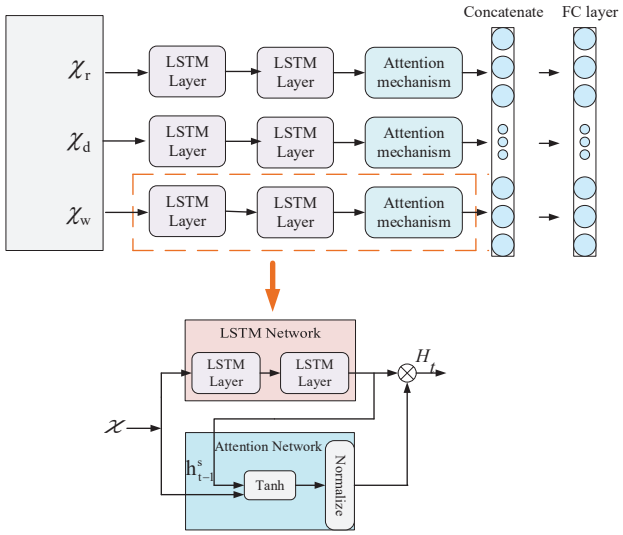


Fig. 6. An illustration of the LSTM network and attention scheme adopted in the time domain.

dataset $\mathcal{X}$, the scores $\boldsymbol{s}_t = (s_1, s_2, \cdots, s_t)$ indicate the importance of the traffic flow sequence at different time slots, which can be obtained as $\boldsymbol{s}_t = U_s \tanh(\boldsymbol{W}_{xs}\mathcal{X} + \boldsymbol{W}_{hs}\boldsymbol{h}^{\boldsymbol{s}}_{\boldsymbol{t-1}} + \boldsymbol{b}_{\boldsymbol{s}})$, where $U_s, \boldsymbol{W}_{xs}, \boldsymbol{W}_{hs}$ are the learnable parameters, $\boldsymbol{b}_{\boldsymbol{s}}$ is the bias parameter and $\boldsymbol{h}^{\boldsymbol{s}}_{\boldsymbol{t-1}}$ is the hidden output from the LSTM network. The weights at different moments, named attention weights, can be expressed as $\alpha_k = \frac{\exp(s_k)}{\sum_{k=1}^{t}\exp(s_k)}$, which normalize the scores. The larger the weight, the more important it is. At each time step t, the output of the LSTM hidden state $\boldsymbol{h}^{\boldsymbol{s}}_{\boldsymbol{t}}$, attention mechanisms can be calculated a vector H_t as a weighted summation $H_t = \sum_{k=1}^{t}\alpha_k h^s_k$. Note that the attention weight α depends on the input $\mathcal{X}$ and the hidden variables $\boldsymbol{h}^{\boldsymbol{s}}_{\boldsymbol{t-1}}$, so it depends not only on the present moment but also on the previous moment. The attention weight α can be regarded as the activation of flow selection gate. The set of gates control the amount of information from each flow to enter the LSTM network. The larger the attention weight, the greater the impact on the prediction results.

V. Capacity Modeling of MBSs and SBSs

In this section, MBSs' and SBSs' capacity functions, $\mathcal{C}_m(\cdot)$ and $\mathcal{C}_s(\cdot)$, are modeled, to reveal the maximum traffic volume (i.e., throughput) that a number of MBSs or SBSs can provide in an area with a characteristic vector $\boldsymbol{r}$. Different from the theoretical capacity function, such as Shannon Theorem, in this section, the neural network is utilized to construct the capacity model.

A. Capacity Impacts of Environment Characteristics

For each square grid, the peak of cellular traffic during a period of 3 months is used to model the capacity of the deployed BSs. This approximation is reasonable, since operators often deploy a specific number of BSs based on peak cellular traffic. The maximum traffic for a square grid (i, j) can be calculated as $z_{(i,j)} = \max_t\{X^{(sms,t)}_{(i,j)} + X^{(call,t)}_{(i,j)} + X^{(internet,t)}_{(i,j)}\}$, $t \in (0, T)$.

As we know, the maximum traffic of an area highly depends on the characteristic of this area, e.g., its functionality and the population distributed. Hence, we investigate the relationship between the maximum wireless traffic and the area characteristic vector $\boldsymbol{r}$. In Fig. 7 we plot the network capacity, deployed BS number, and the total POIs number in each square grid. For clarity, we make the x-axis represents the 1-10000 square grids. From the figure, the three curves show an obvious correlation among them.

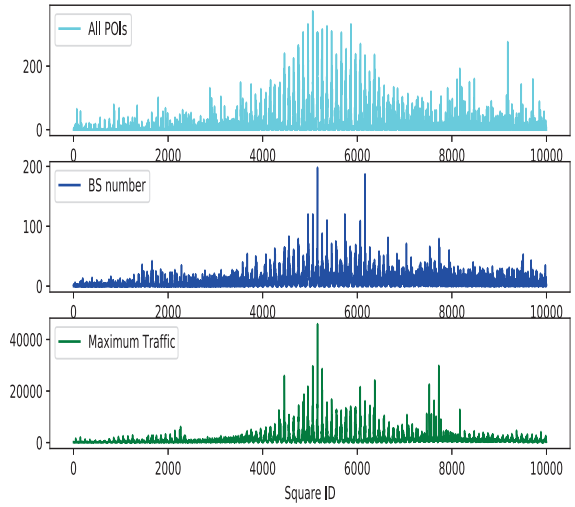


Fig. 7. The number of all kind of POIs, the number of deployed BSs and the maximum traffic in each square grid.

To better quantify the correlation among them, the cross correlation is calculated and plotted, i.e., $R_{km}(\tau) = \frac{\sum_{l=0}^{\mathcal{N}} k(l)m(l+\tau)}{\sigma_{k(l)}\sigma_{m(l)}}$, where the maximum traffic sequence is substituted into $k(l)$ and the number of overall POIs or the number of BSs is substituted into $m(l)$. Note that, around the Square ID 9100 in Fig. 7, there is a spike on the POI number but not on maximum traffic. From map [34], we can see this area is likely to be a mixture of agricultural and industrial areas, where a few main roads and railways have been constructed. Hence, although they are marked with a large number of subway stations, the number of mobile users may be not large. From Fig. 8, we can see that the obvious peak in $\tau = 0$, which implies that in each square grid the maximum traffic highly depends on the number of overall POIs and the BS number, i.e., the environment vector $\boldsymbol{r}$. However, when $\tau \neq 0$ the correlation coefficient rapidly decreases. This phenomena tell us that due to the difference of environment characteristics in different area (i.e., square grids), the deployed BS number, the POIs and the traffic are all different. This difference becomes more significant with the increase of their physical distance.

