IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 13, 2019, accepted March 2, 2019, date of publication March 11, 2019, date of current version March 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2903277

Interest-Driven Outdoor Advertising Display Location Selection Using Mobile Phone Data

MENG HUANG[1], ZHIXIANG FANG[1,2], SHILI XIONG[3], AND TAO ZHANG[4]

[1]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[2]Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China
[3]Department of Advertising, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[4]China Mobile Group Hubei Company Limited, Wuhan 430040, China

Corresponding author: Zhixiang Fang (zxfang@whu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 41771473 and Grant 41231171, and in part by the Fundamental Research Funds for the Central Universities under Grant GK201803049 and Grant 2042017kf0235.

ABSTRACT The modern outdoor advertising industry is introducing new technologies to attract attention and enhance effective reach. However, displaying the pertinent advertisements (ads) and maximizing their coverage are still a challenge, as little is known about the interests of passersby. In this paper, we attempted to fill this gap by 1) identifying the individuals' interests extracted from the mobile phone internet usage data, and 2) analyzing the individuals' mobility patterns acquired from the mobile phone location data. Based on these data, the problem of distributing the outdoor ads to the maximum reach of the potentially interested users was further formulated as a maximal covering location problem (MCLP). Specifically, a continuous space maximal coverage model, the MCLP-complementary coverage (MCLP-CC), was utilized to search for the optimal locations for a given category of advertisements. Finally, the proposed methodology was applied to Wuxue, a city in central China, as a case study. The calculation results show that when setting the same number of places for outdoor advertising, our approach achieves an average of 69% improvement compared with the current dominant ads placement approach (selecting the most crowded location) in reaching the potential target audiences. An average of 15% improvement compared with the classic MCLP model. Overall, this paper demonstrates the value of the mobile phone data for interest-driven outdoor advertising and its potential applications in other mobile services.

INDEX TERMS Interest-driven outdoor advertising, maximal covering location problem, mobile phone data.

I. INTRODUCTION

Despite the fact that the effectiveness of outdoor advertisements (ads) is unsteady from case to case, the use of outdoor advertising has grown steadily in the past years. Since 2010, the global spend on outdoor advertising has increased from $25.54 billion to $33.32 in 2017 [1]. In spite of the increasing spend in outdoor advertising, its effectiveness has been doubted. Outdoor advertisements are making every attempt to compete against floods of information, trying to attract attention from passersby, yet majority of them are ignored [2] due to the lack of interests or personal relevance. On the other hand, interest-driven advertising has proven effective in increasing advertising effectiveness (e.g. the level of attention, and perceived usefulness of information, etc.) [3]–[5]. As a result, advertisers now are inclined to create interest-driven contents for maximal efficiency. However, it is still a challenge to succeed in delivering interest-driven ads to passersby, since it is difficult to acquire detailed interest-related data from potential audiences. Traditional ways for outdoor advertising audience measurements have relied on customer surveys or circulation counts from the Traffic Audit Bureau [6]. But these methods are either dependent on subjective memory that could potentially be distorted due to time distance [7], or limited to only a few professional agencies [6]. The process is also costly, time consuming, and less flexible. Besides, even though some new datasets, such as GPS data and smart card data, can provide more precise traffic data that could be applied to outdoor advertising [7], [8], displaying the right content which coincides with audience's

The associate editor coordinating the review of this manuscript and approving it for publication was Qingchao Jiang.

interest remains challenging, because these data usually do not contain audience's product-related interest or intention information but only provide mobility trajectory. Currently, few tools exist to help marketers precisely target advertising to certain groups. In this paper, we attempt to fill this gap by combining information from two types of mobile phone data: mobile internet usage data and mobile phone location data. The pervasive use of mobile phones enables us to have a complete understanding of both user mobility and interest at a fine scale compared with traditional methods. In this study, people's interests are extracted from mobile internet usage, combined with individuals' mobility patterns acquired from mobile phone location data. For each advertising category, the optimal places to distribute relevant ads are selected using a continuous space maximal covering location problem-complementary coverage model (MCLP-CC) [9]. The results show that we are able to reach a higher portion of interested users using MCLP-CC model compared to classic MCLP model and advertisers' current siting approach.

II. RELATED WORK

A. TRAJECTORY DATA FOR OUTDOOR ADVERTISING

Outdoor advertising plays an important role in reaching the increasingly mobile consumers [8]. There are various outdoor advertising forms, ranging from billboards, digital screen to scrolling panels, street furniture etc. In this paper, we generically refer to them as ''panels''.

There are many factors associated with the success of outdoor advertising, such as location, content design, visibility, among which location is the most important one [10]. There is no doubt that the appropriate locations will significantly increase the target audience exposure. The target audiences are individuals whose interests match the category of ads. Traditional ways for audience measurements have relied on traffic data or rough demographic data [6], [11]. But this kind of data is often coarse-grained and cannot accurately represent daily flows. The lack of validated data for audience measurement is one of the main bottlenecks of outdoor advertising, which has impeded the development of outdoor advertising.

Thanks to the recent advancements of location-aware technologies, a mount of new geo-located data is available now, such as mobile phone data, GPS data and so on, which provides the possibility to solve this problem. As a result, there are a growing number of studies using trajectory data for outdoor advertising in recent years [10], [12]. For example, Liu *et al.* [10] combined billboard location selection and visualization together using large-scale taxi GPS trajectory data. Zhang *et al.* [12] proposed a trajectory-driven model for billboard placement. But few studies combine the user interest and the outdooring advertising. The existing studies aimed at targeted outdoor advertising are almost dependent on social media data or the social events people have likely attended [6], [11], [13]. However, Twitter data cannot accurately represent the whole population and often over represent young adult population [11], [14]. On the other hand, the pervasive use of mobile phones across different age/gender/cultural groups indicates that the mobile phone data is able to provide more information than just within a type of population, and thus enables us to access more generalized users. As a result, mobile phone data is more suitable to be used for interest-driven outdoor advertising.

B. ADVERTISING LOCATION SELECTION AND MAXIMAL COVERING LOCATION PROBLEM (MCLP)

Current interest-driven outdoor advertising placement methods either used the fixed existing billboards [15], or partitioned the study area into grids and selected the top-k grids as the optimal places [13]. In this study, we are aiming at providing a sub-grid method with higher resolution for outdoor ad placement based on target audiences' mobility. We assume that the costs for advertising are the same at different places. Hence, the problem is to find a set of locations to display the ad of category i that would maximize the exposure to the total number of users who are interested in i. From the perspective of location theory, it can be solved based on the maximal covering location problem (MCLP). The digital screens or other forms of outdoor advertising can be seen as facilities providing certain contents as a service to people around them. The target audiences distributed in the study area can be seen as the demands that need to be served or ''covered''. Those who are inside the advertisement's influence range would be declared as covered, since they have the opportunity to see it. The MCLP was first introduced by Church and ReVelle [16] in 1974 . It aims at maximizing the population within the service area of a limit number of facilities. In the MCLP, the demand and the candidate facility sites are both represented as a finite set of discrete points, which are known in advance [16]. The MCLP has achieved many successful applications, such as suggesting the bike sharing stations locations [17], siting fire stations [18], supporting air pollutant monitor [19].

