Measurement-driven Capability Modeling for Mobile Network in Large-scale Urban Environment

Jingtao Ding*, Xihui Liu*, Yong Li*, Di Wu[†‡], Depeng Jin*, Sheng Chen[§],
*Tsinghua National Laboratory for Information Science and Technology (TNLIST),
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[†]Imperial College London, London SW7 2AZ, U.K.
[‡]Hunan University, Changsha 410082, China
[§]Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K.
E-mail: liyong07@tsinghua.edu.cn

***Abstract*—For mobile networks diverse usage scenarios have different capability requirements on *connection density* and *user experienced data rate*, and modeling such capability diversity is crucial to the strategy evaluation in addressing the problem of high traffic load and scalability of network resources. Therefore, it is necessary to build a capability model in two dimensions of *connection density* and *user experienced data rate*. This paper aims at addressing this challenge based on an investigation of network capability in large-scale urban environment. First, our statistical study shows that the spatial distribution of these two parameters can be accurately fitted by log-normal mixture model. Second, we find that only six basic capability patterns exist among the 9,000 cellular base stations. Their connections with the urban functions of geographical locations are also explored in our work. Based on these two discoveries, we build a network capability model which can generate synthetic base stations with diverse *connection density* and *user experienced data rate*. We believe that this flexible and powerful model can help telecommunication operators to design and standardize mobile network in the future.**

I. Introduction

With the tremendous growth in connectivity, density and volume of mobile traffic, both industry and academia are focusing on improving the capability of mobile cellular network. To meet the demands of a fully mobile and connected society, a broad range of usage scenarios for future mobile network are expected and each of them has different network capability requirements. Under these contexts, it is vital to achieve diverse network capabilities in terms of *connection density* and *user experienced data rate*. For example, according to the published white paper [1], in the scenario of *broadband access in dense areas* (e.g., pervasive video), person-to-person or person-to-group video communication with extremely high resolution should be available to every subscriber, where providing such large number of concurrently active connections and high data rate will be a challenge. When it comes to the scenario of *massive Internet of Things*, a single macrocell may need to support 10,000 or more low-rate devices with expected demands in machine-to-machine communication [2]. *Connection density* is a key performance parameter in the scenario of *massive Internet of Things*, while high *user experienced data rate* is more vital in the scenario of *broadband access in dense areas*.

For telecommunication operators, modeling data network capability is extremely valuable in cellular network planning, operation and maintenance, such as performance evaluations of network resources allocation and load balancing. More importantly, the diverse mobile network usage scenarios discussed above require a flexible and powerful model of mobile network capability. Thus it is necessary to build a network capability model on the two-dimensional space of *connection density* and *user experienced data rate*.

However, there exist two challenging problems in modeling mobile network capability:

- How to obtain and analyze *connection density* and *user experienced data rate* of a real cellular network? A large-scale trace data containing these two parameters is vital in the analysis. Also, to build a capability model, we need to consider the spatial distribution of *connection density* and *user experienced data rate*. These tasks are challenging.
- How to extract the key patterns of *connection density* and *user experienced data rate* from the trace data? A suitable clustering method is required, which helps us to understand the network capability in these two dimensions. This is also a difficult task.

To address the first challenge, we carry out a base-station-level analysis of subscriber density and data traffic demand per subscriber in our fine-grained and large-scale trace data, which are collected from a mobile network deployed in *Shanghai*. Subscriber density and data traffic demand per subscriber correspond to the two key parameters of network capability, *connection density* and *user experienced data rate*. In our following study, we use a log-normal mixture model to characterize the spatial distribution of these two metrics. As for the second challenge, we adapt a 2-dimensional clustering method, which is based on Eduardo's work [3]. Moreover, the traffic patterns of cellular base stations do correspond to the urban functions of geographical locations [4]. Inspired by this, we introduce this urban function context information into our capability model. Our key contributions are threefold:

- First, we discover that the spatial distribution of subscriber density and average data traffic demand can be accurately fitted by a log-normal mixture model. Our theoretical proof shows that the product of subscriber density and average data rate, i.e., traffic density, also follows a log-normal mixture distribution spatially, which is further validated by empirical data.
- Next, our extensive analysis provides a precise characterization of individual base station capability and clusters base stations into 6 types according to subscriber density and average data demand. We also explore the relationship between the network capability and urban functional regions where base stations are deployed.

- Finally, we build a network capability model as the function of subscriber density and average data demand. The highlight of our model is that we only need to input the urban function context information, and it can then generate synthetic base stations with realistic diverse capabilities in terms of the two key parameters. Our evaluations demonstrate that this model can reliably and accurately quantify network capability. More importantly, our model provides an insight on how to improve the capabilities of mobile network in diverse usage scenarios.

This paper is structured as follows. In Section II, we detail the utilized mobile network dataset and explain how we extract the useful information, i.e., subscriber density and data traffic demand per subscriber in each cell. In Section III, we analyze the spatial distribution of these two key parameters. In Section IV, using an unsupervised clustering algorithm, we identify the key patterns of network capability. Based on these discoveries, we build a capability model in Section V. After discussing the related work in Section VI, we summarize our work and discuss future investigations in Section VII.

II. Dataset and Key Parameters

A. Dataset

In order to carry out a measurement driven empirical study, we use an anonymous cellular trace from 9,181 cellular base stations deployed in *Shanghai* by one of the major operators in China, within an interval of 31 days in August 2014. Each record of the trace contains detailed mobile data usage of 700,000 subscribers, including the devices ID (anonymized), start-end time of data consumption, base station (BS) ID, BS location and traffic volume (byte). This fine-grained dataset, including both information on subscriber number and data consuming volume, enables us to carry out a comprehensive study on network capability. On the other hand, the large-scale trace, which contains 2.8 petabytes (10^{15}) logs, 92 terabytes (10^{12}) per day and 7 gigabytes (10^{9}) per base station on average, guarantees the credibility of our investigation.

B. Key Parameters

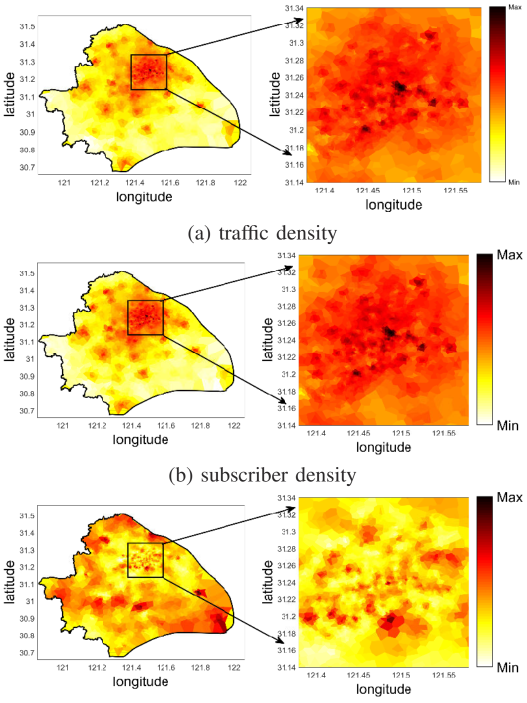
As mentioned previously, subscriber density and average data demand are the two key parameters to describe the network capability. Subscriber density can be computed by counting the number of access devices during a certain period of time. As for data traffic demand per subscriber, it is natural to define it as the total data volume consumed per BS divided by the number of subscribers of the cell.