With these results, we can clear see the impact of the environment characteristics and the number of BSs on wireless maximum traffic, which verifies our modeling of the capacity functions $\mathcal{C}_m(n_m, \boldsymbol{r})$ and $\mathcal{C}_s(n_s, \boldsymbol{r})$. That is, the capacity is highly dependent on $\boldsymbol{r}$. Since the number of BSs deployed in a square grid determines the provided wireless resource,

we choose it as one parameter of the capacity function [35]. Meanwhile, the geographic environment greatly impacts the propagation model of wireless signals [36]. Hence, the area characteristic vector is chosen as another parameter of the capacity function.

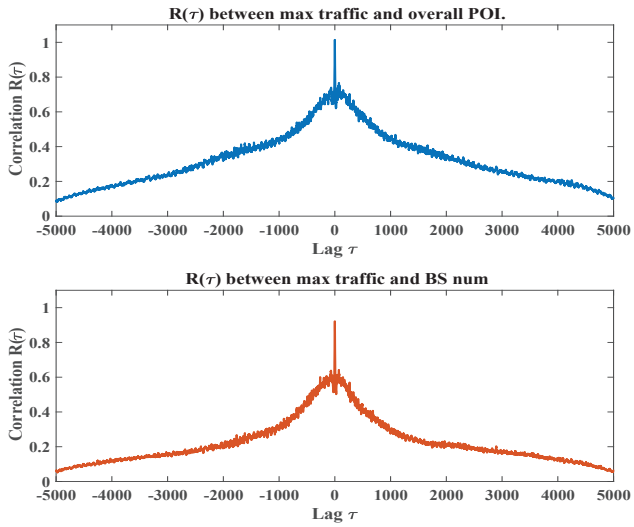


Fig. 8. Lag correlations between maximum traffic and overall POI, BS number.

B. Clustering of Different Regions

From Fig. 1 and 7, we can also see that the wireless traffic is quite different in different geographic locations. The areas close to the city CBD area have a large population, which contributes to a large wireless traffic distributed over a large number of POIs and deployed BSs. In contrast, the sparse population in remote areas leads to a much less traffic volumn and hence a small number of deployed BSs. Considering these differences, a clustering algorithm is used to classify the 10000 square grids into different regions.

In more detail, the number of BSs, the number of POIs, the maximum traffic are collectively used to construct the feature vector for each square grid, then a K-means algorithm is utilized to cluster the whole Milan area into different regions. An unsupervised K-means algorithm is used to implement the clustering, which aims at minimizing the distance within the same cluster and maximize the distance within different clusters.

Fig. 9 shows the result of the adopted clustering algorithm, where the square grids with similar characteristics are grouped into a cluster. Following the typical scenarios defined in 3GPP [36], and considering the region locations in Milan's map, we name the four regions as rural, suburban, urban, and city CBD area. Although the population information is not involved in the classification due to the data unavailability, the relationship between wireless traffic and population ensures the accuracy of classification [37]. We can see that the entire city area is classified into 4 kinds of regions:

- **Rural**: This region mostly include the edge areas on the map, which is sparsely populated. Hence, there are few traffic, few activities and few deployed BSs. As shown in Fig. 10, the number of BSs in the square grids belonging to this region is in the range 1-5 and with a median of 3. Considering the area of each square grid is $0.055/\text{km}^2$, according to the typical inter-site distance of MBSs in 3GPP [36], it can be conclude that in most of the square grids belong to this region only MBSs are deployed.
- **Suburban**: Compared with the rural region, the number of BSs in this area mostly ranges in 6-20, with a median of 12. Comparing with the typical density of MBSs, apparently, in suburban region both MBSs and SBSs are deployed.
- **Urban**: The number of BSs in urban ranges in 19-38, with a median of 27. Due to the large transmission power of MBSs, the interference will quickly increase if the number of MBSs grows. Hence, the increasing number of BSs is mostly contributed by the deployment of SBSs.
- **City CBD**: The number of BSs in city center ranges in 32-63, with a median of 48. This region has the highest level of traffic activity, and thus more SBSs are deployed to support the larger traffic volume than other regions.

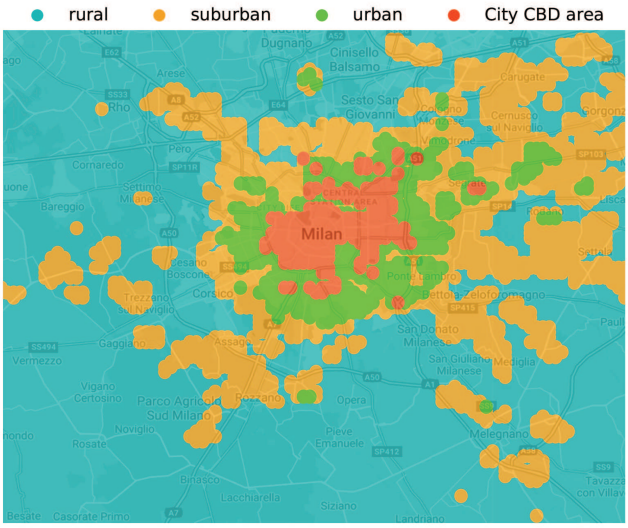


Fig. 9. Clustering result by K-means.

From the above analysis, we can conclude that: *1)* in mostly rural region, only a few MBSs are deployed to ensure a basic network coverage; *2)* for the other three regions, the MBSs and SBSs are both deployed, while the number of SBSs is increasing with the higher traffic requirement.

C. Quantifying MBSs' and SBSs' Capacities by Transfer Learning

As introduced in Section III-C and mentioned above, the original dataset does not differentiate MBSs and SBSs. That is, it only provides the overall traffic and the number of BSs in each square grid. Hence, it is difficult to obtain the capacity functions of MBSs and SBSs directly. However, according to our previous analysis, we know that in most square grids classified as rural regions, only MBSs are deployed. This

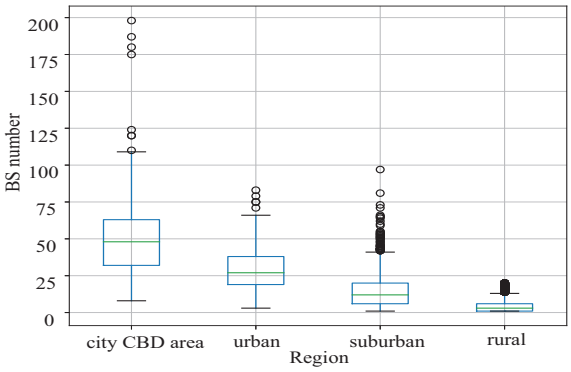


Fig. 10. The distribution of BSs in different regions ("○" means abnormal point).

partition of data can be used to characterize the traffic feature of MBSs, with which the traffic feature of SBSs can be deduced from the data in the other regions.

Following this idea, transfer learning is adopted, which utilizes the knowledge gained from a source domain to efficiently solve a similar problem or the same problem in a target domain. According to [38], the transfer learning can be categorized into three kinds: inductive transfer learning aiming to solve a different but related problem, transductive transfer learning aiming at the same problem but with different data distribution, and unsupervised transfer learning focusing on unsupervised learning tasks.