The MCLP is appropriate when facilities and demands are discrete and pointe-based. However, it is not suitable to use it in our case. On the one hand, the ads panels can be located everywhere in the study area rather than a finite number of locations known in advance as assumed in the MCLP. On the other hand, the target audiences flow is continuous distributed in the study area. The target audiences are not able to be easily represented by serval discrete points, e.g. the locations of mobile phone towers. When a mobile phone tower is within the service area of a panel, it does not necessarily mean that the entire population served by this tower would be covered by the ads due to the large service area of the mobile phone tower. The coverage in the MCLP is binary, which means a demand point unit is either entirely covered or not covered at all. But in reality, many demand zones may only be partially covered. In using the binary coverage assumption, partially covered demand zones will be ignored or assumed to be covered completely as a solution. Many studies have demon-

strated that when dealing with continuous location problem using the MCLP and ignoring partial coverage, significant errors and bias could be introduced into facility location selection [9], [20], [21].

As a result, the MCLP evolves from discrete point-based representation to continuous space MCLP to meet the requirements of more realistic problem situations. Demands are allowed to be represented in more forms, namely lines, polygons. A few MCLP extensions, including MCLP-explicit [22], MCLP-implicit [23], [24], MCLP-complementary coverage (MCLP-CC) [9], which are all allowed polygon-based demand representation, have been developed. Among these models, the MCLP-CC is considered as the most competitive approach, because it is able to achieve the largest coverage with high computation efficiency [25]. Different from the MCLP, the MCLP-CC represents the demand as polygon-based units. It improves the binary specification of coverage by accounting for partial coverage of demands, so it realistically reflects the overall amount of coverage a demand unit receives [9]. Another difference from the MCLP is that the candidate sites for facility placement are not known in advance, instead, the MCLP-CC allows candidates to be sited anywhere in a continuous space. To reduce the search efforts in a continuous space for candidate sites, the polygon intersection point set (PIPS) is often used as a strategy. It is a special set of limited critical locations that can cover more polygons than others, so optimality will not be sacrificed by using it. The infinite potential facility locations are reduced into a finite point set in this way. It also has been proven that the PIPS contains at least one optimal solution for polygon-based demand coverage problem [26]. The MCLP-CC model has been successfully used in siting emergency warning sirens [9], fire stations [25], but has not been applied in advertising. To the best of our knowledge, this is the first work combining mobile phone data and the MCLP-CC model for interest-driven outdoor advertise locating.

III. STUDY AREA AND DATASET

Wuxue is a county-level city in Hubei province, China. According to recent census, Wuxue's resident population is about 300,000 and the city's total area is nearly 1246 km^2 [27]. It is divided into 12 subdistricts, among which Wuxue subdistrict is the center of the city. In total, there are 390 mobile phone towers in Wuxue. As shown in Fig. 1a, the center of the city, Wuxue subdistrict has higher tower densities. The average nearest distance among the mobile phone towers in Wuxue subdistrict is 76 meters. The distance is 984 meters in other areas. Since Wuxue subdistrict has better tower coverage and the majority of people are located there, therefore we decided to focus on this area as a demonstration of our model implementation.

The mobile phone data set used in this study was provided by a major telecom operator in Wuxue. It was comprised of two parts for a time span of 20 days (from 10th of

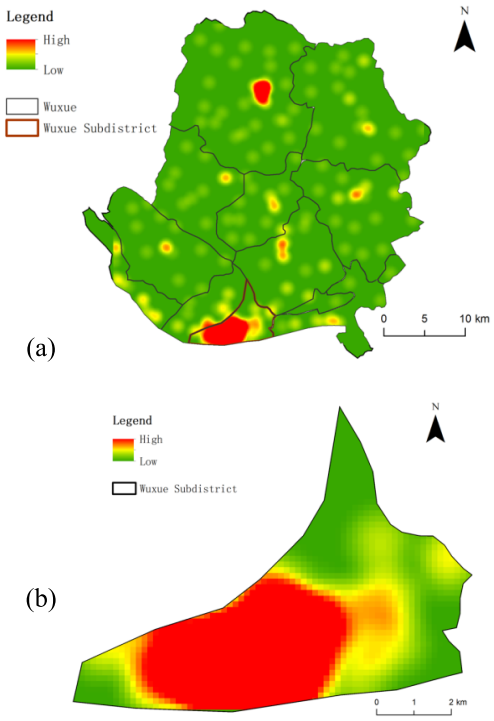


FIGURE 1. Spatial kernel density of the distribution of mobile phone towers in (a) Wuxue; (b) Wuxue subdistrict.

TABLE 1. Sample records of mobile location data.

User ID	Date	Time	Base	Longitude	Latitude
58**	2015-08-10	07:32	11*	115.****	29.****
58**	2015-08-10	08:32	11*	115.****	29.****
⋮	⋮	⋮	⋮	⋮	⋮
58**	2015-08-10	09:27	12*	115.****	29.****

August 2015 to 29th of August 2015). One is the user location data. Every time a user had a mobile phone activity, such as a call, a message, it would be recorded. Besides, when a user did not have any mobile phone activity more than one hour, his location would be actively updated. Table 1 shows an example of an individual's mobile phone location records. Each record comprises an anonymous user ID, recording time, the ID of the mobile phone tower, the latitude and longitude of the mobile phone tower used. This data set consists of a total of 431,928 users.

The other data set is mobile internet usage data. Users' mobile web browsing records include page view counts for different types of mobile apps and websites, the timestamp, the base ID, the traffic consumed. Table 2 gives an example of this data. This data set consists of 152,952 users.

By constraining users to those having both at least one mobile internet usage record every day and location data every day during the whole time span, we end up with 25,568 total users.

TABLE 2. Sample records of mobile internet usage data.