Each BS delivers different coverage for cellular service. As the actual area of cell coverage is difficult to measure, we obtain the area of Voronoi cells [5] drawn by using the locations of BSs. Let X represent the whole network area. Further let K be the set of BS indices and $B = \{b_k, k \in K\}$ be the set of BSs. The Voronoi cell V_k, associated with the BS b_k, is the set of all the points in X whose distances to b_k are not greater than their distances to any other BS b_j with $j \neq k$. Specifically, if $d(x, b)$ denotes the distance between the point x and the BS b, then $V_k = \{x \in X \mid d(x, b_k) \leq d(x, b_j) \text{ for all } j \neq k,\ j, k \in K\}$. In this way, we divide the area into Voronoi cells. The coverage area of each BS b_k is the area of the corresponding Voronoi cell V_k.

In this way, we obtain the subscriber density by dividing the number of subscribers with the area of the corresponding Voronoi cell, denoted as $S^{b_i}(t)$ (subscribers/km^2) for $b_i \in B$ and $1 \leq t \leq 744$, where t is the time sequence index. The length of the duration is 1 hour, which explains why the maximum is $744 = 24 \times 31$. Similarly, we denote the average data demand per subscriber as $D^{b_i}(t)$ (bytes/subscriber) for $b_i \in B$ and $1 \leq t \leq 744$. It is worth noting that the product of subscriber density $S^{b_i}(t)$ and data demand $D^{b_i}(t)$ is the traffic density, denoted as $T^{b_i}(t)$ (bytes/km^2), which represents the degree of traffic load.

III. Network Capability Analysis

In this section, we focus on three metrics of data traffic: traffic density, subscriber density and average demand. By showing the heat maps, we provide a visual view on how they are distributed in the urban area. Then we propose a model to describe the spatial distributions of the empirical data.



(a) traffic density

(b) subscriber density

(c) average data demand

Fig. 1. Geographical distributions of traffic density (a), subscriber density (b), and average data demand (c).

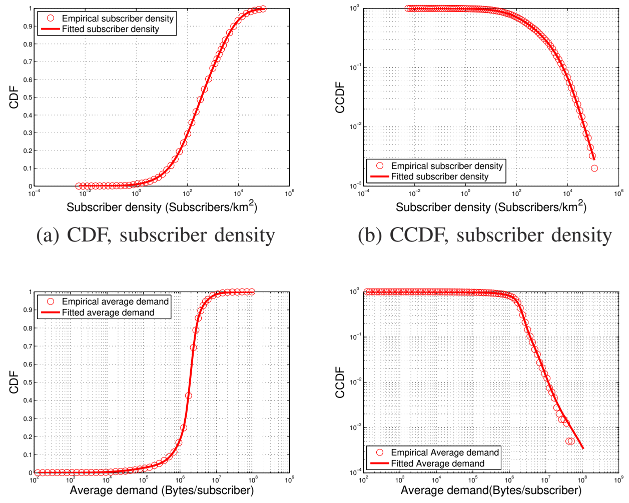
Fig. 1 shows the heat maps of the mean traffic density, subscriber density and average demand in a month. Since the empirical data are highly right-skewed, the log-transformed data are used to draw heat maps. Traffic density and subscriber density are high and concentrated in the city center, while they are relatively low in the rural area. However, the heat map of average demand shows different characteristics: the peak values spread widely, from the city center to rural area.

Our next step is to model the spatial distributions of these three parameters. Researchers [6] found that the spatial distribution of traffic density can be well fitted by a log-normal mixture distribution. We want to know *what is the reason behind this distribution.* Our study also shows that subscriber density and average demand per subscriber can be fitted accurately by log-normal mixture distributions. By theoretical analysis (Proposition 1), we further prove that the product of subscriber density and average demand, i.e., traffic density, also follows a log-normal mixture distribution.

The probability density function (PDF) of the log-normal mixture distribution with l components is:

$$f_X(x) = \sum_{i=1}^{l} p_i \log\mathcal{N}(x; \mu_i, \sigma), \tag{1}$$

where $\log\mathcal{N}(x; \mu_i, \sigma_i)$ is the ith log-normal distribution with location parameter μ_i and scale parameter σ_i, while p_i is the mixture proportion of the ith component and the sum of all the mixture proportions is $\sum_{i=1}^{l} p_i = 1$. The parameters $\{\mu_i, \sigma_i, p_i\}_{i=1}^{l}$ can be obtained for example using the expectation maximization (EM) algorithm [7].



(a) CDF, subscriber density

(b) CCDF, subscriber density

(c) CDF, average demand

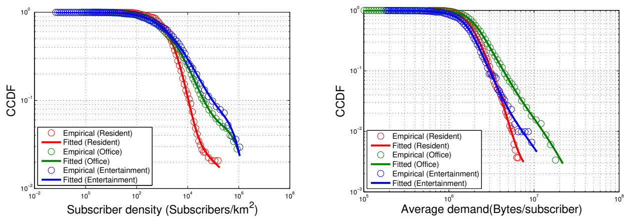
(d) CCDF, average demand

Fig. 2. Log-normal mixture fittings of the spatial distributions of subscriber density and average demand, in the timescale of one hour. The circles represent the empirical data, and the solid lines represent the fitted log-normal mixture distribution.

Log-normal mixture with three components $l = 3$ is used to fit both the subscriber density and average demand in one hour. Fig. 2 shows the cumulative distribution function (CDF) and complementary CDF (CCDF) of the empirical data and fitted model, which indicate that the proposed log-normal mixture distribution fits the empirical data very well. The parameters of the models are listed in Table I. In order to show that this log-normal mixture model is universal in different spatial scales, we also use the log-normal mixture model to fit the empirical distributions in different urban regions, specifically, resident region, office region and entertainment region, as shown in Fig. 3. The Kolmogorov-Smirnov (K-S) test is used to test the goodness of fit [8]. We test the distribution fitting of the cell traffic in every hour in a day at 5% significance level, and find that the log-normal mixtures distribution is accepted all the time.

TABLE I
PARAMETERS OF THE LOG-NORMAL MIXTURE MODELS FOR SUBSCRIBER DENSITY AND AVERAGE DEMAND.

Parameters		Subscriber density	Average demand
Location parameters	μ_1	5.4094	14.5001
	μ_2	5.1033	14.3824
	μ_3	8.2199	12.7958
Scale parameters	σ_1	1.5761	0.2798
	σ_2	2.6085	0.8331
	σ_3	1.2034	1.5647
Mixture proportions	p_1	0.4075	0.4543
	p_2	0.3950	0.4457
	p_3	0.1975	0.1000



(a) CCDF, subscriber density

(b) CCDF, average demand

Fig. 3. Log-normal mixture fittings of the spatial distributions of subscriber density and average demand, in the timescale of one hour, in resident region, office region and entertainment region, respectively. The circles represent the empirical data, and the solid lines represent the log-normal mixture distribution.