In our work, with the learned knowledge about the traffic that a number of MBSs can provide in rural regions, transductive transfer learning can use this knowledge to find the number of deployed MBSs in the other regions. Since each square grid has the same size, $0.055/\text{km}^2$, the number of MBSs in each grid should remain stable.

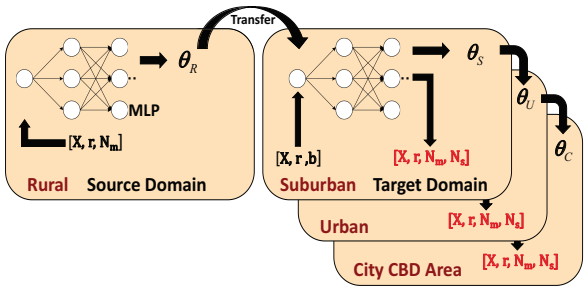


Fig. 11. Transfer Learning with Different Regions.

As shown in Fig. 11, firstly, the traffic features provided by MBSs is captured by an MLP network using the data of rural regions. With the obtained relationship between the traffic $\boldsymbol{X}$, the number of MBSs N_m, and the environment characteristics $\boldsymbol{r}$, the numbers of MBSs deployed in the square grids of the other three regions are obtained by the transfer learning. Due to the similarity measure of environment characteristics among different regions, the transfer learning is proceeded step by step, that is, from rural to suburban, from suburban to urban, and then from urban to city CBD.

During this process, with the numbers of MBSs and SBSs in each square grid, N_m and N_s, the MLP is adopted here to modeling the capacity functions $\mathcal{C}_m(N_m, \boldsymbol{r})$ and $\mathcal{C}_s(N_s, \boldsymbol{r})$. Note that, the maximum traffic z is used to represent the capacity of BSs.

VI. The Optimal BS Sleeping Strategy

In this section, the optimal BS sleeping strategy is investigated which aims to minimize the energy consumption in a given area. Since the data related to each BS's detail deployment parameters and users' traffic are always unavailable due to business confidentiality at the network operators or laws and regulations to protect users' privacy. Hence, it is difficult to obtain such kind of public data sets. In this work, we aim to develop a data-driven BS sleeping strategy based on public data with limited general cellular networks [39], e.g., overall traffic and BSs's number, therefore, we define a cellular energy saving problem to seek the optimal BSs' number in an area.

According to [40], the energy consumption of a BS mainly includes the following parts: 1) the fixed operating energy consumption of a BS, 2) the energy consumption of the circuit components in BS' antennas which is proportional to the number of antennas, 3) the energy consumption of the power amplifier, and 4) the energy consumption on traffic load which is linear with the traffic provided. The efficiency of traffic load consumption of MBSs and SBSs denoted by κ_m and κ_m, respectively. Except the energy related to traffic load, the other three parts of energy consumption can be viewed as a fix consumption for a BS. Denoted by p_m and p_s the fix consumption of an MBS and an SBS, respectively, the energy optimization problem in Eq. (1) can be reformulated as:

$$\min_{n_s, n_m} \; n_m p_m + \kappa_m \mathcal{C}_m(n_m, \mathbf{r}) + n_s p_s + \kappa_s \mathcal{C}_s(n_s, \mathbf{r}), \tag{9a}$$

$$\text{s.t. } \mathcal{C}_m(n_m, \mathbf{r}) + \mathcal{C}_s(n_s, \mathbf{r}) \geq \mu + \Delta, \tag{9b}$$

$$\lceil \frac{N_m}{2} \rceil \leq n_m \leq N_m, \tag{9c}$$

$$0 \leq n_s \leq N_s. \tag{9d}$$

Here, we assume that at least one half of the deployed MBSs should be activated to ensure a reliable umbrella coverage.

1) Global Optimal Solution by Exhaustive Searching: Reviewing the optimization problem in Eq. (1), its formulation and convexity depends on the functions, $\mathcal{C}_s(n_s, \mathbf{r})$ and $\mathcal{C}_m(n_m, \mathbf{r})$. Since they are computed by a neural network, their rigorous math expressions are hard to formulate. Hence, an exhaustive searching algorithm can be used to find the optimal results of n_m and n_s in the feasible region defined by the problem constraints.

2) Optimization with Linear Capacity Model: In many existing research works on cellular network analysis, the network throughput is usually assumed to increase linearly with respect to the number of deployed BSs, especially in sparse scenarios where the aggregated interference is relatively small [41, 42]. This phenomenon is also supported by many industry reports. In Fig. 13, it can be seen that, for a given square grid with characteristics $\boldsymbol{r}$, 1) $\mathcal{C}_s(n_s, \mathbf{r})$ almost increases linearly with n_s in the investigated density range, and 2) $\mathcal{C}_m(n_m, \mathbf{r})$ increases

linearly when the density of MBSs is less than $50/\text{km}^2$, i.e., 3 MBSs in a square grid.

Inspired by these results, linear capacity model can be used to approximately fit the results of $\mathcal{C}_s(n_s, \mathbf{r})$ and $\mathcal{C}_m(n_m, \mathbf{r})$. Then the problem in Eq. (9) can be rewritten as

$$\min_{n_s, n_m} \; n_m p_m + \kappa_m(a_1 n_m + b_1) + n_s p_s + \kappa_s(a_2 n_s + b_2), \tag{10a}$$

$$\text{s.t.} \;\; a_1 n_m + b_1 + a_2 n_s + b_2 \geq \mu + \Delta, \tag{10b}$$

$$\lceil \frac{N_m}{2} \rceil \leq n_m \leq N_m, \tag{10c}$$

$$0 \leq n_s \leq N_s. \tag{10d}$$

Apparently, Problem (10) is a linear programming problem, whose optimal result can be quickly found using math tools like CVX. The complexity is much lower than the exhaustive search.

3) Optimization with Nonlinear Capacity Model: Apparently, the rapidly increasing aggregated inter-cell interference in the network densification will suppress the growth of the network throughput, such kind of results have been verified by many theoretical analysis on dense and ultra-dense networks [43]. Even with various kinds of interference coordination and cancelation techniques, the linear increase of network capacity cannot be achieved in the dense networks.

Also from our results shown in Fig. 13, compared with the SBSs with a low transmission power, the capacity of MBSs increases much slower when the number of deployed MBSs exceeds 3 or 4 per square grid. To better approach the real capacity achieved by the datasets, a non-linear function function to fit the outcome of $\mathcal{C}_m(n_m, \mathbf{r})$, such as $a_1 n_m^2 + b_1 n_m + c_1$. For SBSs, due to the density range in the dataset, this performance degradation does not appear in the outcome. Since the linear model works well and leads to a smallest fitting error, the linear model is still preferred for $\mathcal{C}_s(n_s, \mathbf{r})$. With such models, the optimization problem can be reformulated as

$$\min_{n_s, n_m} \; n_m p_m + \kappa_m(a_1 n_m^2 + b_1 n_m + c_1) + n_s p_s + \kappa_s(a_2 n_s + b_2), \tag{11a}$$

$$\text{s.t.} \;\; a_1 n_m^2 + b_1 n_m + c_1 + a_2 n_s + b_2 \geq \mu + \Delta, \tag{11b}$$

$$\lceil \frac{N_m}{2} \rceil \leq n_m \leq N_m, \tag{11c}$$

$$0 \leq n_s \leq N_s. \tag{11d}$$

Since this optimization is not a convex one. Hence, non-linear programming tools are utilized to find the optimal results. Although the computational complexity is higher than the linear programming for Problem (8), the higher accuracy of this non-linear model is supposed to achieve better performance.