User ID	Date	Time	Base	App	Flow
36**	2015-08-10	08:34	13*	WeChat	0.0107
36**	2015-08-10	08:43	15*	WeChat	0.0017
⋮	⋮	⋮	⋮	⋮	⋮
36**	2015-08-10	11:30	33*	Taobao	1.32

IV. METHODOLOGY

A. OVERALL FRAMEWORK

As analyzed above, interest-driven outdoor advertising using mobile phone data is a continuous maximal covering location problem and candidate panels are allowed to be sited anywhere in a continuous space. The methodological framework of panel siting using the MCLP-CC model is shown in Fig. 2.

The procedure includes three main steps:

1) Discretize the study area into a series of discrete polygon-based object demands and infer users' interests. For a given advertising category, identify its target audiences based on users' interests. Calculate average daily flow of target audiences in each demand object and use it as the weight for each demand. Then estimate the number of people with different interests in each demand object as weights for demand object.
2) Generate refined PIPS to identify potential locations for outdoor advertising.
3) Establish the MCLP-CC model, and solve it by optimization software Gurobi. Evaluate the performance of the model.

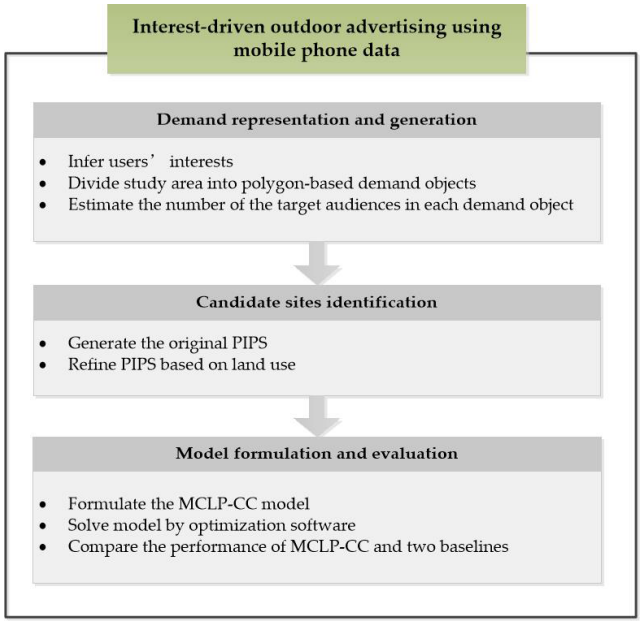


FIGURE 2. Framework of the interest-driven advertising using mobile phone data.

The following sections are organized according to the procedure in Fig. 2, explaining each step of the whole framework.

B. DEMAND REPRESENTATIONAND GENERATION

In location problem, the demand refers to objects that need to be served or covered. In our case, for a given advertising category, the demands are the target audiences, whose interests match the category of ads. Hence, our first goal is to infer users' interests. In this paper, we have a fixed set of ten categories to which both ads and user interests conform: maps and travel, video, music, social networking, game, business, finance, reading, online shopping, news.

The mobile internet browsing history contents are the direct indicator of users' interests. They were used to obtain information about users' degree of interest in these ten categories. For a better understanding of user interests, an application category dictionary was built to classify the applications. Apps and websites were manually categorized into these ten groups according to their function [28]: (1) social networking includes apps like WeChat or Sina Weibo; (2) finance includes mobile banking apps or stock trading apps; (3) maps and travel, e.g. Gaode maps or Baidu maps; (4) games such as angry birds, block puzzle; (5) video consists of Youku, iQiyi, which are popular online video platforms in China; (6) reading, e.g., Sina reading; (7) online shopping includes Taobao or Amazon apps; (8) music contains popular online music apps, such as QQ Music, Xiami Music; (9) business includes email apps, cloud storage apps; (10) news such as Tencent News, TopBuzz. Table 3 shows the number of apps and websites in each category.

TABLE 3. App category.

Category	apps
social networking	52
finance	38
maps and travel	37
game	37
video	34
reading	30
music	29
online shopping	28
news	26
business	26

Inspired by VSM (Vector Space Model) [29], a user's interest was described in the form below.

$$\{(i_1, w_1), ...(i_i, w_i), ...(i_{10}, w_{10})\} \quad (1)$$

where i_i is a certain kind of interest of the user, corresponding to a certain type of mobile phone content (e.g., online shopping, reading) through mobile apps or websites; w_i is the weight of the corresponding interest.

The most direct and simplest way to define a user's interest weight is to assign it based on the consumed traffic or time of this category. But it may result in biased results. Since in general, people tend to use social networking apps more than other categories. If we assign the weight only based on the consumed traffic or usage counts, then the majority of people's interests will be defined as social networking,

which is far from reality. To overcome this problem, TF-IDF (Term Frequency-Inverse Document Frequency) was applied to identify user interest.

TF-IDF is a widely used approach in information retrieval to reflect the importance of a word in a collection of documents. It calculates the weight of a feature term in whole document set according to its term frequency (TF) and inverse document frequency (IDF). The final weight is the product of TF and IDF [6].

In this study, we analogize the problem of finding users' degree of interest in ten categories of contents to the problem of discovering the importance of a word to a document. We regard a user as a document and an app category as a word. The original TF is the ratio number of current word to the number of all words in document. In our context, it is the ratio number of a category's traffic to all traffic a user consumes. IDF (Inverse Document Frequency) is defined in logarithmically scale, as a measure of whether a term is common or rare across all documents. In our context, it is used to reflect whether an app category is popular or rarely used among all users. Finally, TF-IDF is calculated as the product TF and IDF. In sum, the weight of an interest is proportional to the traffic a user consumes of this category's internet content and inversely proportional to the traffic of all users using such apps.

$$TF = \frac{t_{ij}}{\sum_j t_{ij}} \tag{2}$$

$$IDF = log\left(\frac{\sum_i \sum_j t_{ij}}{\sum_i t_{ij}}\right) \tag{3}$$

$$TF - IDF = TF * IDF \tag{4}$$

where w_{ij} stands for the weight of user i of the interest j, t_{ij} is user i's consumed traffic of app category j, $\sum_j t_{ij}$ is the total consumed traffic of user i, $\sum_i t_{ij}$ is the traffic of category j used by all the users, $\sum_i \sum_j t_{ij}$ is the total traffic used by all the users.

Based on (4), each interest was given a weight. A user's dominant interest was assigned to the interest with the highest weight. Users were then classified into ten interest clusters according to their dominant interests.

After users' dominant interests were known, for a given advertising category, the corresponding target audiences were found. Then we tried to represent the target audiences, the so-called demands in location problem, in an appropriate way required by the solution method.