To further reveal the relationships among these three parameters (traffic density, subscriber density and average demand per subscriber), correlation coefficients are used to test the correlations between them. The results show that traffic density and subscriber density are highly correlated, with the correlation coefficients greater than 0.9. By contrast, the correlation coefficients between average demand and the other two parameters are less than 0.1, indicating weak correlation between average demand and subscriber density or traffic density. This observation explains why Fig. 1 shows the similarity between traffic density and subscriber density, but very different patterns for average demand.

Moreover, it can be verified that the product of two independent log-normal mixture distributed random variables also follows a log-normal mixture distribution.

Proposition 1. *Assume that X and Y are independent log-normal mixture distributed variables with m and n components, respectively. Let $Z = XY$, then Z follows a log-normal mixture distribution with $m \times n$ components.*

Proof. Since $X' = \log X$ and $Y' = \log Y$ follow the independent Gaussian mixture distributions with m and n components, respectively, their PDFs are

$$f_{X'}(x') = \sum_{i=1}^{m} p_{X_i} \phi(x'; \mu_{X_i}, \sigma_{X_i}^2), \tag{2}$$

$$f_{Y'}(y') = \sum_{j=1}^{n} p_{Y_j} \phi(y'; \mu_{Y_j}, \sigma_{Y_j}^2), \tag{3}$$

where $\phi(x'; \mu_{X_i}, \sigma_{X_i}^2)$ denotes the Gaussian distribution with mean μ_{X_i} and variance $\sigma_{X_i}^2$. Since $Z' = \log Z = \log X + \log Y = X' + Y'$, we have

$$\begin{aligned}
f_{Z'}(z') &= \int_{-\infty}^{\infty} \left(f_{X'}(u) \cdot f_{Y'}(z'-u)\right) du \\
&= \int_{-\infty}^{\infty} \sum_{i=1}^{m} p_{X_i} \phi(u; \mu_{X_i}, \sigma_{X_i}^2) \sum_{j=1}^{n} p_{Y_j} \phi(z'-u; \mu_{Y_j}, \sigma_{Y_j}^2)\, du \\
&= \int_{-\infty}^{\infty} \sum_{i=1}^{m} \sum_{j=1}^{n} p_{X_i} p_{Y_j} \phi(u; \mu_{X_i}, \sigma_{X_i}^2) \phi(z'-u; \mu_{Y_j}, \sigma_{Y_j}^2)\, du \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} p_{X_i} p_{Y_j} \int_{-\infty}^{\infty} \phi(u; \mu_{X_i}, \sigma_{X_i}^2) \phi(z'-u; \mu_{Y_j}, \sigma_{Y_j}^2)\, du \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} p_{X_i} p_{Y_j} \phi(z'; \mu_{X_i} + \mu_{Y_j}, \sigma_{X_i}^2 + \sigma_{Y_j}^2).
\end{aligned} \tag{4}$$

This proves that Z' follows the Gaussian mixture distribution with $m \times n$ components. Thus Z follows a log-normal mixture distribution with $m \times n$ components, and the parameters of the distribution for Z are given by

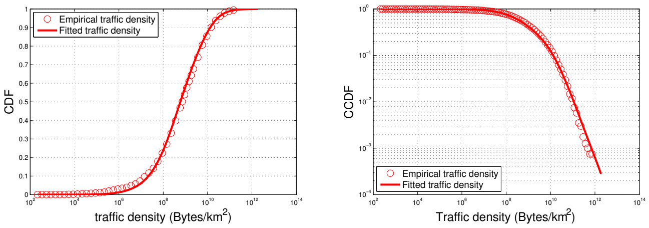
$$\begin{cases} p_{Z_{i,j}} = p_{X_i} p_{Y_j}, \\ \mu_{Z_{i,j}} = \mu_{X_i} + \mu_{Y_j}, \\ \sigma^2_{Z_{i,j}} = \sigma^2_{X_i} + \sigma^2_{Y_j}, \end{cases} \tag{5}$$

for $1 \le i \le m$ and $1 \le j \le n$. □

TABLE II
PARAMETERS OF THE LOG-NORMAL MIXTURE MODEL FOR TRAFFIC DENSITY.

Parameters	Location parameters	Scale parameters	Mixture proportions
Component 1	19.9095	1.6001	0.1851
Component 2	19.7918	1.7828	0.1816
Component 3	19.6034	2.6235	0.1795
Component 4	19.4857	2.7383	0.1761
Component 5	22.7200	1.2355	0.0897
Component 6	22.6023	1.4636	0.0880
Component 7	18.2052	2.6301	0.0408
Component 8	17.8990	3.3522	0.0395
Component 9	21.0156	2.4251	0.0198

As the correlation between subscriber density and average demand is weak, it can be assumed that they are independent. Thus the product of them, i.e., traffic density, follows a log-normal mixture distribution with 9 components. We can use the parameters given in Table I to compute the parameters of the distribution for traffic density, which are listed in Table II. Fig. 4 shows the fitting of the empirical traffic density to the computed log-normal mixture model. Furthermore, the K-S test at 5% significance level also accepts this log-normal mixture distribution. In other words, traffic density also follows a log-normal mixture distribution spatially, which is verified by both empirical data and theoretical proof.



(a) CDF, traffic density

(b) CCDF, traffic density

Fig. 4. Log-normal mixture fitting of the spatial distribution for traffic density, in the timescale of one hour. The log-normal mixture model is calculated by (5). The circles represent the empirical data, and the solid lines represent the theoretical log-normal mixture distribution.

IV. NETWORK CAPABILITY CLUSTERING

In this section, we provide insights into network service capabilities in different urban areas. Firstly, BSs are clustered into six clusters based on the two key parameters: subscriber density and average demand per subscriber which represent network connection density and user experienced data rate. Moreover, we reveal the relationship between these clusters and functional regions of BSs.

A. Clustering Methodology

The key in characterizing network service capability is to find peak hours of BSs. The authors of [4] exploited the spatial information embedded within mobile traffic by identifying key urban functional regions, such as resident region, transport region, office region, entertainment region and comprehensive region. It was found that hourly dynamics of BSs in the same functional region follow the same pattern, with similar peak and non-peak hours. As for BSs in different functional regions, they follow different dynamic patterns and have different peak and non-peak hours. Based on this finding, we define peak hours of BSs in a certain functional region as hours when the average traffic density in this region is over 50% of its peak value. In this way, the averages of $S^{b_i}(t)$ and $D^{b_i}(t)$ for each BS during peak hours are obtained, which are denoted by $Sp_a^{b_i}$ and $Dp_a^{b_i}$, respectively, and they are computed as

$$Sp_a^{b_i} = \frac{1}{|\mathcal{P}_i|} \sum_{t \in \mathcal{P}_i} S^{b_i}(t), \tag{6}$$

$$Dp_a^{b_i} = \frac{1}{|\mathcal{P}_i|} \sum_{t \in \mathcal{P}_i} D^{b_i}(t), \tag{7}$$

where $\mathcal{P}_i$ denotes the set of peak hours for BS b_i.