To better present the procedures adopted in our work, Algorithm 1 is organized, which includes: *1)* using the MGCN-LSTM network to predict the traffic requirement, μ; *2)* K-means algorithm to cluster the square grids into different regions; *3)* transfer learning to obtain the numbers of MBSs and SBSs in each square grid, and MLP to model the capacity functions $\mathcal{C}_m(\cdot)$ and $\mathcal{C}_s(\cdot)$; *4)* for a given square grid, fitting the outcome of $\mathcal{C}_m(:, \boldsymbol{r})$ and $\mathcal{C}_s(:, \boldsymbol{r})$ by linear or non-linear models, and calculate the optimal solutions n_m^{opt} and n_s^{opt}.

Algorithm 1 BS Sleeping Strategy Based on Traffic Prediction and Capacity Modeling.

Offline Stage

1. MGCN-LSTM Model for Traffic Prediction:

Input: adjacency matrix $\boldsymbol{A}^N, \boldsymbol{A}^F, \boldsymbol{A}^S$ based on Eq.(3), Eq.(4) and Eq.(5),cellular traffic data segments $\boldsymbol{\mathcal{X}}_r, \boldsymbol{\mathcal{X}}_d, \boldsymbol{\mathcal{X}}_\omega$ from Eq.(7), Eq.(8) and Eq.(9)

Output: Learned MGCN-LSTM network with parameter $\boldsymbol{\theta}$

Step 1. Randomly initialize the MGCN-LSTM network parameter;

While not end of epoch **do**

Step 2. train the MGCN-LSTM as illustrated in Fig. 3 using $\{\boldsymbol{A}^N, \boldsymbol{A}^F, \boldsymbol{A}^S\}$ and $\{\boldsymbol{\mathcal{X}}_r, \boldsymbol{\mathcal{X}}_d, \boldsymbol{\mathcal{X}}_\omega\}$;

Step 3. update the MGCN-LSTM network parameter $\boldsymbol{\theta}$ by adaptive moment estimation (ADAM) optimizer;

End While

2. Capacity Modeling by Transfer Learning

Input: maximum traffic $\mathbf{z}$, BS number $\boldsymbol{B}$, POI $\boldsymbol{P}$, area characteristic $\mathbf{r}$

Output: capacity models: $\mathcal{C}_m(\cdot)$ and $\mathcal{C}_s(\cdot)$, N_m and N_s

Step 1. K-means uses feature $\{\mathbf{z}, \boldsymbol{B}, \boldsymbol{P}\}$ to cluster the entire city into different regions, R_q;

Step 2. use data of grids $\{z, \mathbf{r}, B\}$ in R_1 as source domain data to train $\mathcal{C}_m(\cdot)$ using MLP;

For region R_q: q from 2 to 4

Step 3. use data of grids $\{z, \mathbf{r}, B\}$ in R_q as target domain data to fine-tuning $\mathcal{C}_m(\cdot)$ and obtain N_m.

Step 4. use data $\{z - \mathcal{C}_m(\mathbf{r}, N_m), \mathbf{r}, B - N_m\}$ to train $\mathcal{C}_s(\cdot)$ using MLP, and further fine-tuning.

End For

Online Stage Part

3. Optimal BS Sleeping Strategy for a Square Grid:

Input: MGCN-LSTM network with parameter $\boldsymbol{\theta}$, capacity models $\mathcal{C}_m(\cdot)$ and $\mathcal{C}_s(\cdot)$; parameters of the square grid: traffic X, BS number N_m and N_s, area characteristic $\mathbf{r}$;

Output: the optimal number of active MBSs and SBSs, n_m^{opt} and n_s^{opt} in this interval

Step 1: use the MGCN-LSTM network to achieve the predicted traffic μ;

Step 2: linear fitting: $\mathcal{C}_m(\mathbf{r}, n_m) \approx \mathcal{C}_m^{\text{L}}(n_m) \triangleq a_1 n_m + b_1$, $\mathcal{C}_s(\mathbf{r}, n_s) \approx \mathcal{C}_s^{\text{L}}(n_s) \triangleq a_2 n_s + b_2$;

Step 3: nonlinear fitting: $\mathcal{C}_m(\mathbf{r}, n_m) \approx \mathcal{C}_m^{\text{N}}(n_m) \triangleq a_1 n_m^2 + b_1 n_m + c_1$;

Step 4: Find n_m^{opt} and n_s^{opt} with different approaches;

Approach 1: use original capacity models $\mathcal{C}_m(\cdot)$ and $\mathcal{C}_s(\cdot)$

For $n_m = \lceil \frac{N_m}{2} \rceil$ to N_m

For $n_s = 0$ to N_s

If $\mathcal{C}_m(\mathbf{r}, n_m) + \mathcal{C}_s(\mathbf{r}, n_s) \geq \mu + \Delta$

$P_{\text{all}} = n_m P_m + n_s P_s$;

record $\{P_{\text{all}}, n_m, n_s\}$;

End If

TABLE III
THE PERFORMANCE ACHIEVED BY MGCN WITH DIFFERENT GRAPHS.

Data	Performance metrics	One graph			Two graphs			Three graphs
		G_N	G_F	G_S	G_N and G_F	G_N and G_S	G_F and G_S	G_N, G_F and G_S
SMS	RMSE	107.14	110.01	105.66	91.60	94.73	97.48	90.14
	MAE	72.08	80.47	76.77	65.32	64.86	68.68	63.88
	R2	0.963	0.958	0.955	0.975	0.969	0.965	0.978
Call	RMSE	109.16	113.65	106.81	84.98	95.32	101.35	72.07
	MAE	77.87	80.40	67.77	58.58	65.37	65.71	48.44
	R2	0.963	0.971	0.965	0.981	0.971	0.967	0.985
Internet	RMSE	561.52	635.69	553.18	519.20	535.43	572.49	493.88
	MAE	457.39	486.62	414.75	424.52	397.19	417.33	359.61
	R2	0.961	0.943	0.957	0.968	0.967	0.957	0.972

End For
End for
$n_m^{\text{opt}}, n_s^{\text{opt}} = \arg\min_{n_m, n_s} P_{\text{all}}$;
Approach 2: with approximated models $\mathcal{C}_m^{\text{L}}(\cdot)$ and $\mathcal{C}_s^{\text{L}}(\cdot)$
solve problem in Eq.(7) by CVX;
Approach 3: with approximated models $\mathcal{C}_m^{\text{N}}(\cdot)$ and $\mathcal{C}_s^{\text{L}}(\cdot)$
solve problem in Eq.(9) using matlab function *fmincon()*;

VII. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, extensive experiments are conducted to evaluate the performance of our cellular traffic prediction model and the proposed dynamic BS sleeping strategy. For MBS and SBS, the parameters are adopted according to [40, 44]: *1*) the fixed operating energy consumption: $10W$ and $5W$; *2*) the number of antennas: 3 and 1, and the circuit components energy consumption of antennas: $1W$ and $0.8W$; *3*) power amplifier efficiency: 0.128 and 0.12, and input power of BS: 43dBm and 33dBm; *4*) the efficiency of load consumption: $1.15 \cdot 10^{-9}$J/bit and $1.05 \cdot 10^{-9}$J/bit.