As mentioned in section II-B, the demand in the MCLP is often represented as points, but the target audiences are continuously distributed in the study area in reality and cannot simply be represented as the location of mobile phone towers. The point-based representation will involve some problems when the demand is continuous. Hence, we decided to use other forms to represent the demand. When dealing with mobile phone data, the study area is often split into Voronoi tessellations generated by the cell phone towers [30]–[32] or square grids [33]–[35]. But some cell phone towers are located very close to each other, which can cause frequent signals jumps between the towers. When estimating the number of mobile phone users by Voronoi partitions, the error was much larger than expected [33]. Therefore, Voronoi partition method is not a suitable approach. We split the area under analysis into a number of contiguous square pixels. The influence of signal switches between close cell phone towers can be reduced in this way [34]. As to the grid size, we use 250×250 m cells as recommended for urban areas in [35], and each cell is tagged with a unique Grid ID.

The demand was thus aggregated by each grid. To reflect a demand object's value or importance, it is often weighted based on the associated attribution such as area, population. In this case, for a given advertising category, the weight of each demand unit was the average daily traffic of the corresponding interest cluster. Irrelevant audiences were ignored. Since we aimed to deliver the ads to the users with a strong preference to the ads' contents rather than have no interest [36]. To achieve this goal, this study first identified the mobile phone towers located in each grid and then summed all the distinct target audience's records from these towers for each day. Then the average target audiences' daily traffic was calculated as the weight of each grid. For grids containing no cell phone tower, the weight will be assigned zero, because population flow could not be counted in grid cells without cell phone towers. Note that only distinct users were counted when measuring the flow. It means that even a user comes into this area serval times, it will be counted only once. The reason for this is that multiple exposures are not needed to drive purchase. Money could be wasted for repeatedly influencing the same group of audiences [37]. Within each demand unit, the flow was assumed to be uniformly distributed.

C. CANDIDATE SITES IDENTIFICATION AND MODEL FORMULATION

There are infinite candidate locations for outdoor advertising, since we assume that panels can be sited anywhere in the study area. How to narrow the continuous space to finite locations for panel placement is a problem. For polygon-based demand, the PIPS is the widely used approach to identify candidate sites. The locations of PIPS are superior to others in providing coverage to polygon-based demands, so the PIPS can reduce the infinite potential panel locations into a finite point set without sacrificing optimality.

Figure 3 is an example of PIPS. ABCD and BEFC are two demand polygons. Centered at each vertex of the demand object ABCD, we draw circular buffers using facility's service distance as the radius. The overlapping area of all vertices' buffer zones is K_{ABCD}. If a facility is located in or on K_{ABCD}, demand polygon ABCD can be covered. This area is defined as the covering boundary of the demand polygon. Through the same procedure, K_{BEFC} is generated. If a facility is located in the overlapping area of K_{ABCD} and K_{BEFC}, demand ABCD and BEFC can both be covered. Hence, the intersection points of all the demand polygons' covering boundaries are critical locations for providing coverage.

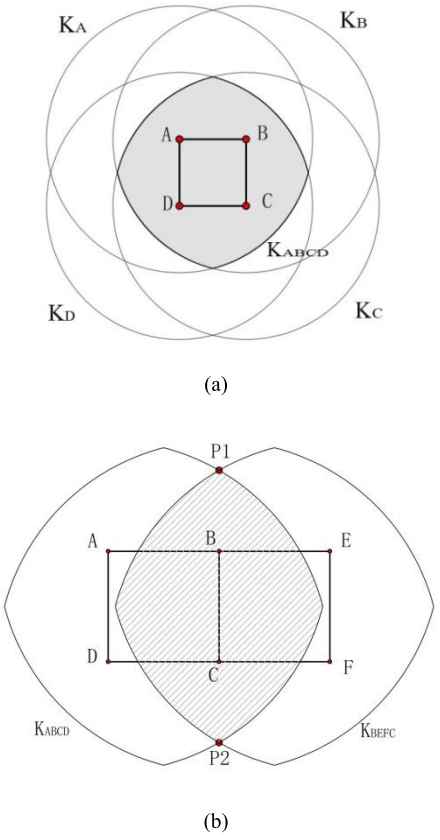


FIGURE 3. The generation of PIPS: (a) Generation of covering boundary. (b) Critical locations.

This intersection point set is defined as PIPS [26]. In the case of Fig. 3, the PIPS contains of two discrete points, namely P1 and P2.

In our context, each advertisement's influence range was represented by a certain radius circle. People inside the circle would be declared as covered, since they have the opportunity to see the advertisement. Different from other facilities, such as hospitals, schools, there is not a fix influence range of outdoor advertising. It relates to many factors, such as the height, the size of the panel and so on. For simplification, the buffer radius of a panel was set as 500 m. The PIPS for locating outdoor advertising can be generated by the following steps [26]:

(1) Extract vertices of all the demand objects, which are 250×250 m grids in our case.
(2) Centered at each vertex, draw circular buffer zones using panel's maximal service distance, which is 500 m.
(3) Extract the overlapping area of vertices' coverage circles as the covering boundary of a demand unit.
(4) Identify the intersection points of all demand objects' covering boundaries as candidate sites for outdoor advertising.

Following the steps above, the original PIPS can be generated. But in reality, some areas are not suitable for outdoor advertising, such as water bodies, wetland, grasslands etc. Hence, based on land use planning data, unfeasible positions of PIPS should be removed. The refined PIPS are potential candidate places for outdoor advertising.

D. MODEL FORMULATION AND EVALUATION

Before introducing the mathematical formulation of this problem, the following notations were introduced:

j = Index of potential panel launch sites where $j = 1, 2, \ldots$ m.

i = Index of demand units where $i = 1, 2, \ldots$ n.

w_i = The weight of demand unit i determined by target audience flow.

p = The number of panels to be located.

λ_i = The overall amount of target users covered in demand unit$i$.

b_{ij} = The amount of target users in demand unit i covered by the panel located at site j

x_j = The number of panels launched from site j

Given this notation, for each advertising category, which is the same as the interest we used to tag users, namely, maps and travel, video, music, social networking, game, business, finance, reading, online shopping, news, the model is stated as follows [26].

$$\text{Maximize} \sum_i \lambda_i \tag{5}$$

$$\text{Subject to} \sum_j x_j = p \tag{6}$$

$$\lambda_i \leq \sum_j b_{ij} x_j, \quad \forall i \tag{7}$$

$$\lambda_i \leq w_i, \quad \forall i \tag{8}$$

$$x_j \in \{0, 1\}, \quad \forall j \tag{9}$$

Objective function (5) is to maximize the total target audiences covered by panels. Constraint (6) specifies that p panels can be sited. Constraints (7) combine all the coverage to a demand unit by summing up all coverage provided by single or multiple panels. Constrains (8) specify the upper limit for the total coverage each demand object can receive. The total coverage accounted for a demand area i cannot exceed the total number of target audiences in demand unit$i$. Finally, constrains (9) impose integer requirements on site selection variables.