Algorithm 1 Agglomerative hierarchical clustering.

```
Input: Base stations number $N$, Threshold value $T$, Traffic data $D_i$, for $i = 1, 2, 3...N$
Output: Labels $L_i$, for $i = 1, 2, 3...N$
Initialize :
$a \leftarrow 0$, $b \leftarrow 0$, $m \leftarrow 0$, $M \leftarrow 0$, $n \leftarrow N$
$c_k \leftarrow [D_k]$ for $k = 1, 2, 3...N$ //Add $D_k$ in the $k$th cluster.
$dist \leftarrow 0$, $Min\_dist \leftarrow Inf$ //$dist$ is between-cluster distance for each two cluster, and $Min\_dist$ is its minimum.
$stop \leftarrow false$
while $stop == false$ do //Find all possible clusters $C$
    $Min\_dist \leftarrow Inf$, $m \leftarrow m + 1$
    $C[m] \leftarrow [c_1, c_2...c_N]$ //$C[m]$ is possible cluster set in this loop
    for $i = 1$ to $n$ do
        for $j = i + 1$ to $n$ do
            $dist \leftarrow compute\_distance(c_i, c_j)$
            if $Min\_dist > dist$ then
                $Min\_dist \leftarrow dist$
                $a \leftarrow i$, $b \leftarrow j$
            end if
        end for
    end for
    $n \leftarrow n - 1$, $c_a \leftarrow merge(c_a, c_b)$
    $C[m].delete(c_b)$, $sort(C[m])$
    if $n == 1$ or $Min\_dist > T$ then
        $M \leftarrow m$
        $stop \leftarrow true$
    end if
end while
$C_{opt} \leftarrow find\_max(SH(C[i]))$ //Compute the Silhouette criterion for each $C[i]$ and find the optimal $C_{opt}$ with highest value.
for $c_i \in C_{opt}$ do //Return label $L_k$ of $C_{opt}$
    for $\forall D_k \in c_i$ do
        $L_k \leftarrow i$
    end for
end for
Return $L$
```

Clustering of BSs are based on $Sp_a^{b_i}$ and $Dp_a^{b_i}$ of each BS. Since the best number of clusters is unknown, we choose the agglomerative hierarchical clustering algorithm [9]. Each BS is regarded as a vertex $\mathcal{V}_i$ with vertex value v_i (the values assigned to vertices will be explained later). Each cluster is a set of vertices (BSs), denoted by C_n. The distance between two vertices $\mathcal{V}_i$ and $\mathcal{V}_j$ is $d(\mathcal{V}_i, \mathcal{V}_j) = |v_i - v_j|$. Using the average linkage criterion, the distance between two clusters C_m and C_n is measured as the average distance between vertices in C_m and vertices in C_n, i.e., $d(C_m, C_n) = \frac{1}{|C_m||C_n|} \sum_{V_i \in C_m, V_j \in C_n} d(V_i, V_j)$. Agglomerative hierarchical clustering starts by considering each vertex as a cluster. During each iteration, it calculates the distances between all pairs of clusters and merges the two clusters with the minimum distance into one cluster. The clusters continue merging until all vertices are included in one cluster. In this way a hierarchical dendrogram is generated. In the next step, Silhouette criterion [10] is used to decide where to cut the hierarchical dendrogram in order to get the best separation among vertices. The details are shown in Algorithm 1.

The cluster process is performed in two rounds both by Algorithm 1. In the first round, each vertex is assigned the value $Sp_a^{b_i}$ of the BS. The first round divides the BSs into two subscriber-density-based clusters, C_1 and C_2, corresponding to low subscriber density (sparse) and high subscriber density (dense). The second round goes inside C_1 and C_2, respectively. Each vertex is assigned the value $Dp_a^{b_i}$ of the BS. In both C_1 and C_2, three average-demand-based sub-clusters are found, with low (light) demand, medium demand, and high (heavy) demand, respectively. Finally, combining the two rounds of clustering process, we obtain six clusters. They are identified as DH (dense heavy) cluster, DM (dense medium) cluster, DL (dense light) cluster, SH (sparse heavy) cluster, SM (sparse medium) cluster, and SL (sparse light) cluster.

B. Analysis of Clustering Results

In this part, we further analyze the clustering results, and reveal the relationship between BS clusters and their geographical functional regions.

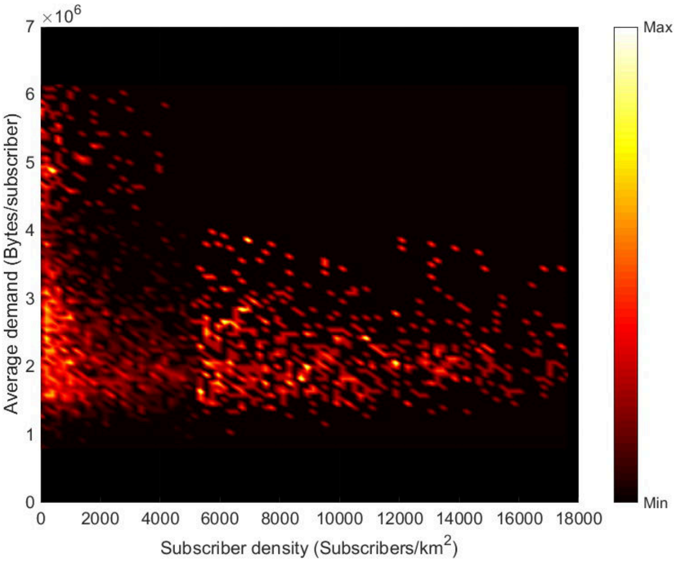


Fig. 5. Clustering result in the space of subscriber density and average demand, in the view of density.

The clustering results of BSs are shown in Fig. 5, where each BS is mapped onto the (S^{b_i}, D^{b_i}) space. Rather than plotting in scatter form, we depict the number of BS samples in the unit area of the (S^{b_i}, D^{b_i}) space, i.e., density, by the brightness of color bar. As can be observed in this map, we can clearly identify 6 brightest areas each of which represents the centroid of each cluster.

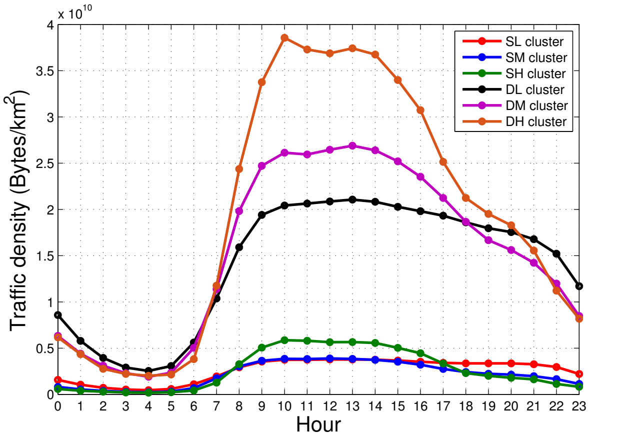


Fig. 6. Temporal dynamics of average traffic density in each cluster. Each curve represents the average 24-hour dynamics of one BS cluster.