A. Traffic Prediction by MGCN-LSTM

1) Evaluation metrics and prediction performance of the proposed model: Three performance metrics are investigated in our work: root mean square error (RMSE), mean absolute error (MAE) and R-square (R2), which are formally written as: RMSE $= \sqrt{\frac{\sum_{t=1}^{n}(\mu^t - X_{(i,j)}^t)^2}{n}}$, MAE $= \sqrt{\frac{\sum_{t=1}^{n}|\mu^t - X_{(i,j)}^t|}{n}}$, R2 $= 1 - \frac{\sum_{t=1}^{n}(\mu^t - X_{(i,j)}^t)^2}{\sum_{t=1}^{n}(\bar{\mu} - X_{(i,j)}^t)}$, where $X_{(i,j)}^t$ represents the ground-truth traffic flow and μ is the corresponding predicted traffic flow, n is the size of traffic flow, and $\bar{\mu}$ is the average of μ. Note that, the smaller the MSE and RMSE values are, the more accurate the performance will be, while R2 indicates in an opposite way.

2) Prediction performance achieved by MGCN: In Table III, we show the performance achieved by the MGCN when different types of graphs are joined. Generally, it can be seen that the more graphs are involved, the better performance can be achieved.

For instance, in the prediction of SMS traffic, compared with an one-graph convolution network, the two-graph structure achieves a better performance, and the three-graph structure is superior to the former two. In more detail, for the SMS traffic prediction, MGCN with a three-graph structure improves the RMSE and MAE by 7.5% and 7.0%, respectively, compared with that achieved by MGCN with a two-graph structure (G_F and G_S). In addition, MGCN with three-graph improves the RMSE and MAE by 12% and 15.3%, respectively, compared with that achieved by MGCN with the one graph (G_N).

To verify the convergence of proposed prediction model, the loss function in each training and testing epoch are plotted in Fig. 12. It can be observed that the value of the loss function quickly decreases and converges to a small value, after 10 epochs, the loss performance is relatively stable. And the training and testing loss performance tend to be stable after 40 epochs, indicating that MGCN-LTSM can converge and the training process is time efficient.

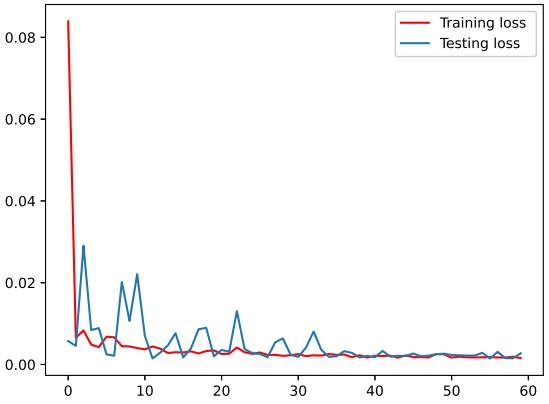


Fig. 12. The convergence illustration.

3) Comparison with other algorithms: To guarantee the comparison fairness, the network parameters and training data are kept consistent. In Fig. 13, the proposed MGCN-LSTM algorithm is compared with several conventional prediction methods: autoregressive integrated moving average (ARIMA), LSTM networks and ConvLSTM.

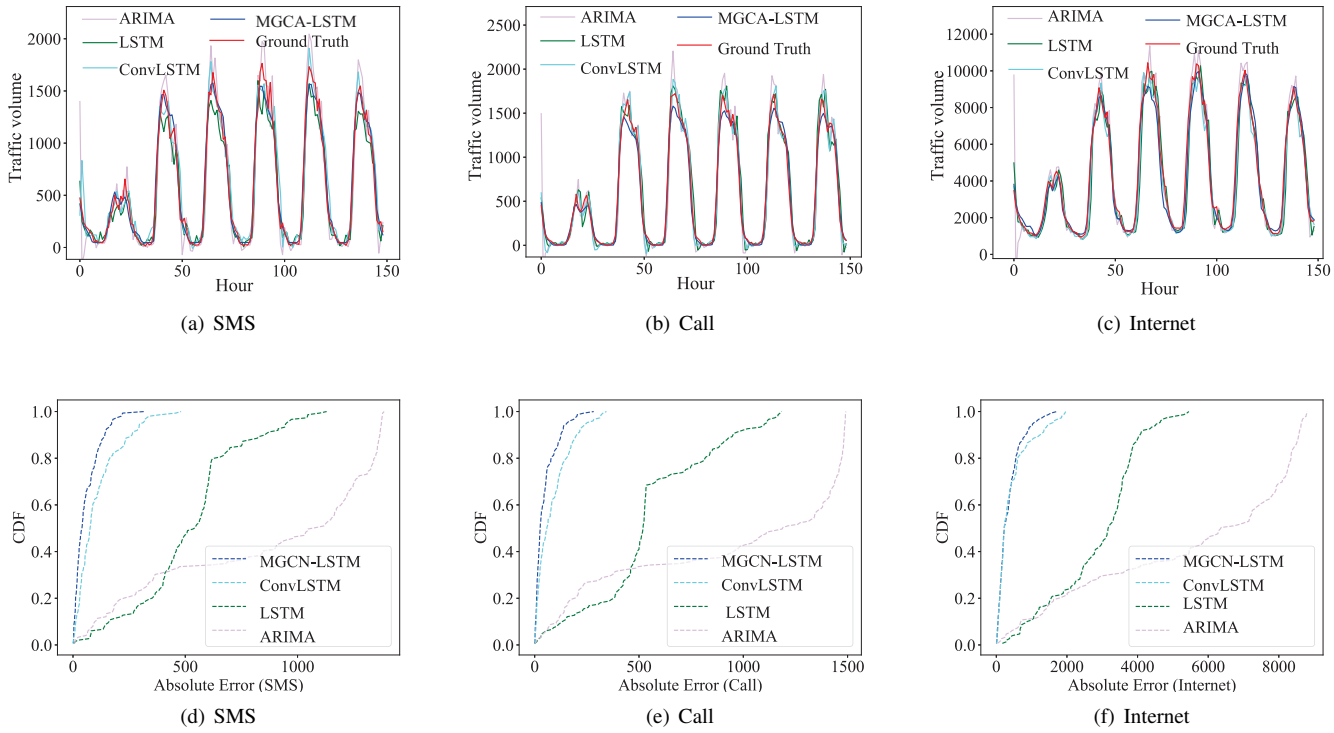


Fig. 13. Comparisons of prediction versus ground truth for all methods on dataset.