The ideal way to evaluate our method is to create a real advertising campaign for each category, which is beyond our ability. As an alternative, we used two baselines to compare the performance with our approach. One is using the classic MCLP model to suggest location for each category of advertisements. As mentioned in Section II, the MCLP is a point-based representation of demand approach, so the mobile phone towers are used as both demand points and candidate locations of the panels in this approach. The weight at each demand point (i.e., mobile phone tower) is derived from target audience served by this mobile phone tower. We denote this method as the interest-driven MCLP.

The second baseline directly selects the most crowed places for outdoor advertising. Using high traffic areas for outdoor advertising is a common strategy in practice [13], [37], [38]. Based on mobile phone data, towers with the maximum average daily flow were selected for advertising. In this approach, the selected locations were the same for various categories of advertisements. We name this baseline as the total flow-driven approach.

The coverage in this case is the percentage of unique members of target audiences who have had an opportunity to see the advertisement. In the advertising industry, it is also called reach, as a typical measure for the performance of an advertising campaign [37]. In this study, we use it as the indicator to describe how well the selected locations could cover the target users. A high reach indicates that a high proportion of the target audiences can be potentially reached, so the selected places are effective for outdoor advertising. We assume that flow is uniformly distributed in each grid. Reach is proportional to the total weighted coverage area of the panels. Equation (10) is the way to calculate reach.

$$r_j = \frac{\sum_i \frac{a_i}{A_i} w_i}{T_j} \tag{10}$$

where j is index of the category of ads, and r_j is the reach of the advertisement of category j. i is index of demand units. a_i is the covered area in demand unit i. A_i is the overall area of the demand unit$i$. w_i is the weight of demand unit. T_j *is* the total population of target audience.

V. ANALYSIS RESULTS

A. DERIVING AGGREGATED SPACE-TIME RHYTHM OF APP USE

First we analyzed some general characteristics of the users' app usage. We aggregated the usage count and traffic volume of each category apps. In Fig.4, it shows that the usage of different categories varies widely. Social networking is the most popular category in terms of both usage count and traffic volume. It contributes nearly half of the traffic volume of the total network and more than half of the total usage counts, whereas the least popular category, game, accounts for only one percent of the traffic volume.

The number of users who use certain type of apps was also aggregated. Different from unbalanced usage pattern, each category has similar-size users, except for game and reading. In Fig.5, it shows that social networking is the most widely used app category, followed by news, business. Note that not all users use all these ten categories of apps. If a user doesn't use a certain kind of apps, we think it indicates that the user isn't interested in contents of this category, so the related interest weight is set to 0.

We also studied the spatial and temporal usage patterns of different categories of apps. We first investigated the diversity of apps being used by subscribers in different geographic locations. It is of interest to analyze whether a category of applications is intensively used at a certain location or is

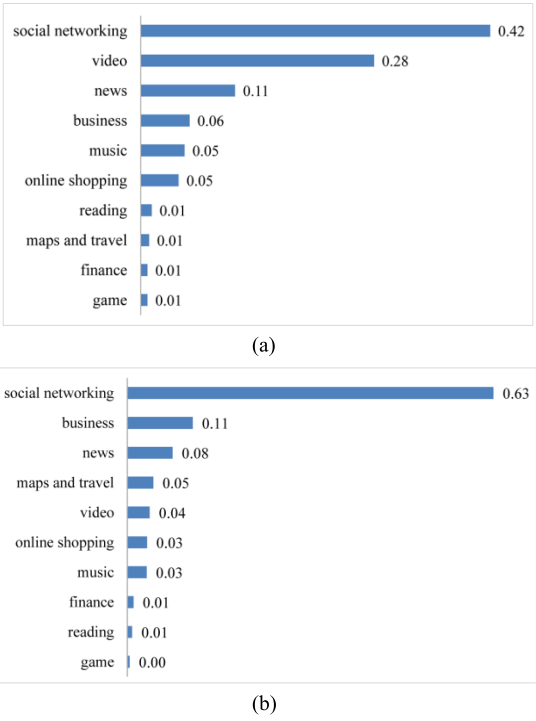


FIGURE 4. (a) Percentage of traffic contribution of each category; (b) Percentage of usage count of each category.

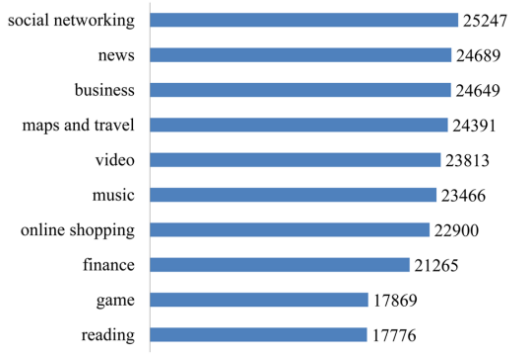


FIGURE 5. The number of users for each category.

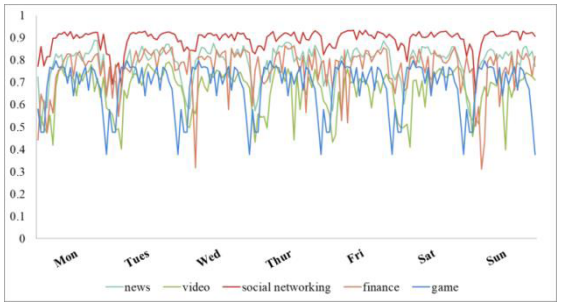
used geographic uniformly and we are also curious about the difference between different kinds of applications. Inspired by [39]–[41], the entropy was applied to measure the geographic uniformity of app usage, which is defined as follows:

$$E(i) = -\sum_{t=1}^{N} \frac{p_i^t log p_i^t}{log(N)} \tag{11}$$

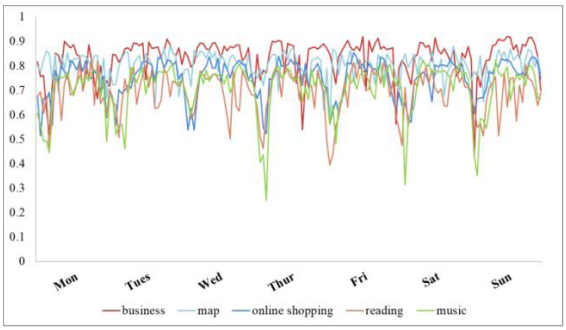
where N is the number of different locations, in this case, it is the total number of used mobile phone towers. P_i^t is the probability of the consumed traffic of app category i at location t.