The six clusters exhibit different characteristics, in terms of network capability. Fig. 6 presents the temporal dynamics of average traffic density in each cluster. It can be seen that all the six clusters have peak hours when the traffic density is high, but there are significant differences on the time and duration of peak hours as well as th peak values among different clusters. Further analysis on the hourly dynamics of subscriber density and average demand in each cluster reveals that subscriber density and average demand also show identical dynamic patterns in each cluster. Another interesting observation is that subscriber density rises and reaches its peak hours during daytime, and then drops to low values at night, while average demand rises during daytime and does not drop to low values until midnight. This reveals that although subscriber density tends to be low at night, subscribers still consume much data traffic during night. These insights indicate that different peak and non-peak time periods are needed to characterize the hourly dynamics of subscriber density and average demand in each cluster, which is discussed in the next section.

TABLE III
DISTRIBUTION OF BSS IN DIFFERENT CLUSTERS AND FUNCTIONAL REGIONS.

Types of BS	SL	SM	SH	DL	DM	DH	Total
Resident	381	40	3	3	106	1	534
Transport	9	46	1	9	1	1	67
Office	481	340	103	85	145	23	1177
Entertainment	82	48	11	18	45	4	208
Comprehensive	359	109	13	30	170	4	685
Total	1312	583	131	145	467	33	2671

Table III shows the numbers of BSs in each cluster and each functional region. A highly asymmetric characteristic is observed: almost 50% of BSs base are in the SL cluster (low subscriber density and low average demand), which is consistent with the right-skewed log-normal mixture distribution we mentioned previously.

To further investigate the relationship between the (S^{b_i}, D^{b_i}) profile and the geographical location of BSs, several parameters are defined to measure the relationship between the six clusters and the functional regions. Let $N_{m,n}$ denote the number of BSs in the mth functional region and the nth cluster. Then $\sum_{j=1}^{6} N_{m,j}$ is the total number of BSs in the mth functional region. The mapping relation, $P_{m,n}$, is defined to represent the proportion of the nth cluster of BSs in the mth region, which is given by

$$P_{m,n} = \frac{N_{m,n}}{\sum_{j=1}^{6} N_{m,j}}. \tag{8}$$

We will use this parameter in the next section when we use our model to generate synthetic BSs in different urban areas. Table IV shows the mapping relation $P_{m,n}$. It is clear that the SL cluster is the main cluster in every functional regions, except for transport region, where the SM is the main cluster.

TABLE IV
MAPPING RELATION (%).

Types of BS	SL	SM	SH	DL	DM	DH
Resident	71.35	7.49	0.56	0.56	19.85	0.19
Transport	13.43	68.66	1.49	13.43	1.49	1.49
Office	40.87	28.89	8.75	7.22	12.32	1.95
Entertainment	39.42	23.08	5.29	8.65	21.63	1.92
Comprehensive	52.41	15.91	1.90	4.38	24.82	0.58

In order to reveal the differences among various functional regions, the relative proportion $P'_{m,n}$ is defined as follows. First compute

$$R_{m,n} = \frac{N_{m,n}}{\sum_{i=1}^{5} N_{i,n}}, \forall m, n. \tag{9}$$

Then

$$P'_{m,n} = \frac{R_{m,n}}{\sum_{j=1}^{6} R_{m,j}}. \tag{10}$$

The definition of relative proportion $P'_{m,n}$ eliminates the differences in the absolute numbers of BSs in different clusters, and thus it characterizes different patterns of subscriber density and demand in different functional regions. The values of $P'_{m,n}$ are listed in Table V. Compared with Table IV, several different observations are made:

- In office regions, the relative proportion of SH cluster is the highest, followed by DH cluster. This indicates that office regions tend to handle heavier average traffic demand by each subscriber.
- In entertainment regions, the relative proportion of DM cluster is the highest, followed by DH cluster, indicating that entertainment regions handle larger numbers of subscribers and relatively heavier demand. This is consistent with our common sense that entertainment regions tend to have denser population and that people consume more data traffic in entertainment regions.
- In comprehensive regions, the relative proportion of DL cluster is the highest, followed by SL cluster, which indicates that the subscriber density in comprehensive regions is relatively higher than that in other regions.

TABLE V
RELATIVE PROPORTION (%).

Types of BS	SL	SM	SH	DL	DM	DH
Resident	44.01	10.40	3.47	34.40	3.14	4.59
Transport	3.65	41.99	4.06	1.14	33.03	16.13
Office	11.01	17.51	23.61	9.32	17.61	20.93
Entertainment	10.96	14.43	14.72	16.89	21.76	21.25
Comprehensive	21.86	14.93	7.93	29.08	16.53	9.68

The above analysis shows how different clusters are related with geological regions. Since different regions have different proportions for each cluster (measured by the mapping function $P_{i,j}$), it is obvious that these regions have different characteristics in subscriber density and average data demand, and thus they have different network capabilities. However, when comparing different urban areas, the network capability is similar in the same functional region, which means that the proportion of each BS cluster in a certain urban functional region, i.e., the mapping relation, can be treated as constant for different urban areas. This observation will be utilized in the next section to model the network capability in urban areas.

V. NETWORK CAPABILITY MODELING

To build an accurate model for urban network capability, both traffic density dynamics $T^{b_i}(t)$ and numbers of synthetic BSs need to be consistent with those of real BSs. Recall that BSs belonging to different types (DH, DM, DL, SH, SM and SL) have their own characteristics in terms of the number of access subscribers during a certain period (subscriber density $S^{b_i}(t)$) and traffic volume they consumed (data demand $D^{b_i}(t)$). Meanwhile, our analysis in Section IV reveals the relationship between network capability and geographical context of BSs, i.e., mapping relation $P_{m,n}$. More specifically, this relationship is consistent in different urban areas. Furthermore, in a certain type of BS, both $S^{b_i}(t)$ and $D^{b_i}(t)$ have distinct dynamics during different time periods. Thus, to obtain a fine-grained model, it is necessary to take into account all the above considerations/observations.

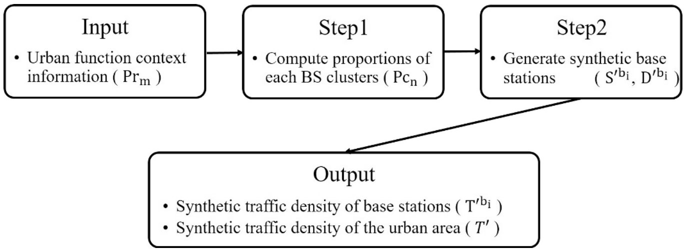

Input
• Urban function context information (Pr_m)
Step1
• Compute proportions of each BS clusters (Pc_n)
Step2
• Generate synthetic base stations (S'^{b_i}, D'^{b_i})
Output
• Synthetic traffic density of base stations (T'^{b_i})
• Synthetic traffic density of the urban area (T')

Fig. 7. Modeling Methodology.