Fig. 13(a), 13(b) and 13(c) show the prediction results of the traffic in SMS, CALL and Internet, respectively. It can be seen that the ARIMA model has the largest deviation since ARIMA pays attention to the average value of the past moments. Moreover, LSTM outperforms ARIMA since LSTM is able to capture the temporal correlation. Furthermore, ConvLSTM can simultaneously extract spatial-temporal characteristics, so it has a better prediction performance than LSTM. Most importantly, the proposed MGCN-LSTM exhibits the best performance, especially in terms of peak and valley of traffic prediction. This can be attributed to the fact that MGCN-LSTM utilizes multi-graph convolution, which captures spatial features from various aspects.

To show the performance comparison in a clearer way, Fig. 13 (d)-(f) plot the cumulative distribution function (CDF) of the predicted error. From Fig. 13(d), we can see that for SMS traffic that 80% prediction errors with ARIMA, LSTM, ConvLSTM and MGCN-LSTM are less than 1335, 622, 158, 103, respectively. Compared with ConvLSTM, MGCN-LSTM achieves a performance improvement of 34.8%. Similarly, in Fig. 13(e) and Fig. 13(f), MGCN-LSTM achieves about 47% and 11.6% improvement compared with ConvLSTM, respectively.

In addition, the detail values of the performance metrics related to these 4 algorithms are listed in Table IV. We can see that, for SMS, MGCN-LSTM brings about 56.7%, 43.3% and 43.5% improvements in terms of RMSE compared to ARIMA, LSTM and ConvLSTM, respectively. The improvements in Call traffic are 59.4%, 56.1% and 39.1%, and in Internet traffic, 46.2%, 45.6% and 22.7%. For MAE and R2, a significant performance improvement brought by MGCN-LSTM can also be observed.

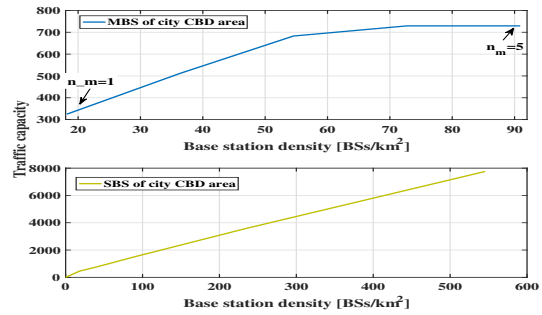
TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT PREDICTION ALGORITHMS

Dataset	Metrics	ARIMA	LSTM	Conv-LSTM	MGCN-LSTM
SMS	RMSE	189.00	144.22	144.86	81.79
	MAE	135.31	105.96	106.77	58.74
	R2	0.889	0.917	0.932	0.978
Call	RMSE	177.67	164.06	118.42	72.07
	MAE	115.04	109.09	85.31	48.44
	R2	0.917	0.937	0.967	0.985
Internet	RMSE	919.91	910.65	640.22	494.95
	MAE	579.67	611.68	431.53	374.18
	R2	0.908	0.912	0.953	0.972

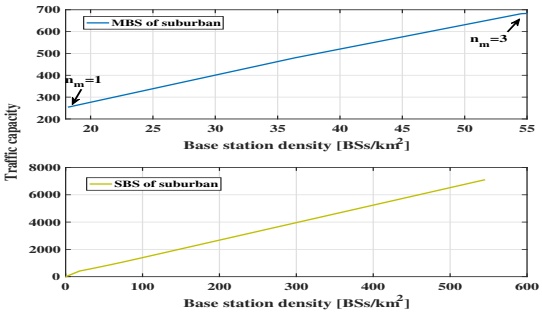
The advantage of the proposed algorithm is achieved by the following reasons: 1) The multiple spatial features are extracted by multi-graph through the GCN module; 2) Multi-domain data sets: POIs, BSs, social activity and multiple periodic features are involved in time-domain prediction; 3) The attention mechanism is adopted to optimize the features extracted.

B. Capacity Modeling of MBSs and SBSs

Fig. 14 plots the capacity functions of MBSs and SBSs obtained in two square grids, where the first grid belongs to the city CBD and the second grid belongs to suburban region. Firstly, it is observed that although the environment characteristics $\boldsymbol{r}$ in two grids are different, the capacity curves vs. the BS number are quite similar. That is, the network



(a) city CBD area.



(b) suburban.

Fig. 14. The capacity modeling of MBSs and SBSs in city CBD area and suburban.

capacity mainly depends on the number of deployed BSs, and the performance impact caused by environment difference is limited.

Secondly, in both grids, with the increasing number of deployed MBSs, the capacity provide firstly increases almost linearly with the BS number. However, in the grid belonging to city CBD, with the further increase of MBSs' number, the capacity increase slows down, which is consistent with the analysis in [43]. The main reason behind is the sever interference aggregated during the process of densification. In contrast, this phenomenon dose not show in the capacity curves of SBSs learned by the Milan's data. Due to the low transmission power of SBSs, the degradation caused by interference appears in ultra-dense scenarios. For the investigated data of Milan at several years ago, it seems that the deployment of SBSs has not entered this regime.

C. Performance of BS Sleeping Strategy

In Fig. 15, we plot the predicted traffic flow, the optimal numbers of active MBSs and SBSs based on 1) search using the obtained $\mathcal{C}(n_m, \boldsymbol{r})$ and $\mathcal{C}(n_s, \boldsymbol{r})$, 2) linear programming using linear approximations of $\mathcal{C}(n_m, \boldsymbol{r})$ and $\mathcal{C}(n_s, \boldsymbol{r})$, and 3) non-linear programming using approximate quadratic function for $\mathcal{C}(n_m, \boldsymbol{r})$. Here, the results in square grids $(50, 61)$ and $(38, 61)$ are shown, which belong to city CBD region and suburban region, respectively.

Firstly, we can see that no matter what kind of method is used and what region the square grid belongs to, the number of active SBSs changes with the traffic flow, while the number of active MBSs changes little. Since MBSs aim at ensuring the cell coverage and handling the user mobility, a minimum number of MBSs is guaranteed to power on in our strategy. When the traffic requirement exceeds this basic level, more SBSs and MBSs are activated to enhance the network capacity. Since the power consumption of SBSs is much less than MBSs, with the aim of energy saving, SBSs will be activated firstly to boost the network throughput.

Secondly, compared the three methods, the optimal numbers of MBSs and SBSs achieved by the approximated nonlinear model are very close to those obtained by exhaustive searching. In more detail, the approximated linear model leads to a more frequent change of MBSs' number, and a bigger number of active SBSs in peak times. From these results, it can be concluded that the approximate quadratic function is more accurate than the linear function.