Entropy is a number between 0 and 1. The bigger it is, the more geographic uniformly the app is used. On the contrary, the smaller the entropy is, the more geographic biased the usage is. Namely, the majority of a category's traffic comes

from the same region. In order to avoid the noise of the day to day random variation, we utilized the weekly average of the dataset for the analysis, which means the data point at 8:00 a.m. on Tuesday in Fig. 6 is an average of all the observations at 8:00 a.m. on Tuesday in the dataset. Fig. 6 shows the entropy of each category for every hour in the composite week. In order to clearly present the temporal dynamics of the entropy for each category, we separated the ten categories into two figures (Fig. 6a and Fig. 6b) to show the difference between the different applications.



(a)

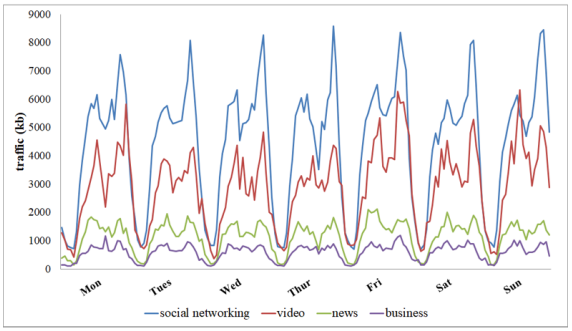


(b)

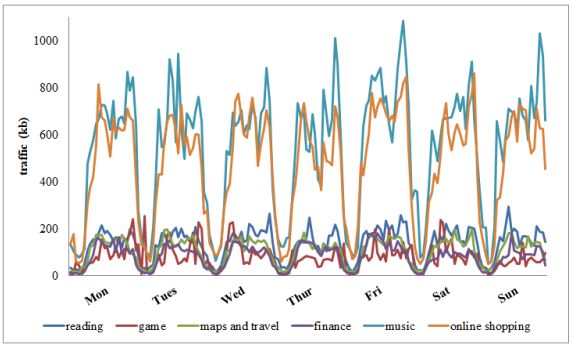
FIGURE 6. Entropy of each category: (a) news, video, social networking, finance, game. (b) business, map, online shopping, reading, music.

Both Fig. 6a and Fig. 6b show the strong diurnal characteristics in all applications. The average value of the entropy for social networking is 0.89, which is the largest among all the application categories, indicating that the usage of social networking apps is widely and uniformly distributed across different geographic locations. The average value of entropies for game and video are 0.62 and 0.66 respectively, which are smaller than any other application, implying a more biased geographic distribution. The possible reason for it is that the required traffic volume by game, video applications is usually large, so people may prefer to watch mobile video and play mobile game at the areas with better internet connections, causing clustering effects.

Then hourly variations in the traffic volume of all applications were analyzed for the weekly averaged dataset. In order to clearly present the temporal dynamics of aggregate traffic for each category, we also separated the results into two graphs to show the difference between different applications in Fig. 7a, Fig. 7b.



(a)



(b)

FIGURE 7. Temporal dynamics of aggregate traffic: (a) Social networking, video, news, business. (b) Reading, game, maps and travel, finance, music, online shopping.

We can observe a clearly daily pattern in the aggregate traffic. However, there are some differences across these applications. For instance, it shows that the traffic volumes of social networking and music reach the peaks around evening, while the peaks for the traffic volume of news and finance are in the morning. The popularity of a given category varies during the day time.

B. MODEL IMPLEMENTATION

Using the method in section IV, we grouped the users into different interest clusters according to their primary interest. Fig. 8 shows the percentage of each interest cluster. It clearly shows that the population interested in maps and travel has the highest percentage, followed by social networking, news, business; users who are addicted to game account for the least percentage.

In order to apply the MCLP-CC model, the study area, Wuxue subdistrict, was first divided into 655 demand polygons with mostly square polygons in 250 × 250 m (Fig. 9). After grouping the users into ten interest clusters, we extracted a grid cell-based average daily flow of each interest cluster. It serves as the weight for the demand objects.

In addition, the original PIPS were generated following the procedures in Section IV. Some regions are not suitable for advertising, so the PIPS located in unfeasible land use type, including water bodies, wetland and grasslands were removed based on land use planning data of Wuxue. Finally,

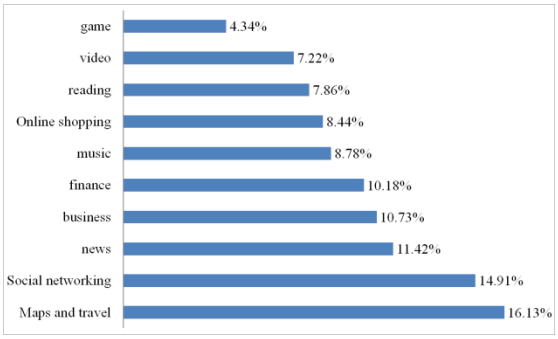


FIGURE 8. User percentage of each category.

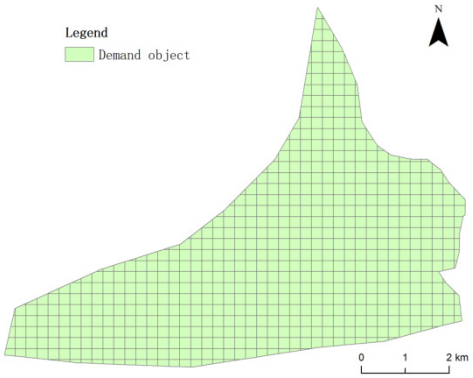


FIGURE 9. Demand objects.

7649 critical locations were identified as candidate places for outdoor advertising as shown in Fig. 10.

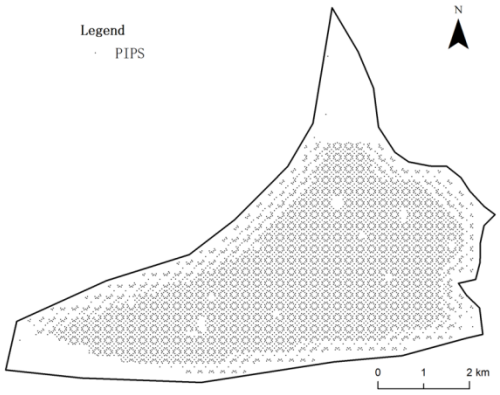


FIGURE 10. Refined PIPS.