The schematic of our model is illustrated in Fig. 7. The idea is to build a capability model which can generate different types of synthetic BSs in terms of user density $S^{b_i}(t)$ and average data demand $D^{b_i}(t)$, i.e., DH, DM, DL, SH, SM and SL. Thus the first step is to compute the proportion of each BS type in a given urban area with the input of urban function context information, as given in Section V-A. Next for each type of BSs, a certain number of synthetic BSs are generated, as discussed in Section V-B. Then we can obtain the dynamics of traffic density in the whole urban area. Moreover, We conduct an evaluation on the accuracy of our model compared to the original empirical data, in Section V-C.

A. Base Station Proportion Computation

Building a model of network capability in the given urban area first requires us to generate different BS types. We compute the proportion of each BS type, i.e., the probability used in synthetic BS generation, which is denoted as Pc_n for $1 \leq n \leq 6$, corresponding to DH, DM, DL, SH, SM and SL, respectively. The input is the proportion of the BSs deployed in different urban functional regions, denoted as Pr_m for $1 \leq m \leq 5$, corresponding to resident, transport, office, entertainment and comprehensive regions, respectively. Recall that we already obtain the mapping relation $P_{m,n}$, i.e., the proportion of the nth type of BSs in the mth region, which are listed in Table IV. Pc_n can be computed as follows

$$Pc_n = \sum_{m=1}^{5} Pr_m \times P_{m,n}. \tag{11}$$

Then we generate a set S of synthetic BSs with $|S| = |B|$. Each type of synthetic BSs in S have the same proportion as those in the original BS set B. The details of synthetic BS generation are given next.

B. Synthetic Base Station Generation

In Section III, we use the log-normal mixture model to fit the spatial distributions of subscriber density $S^{b_i}(t)$ and data demand per subscriber $D^{b_i}(t)$ in the original data. We also show that this model is universal in different spatial scales (Fig. 3). Thus when generating the synthetic spatial distributions of $S^{b_i}(t)$ and $D^{b_i}(t)$ in each BS type, we naturally choose this log-normal mixture model.

In our model, we focus on the one-day dynamics of $S^{b_i}(t)$ and $D^{b_i}(t)$. Thus we compute the average one-day sequences in whole month, denoted as $S_a^{b_i}(t_h)$ and $D_a^{b_i}(t_h)$, as

$$S_a^{b_i}(t_h) = \frac{1}{31} \sum_{j=1}^{31} S^{b_i}(t_h + (j-1) \times 24), \tag{12}$$

$$D_a^{b_i}(t_h) = \frac{1}{31} \sum_{j=1}^{31} D^{b_i}(t_h + (j-1) \times 24), \tag{13}$$

for $1 \leq t_h \leq 24$, where $(t_h + (j-1) \times 24)$ represents the t_hth hour in the jth day. Considering the temporal correlations of both $S_a^{b_i}(t_h)$ and $D_a^{b_i}(t_h)$, we divide one day into 3 periods, denoted as $Idle$, $Busy$ and $Tail$. More specifically, we set 80% of $\max\{S_a^{b_i}(t_h)\}$ and 70% of $\max\{D_a^{b_i}(t_h)\}$ as thresholds. $Idle$ periods represent periods when both $S_a^{b_i}(t_h)$ and $D_a^{b_i}(t_h)$ are under the related thresholds, while in $Busy$ periods they are both above the thresholds. In $Tail$ periods, $S_a^{b_i}(t_h)$ are under its threshold and $D_a^{b_i}(t_h)$ are above its threshold. Note that the fourth case will not happen because the above-threshold periods of $D_a^{b_i}(t_h)$ cover those of $S_a^{b_i}(t_h)$.

When fitting the empirical distributions of $S_a^{b_i}(t_h)$ and $D_a^{b_i}(t_h)$, we use the 2-dimensional log-normal mixture model to preserve the correlation between subscriber density and average data demand. In other words, we fit $S_a^{b_i}(t_h)$ and $D_a^{b_i}(t_h)$ together. The inputs of our fitting are the 18 CDFs of $(S_a^{b_i}(t_h), D_a^{b_i}(t_h))$ for 3 types of periods and 6 types of BSs. Since the log-normal mixture fitting is already detailed in Section III, we skip the distribution fitting for $(S_a^{b_i}(t_h), D_a^{b_i}(t_h))$ for space economy reason. Finally we obtain the fitting parameters needed in the network capability model.

We now briefly describe how to generate a synthetic BS using the capability model. After obtaining the BS type, for a given hour t_h, we randomly sample a pair of values ${S'}_a^{b_i}(t_h)$ and ${D'}_a^{b_i}(t_h)$ according to the fitted distribution functions. The subscriber density ${S'}_a^{b_i}(t_h)$ and average data demand ${D'}_a^{b_i}(t_h)$ describe the network capability of this BS. By multiplying ${S'}_a^{b_i}(t_h)$ and ${D'}_a^{b_i}(t_h)$, we obtain the traffic density ${T'}_a^{b_i}(t_h)$.

C. Model Validation

We first evaluate the accuracy of our method of building model. Then we use the urban environment of *Shanghai* as a case study for modeling network capability.

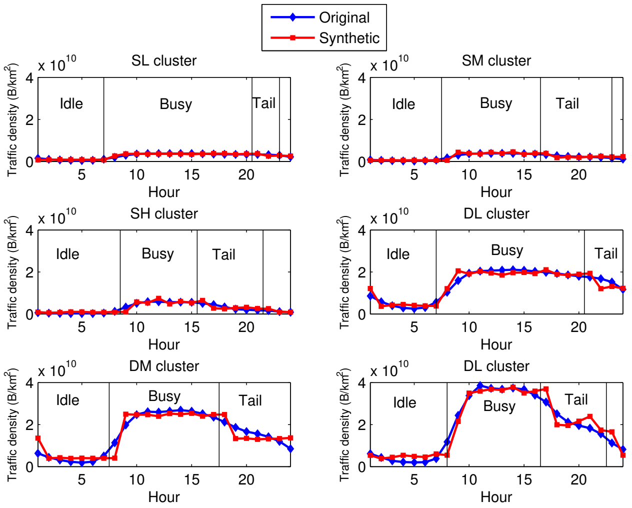


Fig. 8. Performance on modeling dynamics. Synthetic traffic density dynamics for each type of BSs are plotted, along with the original empirical ones.

1) Model Building Method Evaluation: For each type of DH, DM, DL, SH, SM and SL, we generate a set of 10,000 synthetic BSs based on the fitted log-normal mixture distribution. Each synthetic data set contains the subscriber density and average demand of BSs in a certain cluster in 24 hours. In Fig. 8, we evaluate the performance of our BS model by comparing the synthetic traffic density dynamics for each type generated by the model with the original empirical dynamics. By dividing one day into 3 periods, i.e., $Idle$, $Busy$ and $Tail$, we maintain the temporal correlation of dynamics. It can be seen from Fig. 8 that the traffic density of synthetic BSs have similar patterns to those of the original real BSs.