Thirdly, comparing the curves in the square grids of the suburban region with those of the city CBD regions, we can see that the traffic is much less than that in city CBD, and hence the deployed MBSs and SBSs are both less than those deployed in city CBD.

Moreover, from the figure, even the number of active SBSs changes more quickly than MBSs, on average it changes every 2-3 hours in a relatively regular pattern . That is, the curves oscillate on a multi-hour basis. This frequency is more slower than the strategies proposed in many previous works by detecting the real-time cell load.

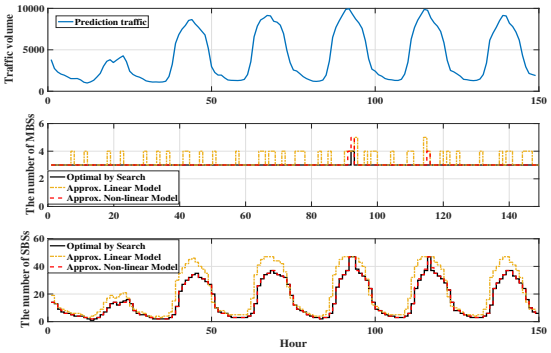
To show the energy consumption with these three methods in the two square grids, Fig. 16 is plotted, apparently, the energy consumption is highly related to the numbers of active MBSs and SBSs in this square grid. Hence, similar with the results in Fig. 15, the nonlinear model, i.e., quadratic function fitting, for $\mathcal{C}_m(n_m, \boldsymbol{r})$ achieves near-optimal performance in energy consumption. Since the energy consumption also depends on the traffic load, the curve is continuously

Moreover, compared with the energy consumption with all active MBSs and SBSs as 100%, the energy saved by the BS sleeping strategy using non-linear model is more obvious in city CBD. In more detail, the energy saved in city CBD approaches 63% in one week, and near 54% of the energy is saved in suburban regions. This is because the large number of BSs deployed and the traffic disparity between the peak and off-peak time.

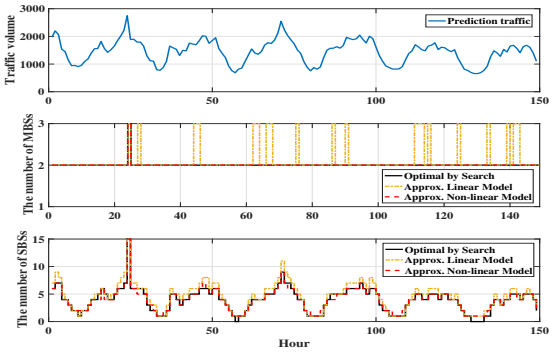
Moreover, we compare the saved energy achieved by our proposed algorithm with the strategy in [45]. In detail, square grid with index 5061 in city CBD region and grid 3859 in suburban region are investigated, with the all active power consumption as 100%. As shown in Table V, it can be observed that in both regions our proposed algorithm has better performance, and in the city CBD region with more BSs, the effect of energy-saving is more obvious.

TABLE V
THE AVERAGE ENERGY SAVED(%) IN DIFFERENT ALGORITHMS.

	comparison algorithm	our proposed algorithm (%)
Suburban	52%	55%
City CBD	57%	63%

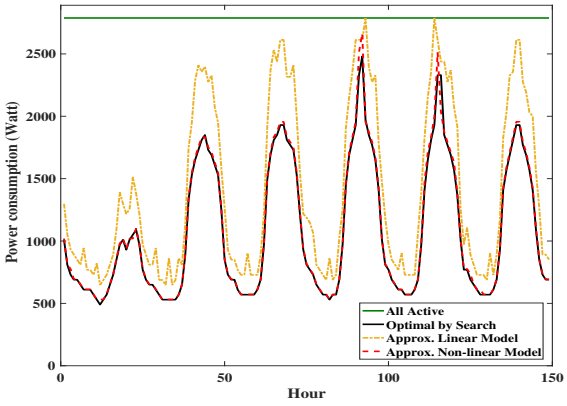


(a) City CBD area.

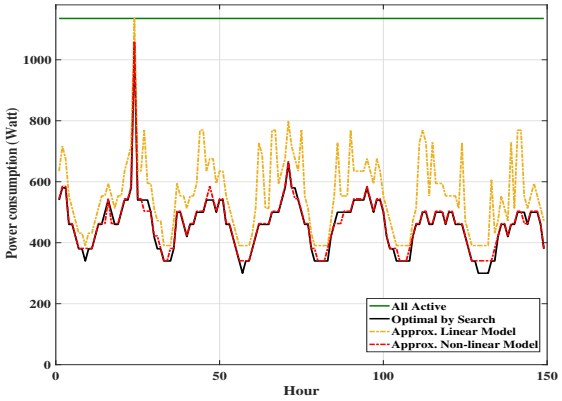


(b) Suburban.

Fig. 15. The top subgraph is the predicted traffic flow, the middle and bottom are the number of active MBSs and SBSs, respectively.



(a) City CBD area.



(b) Suburban.

Fig. 16. Performance comparison of different algorithms for power consumption.

VIII. CONCLUSION

In this paper, a data-driven BS sleeping control strategy was designed based on the traffic prediction and capacity modeling. To better extract the spatial characteristics in different domains, a MGCN framework was introduced together with the LSTM network in time domain prediction and an attention mechanism. The proposed MGCN-LSTM framework achieved a favorable performance in traffic prediction. The capacity models of MBSs and SBSs in different environments were learned by transfer learning using the data in rural regions as the source domain. At last, optimal BS sleeping algorithms were proposed to minimize the power consumption. From the simulation results, more energy can be saved by in the regions with a large number of deployed BSs or with a more sever traffic fluctuation. The approximate non-linear model of capacity function achieved a near-optimal performance with a relatively low complexity.

REFERENCES

[1] X. Guo, Z. Niu, S. Zhou, and P. R. Kumar, "Delay-constrained energy-optimal base station sleeping control," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1073–1085, 2016.

[2] M. Li, P. Li, X. Huang, Y. Fang, and S. Glisic, "Energy consumption optimization for multihop cognitive cellular networks," *IEEE Transactions on Mobile Computing*, vol. 14, no. 2, pp. 358–372, 2015.

[3] J. Kim, W. S. Jeon, and D. G. Jeong, "Effect of base station-sleeping ratio on energy efficiency in densely deployed femtocell networks," *IEEE Communications Letters*, vol. 19, no. 4, pp. 641–644, 2015.

[4] M. A. Marsan, A. F. Anta, V. Mancuso, B. Rengarajan, P. R. Vasallo, and G. Rizzo, "A simple analytical model for energy efficient ethernet," *IEEE Communications Letters*, vol. 15, no. 7, pp. 773–775, 2011.

[5] A. Baiocchi, L. Chiaraviglio, F. Cuomo, and V. Salvatore, "Joint management of energy consumption, maintenance costs, and user revenues in cellular networks with sleep modes," *IEEE Transactions on Green Communications and Networking*, vol. 1, no. 2, pp. 167–181, 2017.