The MCLP-CC model was formulated in Python, using the Arcpy library associated with ArcGIS. Then the optimization problems were solved using a commercial linear programming solver Gurobi. The baseline, the MCLP, was performed using the maximum coverage location-allocation model in ArcGIS 10.1

C. MODEL EVALUATION

We defined two different scenarios to compare the performance of the interest-driven MCLP-CC with the baselines.

Scenario One: for each advertising category, a fixed number of panels ($p = 5$) are sited. The reaches of various algorithms were compared. Table 4 summarizes the reach values of the solutions for all the categories.

TABLE 4. Reach of different approaches for each category.

Category	Interest-driven MCLP-CC	Interest-driven MCLP	Total flow-driven
video	58.71%	51.10%	34.96%
social	62.32%	54.84%	37.13%
reading	58.72%	51.03%	34.36%
music	64.23%	57.12%	36.75%
news	59.41%	51.01%	33.35%
maps and travel	58.58%	50.77%	35.04%
game	62.52%	55.47%	37.53%
finance	63.15%	55.79%	37.69%
business	60.03%	49.39%	36.05%
online shopping	60.42%	52.69%	36.65%

The two interest-driven approaches both show an obvious improvement compared with the total flow-driven approach, which is currently widely adopted approach by advertisers. The interest-driven MCLP-CC achieves the highest reach for all the categories, followed by the interest-MCLP method. Simulation results show that the interest-driven MCLP-CC achieves an average of 69% improvement compared with the total flow-driven approach while an average around 15% improvement compared with the interest-driven MCLP in reaching more target audience. The estimated spatial coverage areas of all the three models are shown in Fig. 11.

For every category of advertisement, each demand unit is weighted by the daily flow of the corresponding interest cluster. As shown in Fig. 11, the daily flow of various interest clusters is different in the same demand object. Interest-driven MCLP-CC has the capability of identifying the locations where the size of the target audience is the largest. As a result, more target audience can be reached by the interest-driven MCLP-CC method.

Scenario Two: varying the number of panels from 1 to 25, we compare the reaches achieved by different approaches for the categories with more users, namely maps and travel, social networking until nearly 100% reach is provided.

Fig. 12 and Fig.13 show how the total reaches based on the three models respectively grow with the increasing number of the panels. As expected, the two interest-driven approaches show the superiority of reaching the target audience for all the different numbers of panels. For the category of maps and travel, the reach achieved by interest-driven MCLP-CC ranges from 18.59%, when only one panel is utilized, to 98.41%, when 25 are sited. For the category of social

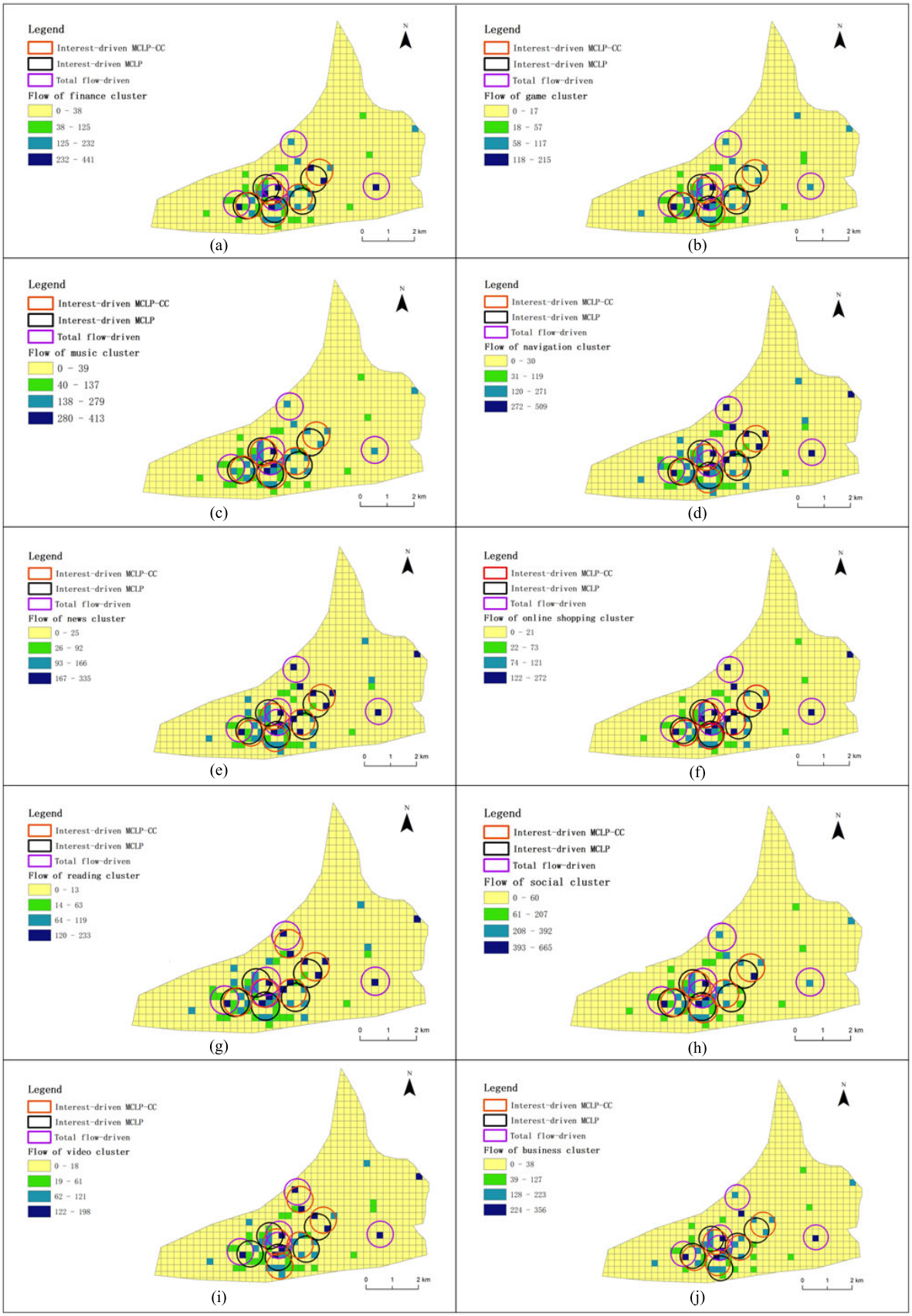


FIGURE 11. Coverageareas of interest-driven MCLP-CC, interest-driven MCLP, total flow-driven approach.