Next we evaluate how consistent the 6-type synthetic traffic densities are by comparing their distributions with those of the original real BSs in the set B. To this aim, we use the Bhattacharyya (BH) measure or distance [3], which quantifies the similarity between two probability distributions $p(x)$ and $p'(x)$. For discrete probability distributions, the BH measure is defined by

$$\rho(p, p') = \sum_{x \in X} \sqrt{p(x)p'(x)}, \tag{14}$$

while for continuous probability distributions, it is given by

$$\rho(p, p') = \int \sqrt{p(x)p'(x)}\, dx. \tag{15}$$

In order to satisfy all the metric axioms, we use an alternative distance metric based on the BH measure defined as

$$d(p, p') = \sqrt{1 - \rho(p, p')}. \tag{16}$$

Note that $d(p, p') = 0$ iff $p = p'$, indicating two identical distributions.

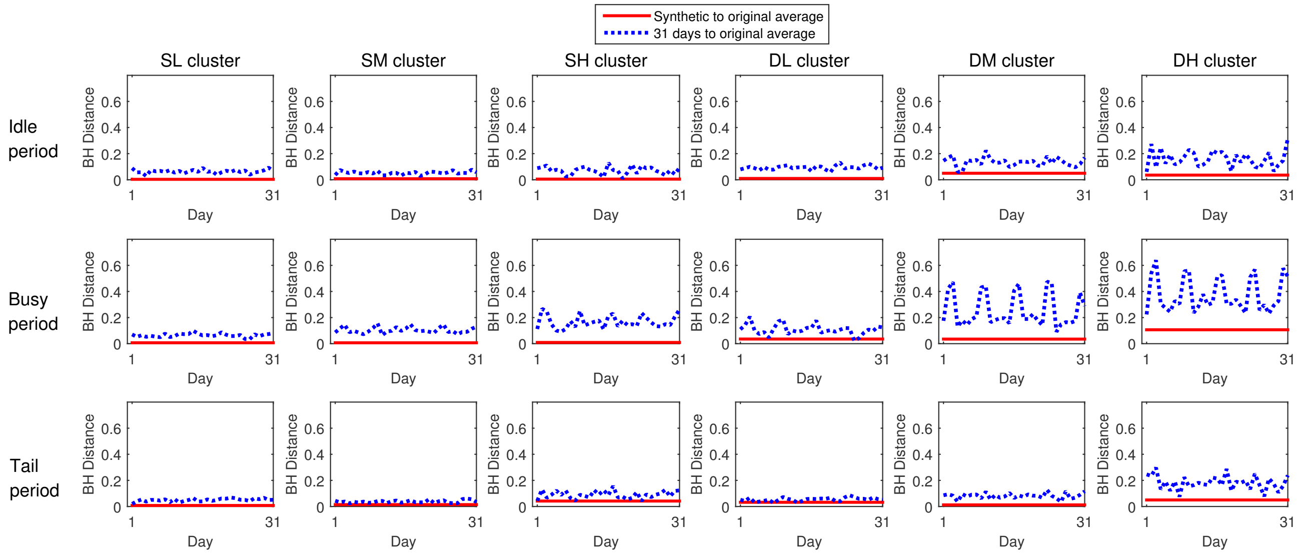


Fig. 9. Performance on modeling statistical distributions. Horizontal solid lines are Bhattacharyya distances between the traffic density distributions of the original average and the synthetic day, while dashed curves are Bhattacharyya distances between the traffic density distributions of the original average and the 31 days in the original trace.

Let $\mathbb{M}$ denote the set of 31 days in the dataset (from 1st August to 31st August), and $\mathbb{H}$ denote the set 24 hours in a day. $\mathbb{H}$ is further divided into 3 subsets, denoted as $\mathbb{H}_m$ for $m = 1, 2, 3$, which correspond to $Idle$, $Busy$ and $Tail$ periods, respectively. In the sequel, the PDF of traffic density is denoted by $p_n^{X_m}(x)$, where m represents the time period $\mathbb{H}_m$, n denotes the BS type (DH, DM, DL, SH, SM or SL), and X represents the dataset.

For each BS in type n, a synthetic dataset of subscriber density and average demand for one day is generated, based on the log-normal mixture distribution model. We then obtain the synthetic traffic density of each BS by multiplying the subscriber density and average demand. The PDFs of traffic density in the synthetic dataset are denoted as $\left\{p_n^{S_m}\right\}$, while the PDFs of average traffic density over 31 days in the original dataset are denoted as $\left\{p_n^{A_m}(x)\right\}$ and the PDFs of the original traffic density in a given day $D \in \mathbb{M}$ are denoted as $\left\{p_n^{D_m}\right\}$.

To evaluate our traffic density model, we first compute $d\left(p_n^{A_m}, p_n^{S_m}\right)$, the distance between the traffic density distributions of the averaged original data and the synthetic day S, in terms of BS types n and time periods m. Then, we compute $d\left(p_n^{A_m}, p_n^{D_m}\right)$ for $D \in \mathbb{M}$, the distance between the distributions of the averaged original data and the 31 days $D \in \mathbb{M}$ in the original trace. Fig. 9 plots $d\left(p_n^{A_m}, p_n^{S_m}\right)$ and $d\left(p_n^{A_m}, p_n^{D_m}\right)$, where there are 18 figures for 6 BS types and 3 time periods. Finally, we also compute the mean and the 95% confidence interval of $d\left(p_n^{A_m}, p_n^{D_m}\right)$. It can be verified that $d\left(p_n^{A_m}, p_n^{S_m}\right)$ is within the confidence interval of $d\left(p_n^{A_m}, p_n^{D_m}\right)$, indicating that the error of our model, i.e., $d\left(p_n^{A_m}, p_n^{S_m}\right)$, is sufficiently small.

Based on the above evaluations, we have demonstrated that our model performs well on characterizing the traffic density of BSs, in terms of both dynamics and statistical distribution.