[6] J. Wu, S. Zhou, and Z. Niu, "Traffic-aware base station sleeping control and power matching for energy-delay tradeoffs in green cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 8, pp. 4196–4209, 2013.

[7] J. T. Louhi, "Energy efficiency of modern cellular base stations," in *INTELEC 07 - 29th International Telecommunications Energy Conference*, 2007, pp. 475–476.

[8] P. Frenger, P. Moberg, J. Malmodin, Y. Jading, and I. Godor, "Reducing energy consumption in lte with cell dtx," in *2011 IEEE 73rd Vehicular Technology Conference (VTC Spring)*, 2011, pp. 1–5.

[9] A. T. Koc, S. C. Jha, R. Vannithamby, and M. Torlak, "Device power saving and latency optimization in LTE-A networks through DRX configuration," *IEEE Transactions on Wireless Communications*, vol. 13, no. 5, pp. 2614–2625, 2014.

[10] J. Liu, B. Krishnamachari, S. Zhou, and Z. Niu, "Deepnap: Data-driven base station sleeping operations through deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4273–4282, 2018.

[11] P. Dong-Chul, "Structure optimization of bilinear recurrent neural networks and its application to ethernet network traffic prediction," *Information Sciences*, vol. 237, pp. 18–28, 2013.

[12] Y. Hou, L. Zhao, and H. Lu, "Fuzzy neural network optimization and network traffic forecasting based on improved differential evolution," *Future Generation Computer Systems*, vol. 81, no. APR., pp. 425–432, 2017.

[13] L. Nie, D. Jiang, S. Yu, and H. Song, "Network traffic prediction based on deep belief network in wireless mesh backbone networks," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 1–5.

[14] W. Huang, G. Song, H. Hong, and K. Xie, "Deep architecture for traffic flow prediction: Deep belief networks with multitask learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2191–2201, 2014.

[15] J. Wang, J. Tang, Z. Xu, Y. Wang, G. Xue, X. Zhang, and D. Yang, "Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, 2017, pp. 1–9.

[16] Y. Wu, H. Tan, L. Qin, B. Ran, and Z. Jiang, "A hybrid deep learning based traffic flow prediction method and its understanding," *Transportation research*, vol. 90, no. 5, pp. 166–180, 2018.

[17] L. Chen, D. Yang, D. Zhang, C. Wang, J. Li, and T.-M.-T. Nguyen, "Deep mobile traffic forecast and complementary base station clustering for c-ran optimization," *Journal of Network and Computer Applications*, vol. 121, pp. 59 – 69, 2018.

[18] V. Balachandran, R. Kadarkarayandi, J. NZ, and V. Sahil, "An attention-based deep learning model for traffic flow prediction using spatiotemporal features towards sustainable smart city," *International Journal of Communication Systems*, 2020.

[19] H. Zheng, F. Lin, X. Feng, and Y. Chen, "A hybrid deep learning model with attention-based Conv-LSTM networks for short-term traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2020.

[20] C. Zhang, H. Zhang, J. Qiao, D. Yuan, and M. Zhang, "Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1389–1401, 2019.

[21] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 12 2016.

[22] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li, "T-gcn: A temporal graph convolutional network for traffic prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 9, pp. 3848–3858, 2020.

[23] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 7 2018.

[24] X. Geng, Y. Li, L. Wang, L. Zhang, and Y. Liu, "Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[25] D. Chai, L. Wang, and Q. Yang, "Bike flow prediction with multi-graph convolutional networks," in *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 11 2018.

[26] C. Jia and T. J. Lim, "Resource partitioning and user association with sleep-mode base stations in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 7, pp. 3780–3793, 2015.

[27] Z. Niu, X. Guo, S. Zhou, and P. R. Kumar, "Characterizing energy-delay tradeoff in hyper-cellular networks with base station sleeping control," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 4, pp. 641–650, 2015.

[28] L. Wang and S. Cheng, "Self-organizing ultra-dense small cells in dynamic environments: A data-driven approach," *IEEE Systems Journal*, vol. 13, no. 2, pp. 1397–1408, 2019.

[29] H. Jiang, S. Yi, L. Wu, H. Leung, Y. Wang, X. Zhou, Y. Chen, and L. Yang, "Data-driven cell zooming for large-scale mobile networks," *IEEE Transactions on Network and Service Management*, vol. 15, no. 1, pp. 156–168, 2018.

[30] M. Zhang, H. Fu, Y. Li, and S. Chen, "Understanding urban dynamics from massive mobile traffic data," *IEEE Transactions on Big Data*, pp. 1–1, 2017.

[31] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri, "A multi-source dataset of urban life in the city of milan and the province of trentino," *Scientific Data*, 2015.

[32] F. Xu, L. Yong, H. Wang, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 1147–1161, 2017.

[33] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 129–150, 3 2011.

[34] A. Bernini, A. Toure, and R. Casagrandi, "The time varying network of urban space uses in milan," *Applied Network Science*, vol. 4, 12 2019.

[35] Y. Teng, M. Liu, F. R. Yu, V. C. M. Leung, M. Song, and Y. Zhang, "Resource allocation for ultra-dense networks: A survey, some research issues and challenges," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2134–2168, 2019.

[36] 3GPP, "TR 36.814: Further advancements for E-UTRA physical layer aspects," Mar. 2010.

[37] G. Khodabandelou, V. Gauthier, M. Fiore, and M. A. El-Yacoubi, "Estimation of static and dynamic urban populations with mobile network metadata," *IEEE Transactions on Mobile Computing*, vol. 18, no. 9, pp. 2034–2047, 2019.

[38] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[39] U. Paul, J. Liu, S. Troia, O. Falowo, and G. Maier, "Traffic-profile and machine learning based regional data center design and operation for 5g network," *Journal of Communications and Networks*, vol. 21, no. 6, pp. 569–583, 2019.

[40] E. Björnson, L. Sanguinetti, and M. Kountouris, "Deploying dense networks for maximal energy efficiency: Small cells meet massive MIMO," vol. 34, no. 4, 2016, pp. 832–847.

[41] R. Tanbourgi, S. Singh, J. G. Andrews, and F. K. Jondral, "A tractable model for noncoherent joint-transmission base station cooperation," *IEEE Transactions on Wireless Communications*, vol. 13, no. 9, pp. 4959–4973, 2014.

[42] M. Ding, P. Wang, D. López-Pérez, G. Mao, and Z. Lin, "Performance impact of los and nlos transmissions in dense cellular networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 2365–2380, 2016.

[43] M. Ding, D. López-Pérez, Y. Chen, G. Mao, Z. Lin, and A. Zomaya, "UDN: A holistic analysis of multi-piece path loss, antenna heights, finite users and bs idle modes," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2019.

[44] M. Deruyck, W. Joseph, and L. Martens, "Power consumption model for macrocell and microcell base stations," *European Transactions on Telecommunications*, vol. 25, no. 3, pp. 320–333, 2014.

[45] S. Zhao, "Traffic prediction based power saving in cellular networks: A machine learning method," in *the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 11 2017.