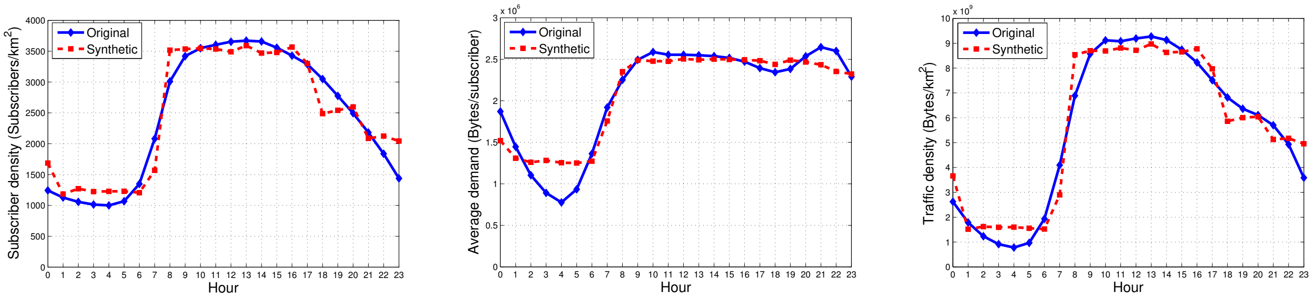
2) Case Study: The soundness of our method has been verified in the above evaluations. We now provide a case study on modeling the network capability in *Shanghai*. The input is the distribution of urban functional regions and the number of BSs deployed in each region, i.e., Pr_m listed in Table VI. Using the mapping relation $P_{m,n}$ listed in Table IV, the probabilities of 6 BS types Pc_n can be computed by (11). Then synthetic BSs' subscriber density and data demand per subscriber $(S'^{b_i}_a(t_h), D'^{b_i}_a(t_h))$ are generated. The output is the traffic density dynamics in the scales of BS and whole urban area, i.e., $\{T'^{b_i}_a(t_h)\}$ and $T'_a(t_h)$.

TABLE VI
PERCENTAGE OF BSS DEPLOYED IN EACH REGION.

Functional Regions	Index	Percentage
Resident	1	17.55%
Transport	2	2.58%
Office	3	45.72%
Entertainment	4	9.35%
Comprehensive	5	24.81%

The results are shown in Fig. 10, which plots the subscriber density, average demand and traffic density of the whole urban area, i.e., averaging over all the BSs. More specifically, we compare the synthetic subscriber density, average demand and traffic density $\{S'^{b_i}_a(t_h), D'^{b_i}_a(t_h), T'^{b_i}_a(t_h)\}$, generated by our model, with the empirical $\{S'^{b_i}_a(t_h), D'^{b_i}_a(t_h), T'^{b_i}_a(t_h)\}$, produced from the original trace data. It can be seen from Fig. 10 (a) and (b) that our model accurately describes the one-day dynamics of subscriber density and average demand. By multiplying subscriber density and average demand, we obtain the similar results in traffic density, as can be seen from Fig. 10 (c). Since traffic density represents the degree of traffic load in a cell, the traffic density plotted in Fig. 10 (c) describes the average load among all the BSs deployed in *Shanghai*.

This case study has verified the accuracy of our network capacity model. Based on this model, telecommunication operators can conduct network resources allocation, load balancing and infrastructure testing without the need of large real traffic data records, which is costly to collect. Moreover, our concepts of subscriber density and average data demand are generic which guarantees our model is universally applicable to diverse urban environments and, therefore, ensures the applicability of our model to new usage scenarios in future mobile communication. For example, urban function information of a city



(a) (b) (c)

Fig. 10. Network capacity modeling performance: (a) subscriber density, (b) average demand, and (c) traffic density. Dashed curve represents the result based on the synthetic BS set S, while solid curve represents the result based on the original BS set B.

can be obtained from the government. Using this information together with statistics $P_{m,n}$, we can generate synthetic BSs in each urban functional region.

VI. Related Work

Cellular BS traffic patterns have been extensively investigated for understanding various perspectives of cellular networks. Wang *et al.* [4] extracted the traffic patterns of large-scale BS towers by combining three dimensional information (time, locations of towers and traffic frequency spectrum) together. Similarly, Shafiq *et. al.* [11] also focused on modeling traffic patterns of BSs. Unlike the work of [4], however, the study [11] was based on the knowledge of the application distributions in each cell. As for modeling traffic load in each cell, Lee *et al.* [6] demonstrated that the spatial distribution of the traffic density can be accurately modeled by a log-normal mixture distribution, while Wang *et al.* [12] found that mobile traffic volume (not density) followed a trimodal distribution on both spatial and temporal dimensions. However, these modeling methods are only for statistical fittings, and they are not suitable for considering temporal correlations in traffic dynamics.

In addition to BS traffic modeling, there exist a few related works on modeling traffic usage patterns of mobile devices or subscribers. Shafiq *et al.* [13] proposed a Zipf-like model to capture the volume distribution of application traffic in celluar devices and then used a Markov model to characterize the dynamics. Oliveira *et al.* [3] classified subscribers into 4 profiles according to session number and traffic volume in a certain period. Then a traffic usage model was build for each profile of subscribers in peak and non-peak time periods, respectively. By contrast, our work considers higher-level capability modeling of mobile networks. For our problem of modeling network capability, we adapt Oliveira's methods of clustering.

In particular, we model the mobile network capability in the two-dimensional space of subscriber density and average data demand. Our analysis focuses on the spatial and temporal distributions of subscriber density and average data demand, which is lacked in the previous works of [4], [6], [12]. Furthermore, by decomposing traffic density into subscriber density and average data demand, we can explain Lee's observation that traffic density follows a log-normal mixture distribution [6]. To the best of our knowledge, our work is the first trial to build a network capability model.

VII. Conclusion

In this paper, we investigate the capability of mobile cellular data network in large-scale urban environment. Our investigation reveals two important discoveries. First, the spatial distribution of both subscriber density and average traffic demand in each cell can be accurately fitted by log-normal mixture model. Second, using an unsupervised clustering method, we find that large scale base stations can be clustered into 6 distinct types according to subscriber density and average traffic demand. Inspired by those two observations, we build a data network capability model and use this to generate real base stations with diverse network capabilities. Our evaluations show that the synthetic trace presents a consistent behavior with the original dataset, which demonstrates that our model is precise and flexible.

References

[1] "NGMN 5G white paper," available online at https://www.ngmn.org/5g-white-paper.html, confirmed in Nov. 2015.

[2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5g be?" *Selected Areas in Communications, IEEE Journal on*, vol. 32, no. 6, pp. 1065–1082, 2014.

[3] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute, "Mobile data traffic modeling: Revealing temporal facets," Ph.D. dissertation, INRIA, 2014.

[4] H. Wang, F. Xu, Y. Li, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," in *ACM IMC*, 2015.

[5] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan, *Statistical analysis and modelling of spatial point patterns*. John Wiley & Sons, 2008, vol. 70.

[6] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang, "Spatial modeling of the traffic density in cellular networks," *Wireless Communications, IEEE*, vol. 21, no. 1, pp. 80–88, 2014.

[7] G. McLachlan and D. Peel, *Finite mixture models*. John Wiley & Sons, 2004.

[8] R. B. D'Agostino, *Goodness-of-fit-techniques*. CRC press, 1986, vol. 68.

[9] R. R. Sokal, "A statistical method for evaluating systematic relationships," *Univ Kans Sci Bull*, vol. 38, pp. 1409–1438, 1958.

[10] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[11] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3g cellular data network," in *IEEE INFOCOM*, 2012.

[12] H. Wang, J. Ding, Y. Li, P. Hui, J. Yuan, and D. Jin, "Characterizing the spatio-temporal inhomogeneity of mobile traffic in large-scale cellular data networks," in *ACM HotPOST*, 2015.

[13] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," in *ACM SIGMETRICS*, 2011.