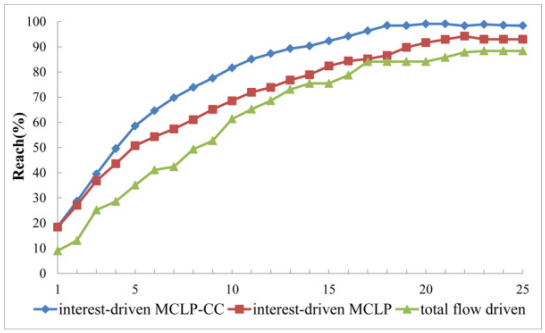


FIGURE 12. Comparison of performance between interest-driven MCLP-CC, interest-driven MCLP and total flow-driven approach for maps and travel.

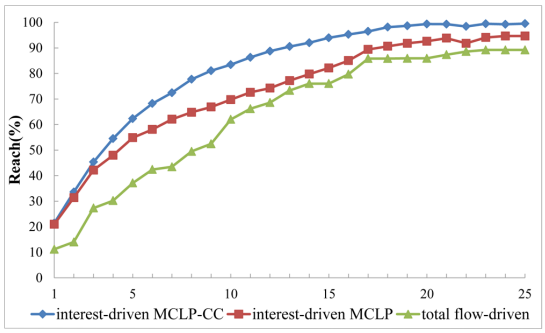


FIGURE 13. Comparison of performance between interest-driven MCLP-CC, interest-driven MCLP and total flow-driven approach for social networking.

networking, the results of interest-driven MCLP-CC ranges from 21.4% to 99.51%. In general, it shows that the solutions given by the interest-driven MCLP-CC method always outperform the other baselines no matter how many panels to be located, but the gap narrows as the number of panel increases. As to interest-driven MCLP, it outperforms total flow-driven approach for various panels.

VI. DISCUSSION AND CONCLUSIONS

Using mobile phone data in Wuxue, China, as a case study, this study demonstrates how mobile phone data can be applied to infer users' interests and provide suggestions to the locations of outdoor advertising. In sum, the main contributions of this study are as follows:

- We analyzed the fine-grained spatial-temporal dynamics of users' mobile internet usage, and proposed a method combining TF-IDF to identify users' interests. Users were then classified into ten interest clusters.
- We formulated the interest-driven outdoor advertising placement as a maximal covering location problem (MCLP), which aims at maximizing the exposure of target users. Based on the daily flow of each interest cluster, a continuous space maximal coverage model, the MCLP-CC was used to suggest the optimal places to distribute a relevant advertisement. The performance of our approach was compared with two baselines under two scenarios. The results showed that this approach can reach more "right" people.

Although we have made a successful attempt to apply mobile phone data for the interest-driven outdoor advertising locating, there are several aspects of this study that could be enhanced in future studies. First, in this study, the weight of demand object is solely depending on the flow of people with certain interest. If more data were available in the future, such as user's income, age, gender and so on, different weighting schemes can be proposed and compared for more precise marketing. Second, when locating the panels, we didn't consider the existing outdoor advertisements. If there are a high number of co-occuring ads, it will result in an overcluttering effect, and consequently, the effectiveness of advertisements will be impaired. Therefore, when positioning the outdoor advertisements in a real case, more factors should be comprehensively considered. Third, in this study, we only considered the most dominant interest and ignored users with multiple interests. However, diversity is an important nature of user interests [42]. Hence, in our following study, we are aiming at extending our current work by constructing a more comprehensive user interest profile and linking advertising contents to user interests better.

REFERENCES

[1] Statista. (2018). *Global OOH Ad Expenditure 2020|Statistic*. Accessed: Nov. 11, 2018. [Online]. Available: https://www.statista.com/statistics/273716/global-outdoor-advertising-expenditure/

[2] R. Pieters, L. Warlop, and M. Wedel, "Breaking through the clutter: Benefits of advertisement originality and familiarity for brand attention and memory," *Manage. Sci.*, vol. 48, no. 6, pp. 765–781, 2002.

[3] C. Adolphs and A. Winkelmann, "Personalization research in e-commerce : A state of the art review (2000–2008)," *J. Electron. Commerce Res.*, vol. 11, no. 4, p. 326, 2010.

[4] K. Y. Tam and S. Y. Ho, "Understanding the impact of web personalization on user information processing and decision outcomes," *MIS Quart.*, vol. 30, no. 4, pp. 865–890, 2006.

[5] M. Malheiros, C. Jennett, S. Patel, S. Brostoff, and M. A. Sasse, "Too close for comfort: A study of the effectiveness and acceptability of rich-media personalized advertising," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, May 2012, pp. 579–588.

[6] D. Quercia, G. D. Lorenzo, F. Calabrese, and C. Ratti, "Mobile phones and outdoor advertising: Measurable advertising," *IEEE Pervasive Comput.*, vol. 10, no. 2, pp. 28–36, Apr./Jun. 2011.

[7] S. Dilchert, D. S. Ones, C. Viswesvaran, and J. Deller, "Response distortion in personality measurement: Born to deceive, yet capable of providing valid self-assessments?" *Psychol. Sci.*, vol. 48, no. 3, pp. 209–225, 2006.

[8] R. T. Wilson and B. D. Till, "Effects of outdoor advertising: Does location matter?" *Psychol. Marketing*, vol. 28, no. 9, pp. 909–933, 2011.

[9] D. Tong, "Regional coverage maximization: A new model to account implicitly for complementary coverage," *Geograph. Anal.*, vol. 44, no. 1, pp. 1–14, 2012.

[10] D. Liu *et al.*, "SmartAdP: Visual analytics of large-scale taxi trajectories for selecting billboard locations," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 1–10, Jan. 2017.

[11] J. Lai, T. Cheng, and G. Lansley, "Improved targeted outdoor advertising based on geotagged social media data," *Ann. GIS*, vol. 23, no. 4, pp. 237–250, 2017.

[12] P. Zhang, Z. Bao, Y. Li, G. Li, Y. Zhang, and Z. Peng. (2018). "Trajectory-driven influential billboard placement." [Online]. Available: https://arxiv.org/abs/1802.02254

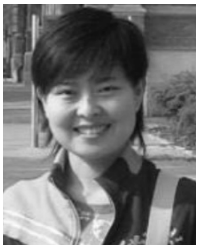
[13] A. Anagnostopoulos, F. Petroni, and M. Sorella, "Targeted interest-driven advertising in cities using Twitter," *Data Mining Knowl. Discovery*, vol. 32, no. 3, pp. 737–763, 2018.

[14] P. A. Longley, M. Adnan, and G. Lansley, "The geotemporal demographics of Twitter usage," *Environ. Planning A: Economy Space*, vol. 47, no. 2, pp. 465–484, 2015.
[15] J. Davies, S. Joseph, and S. Joseph. (Sep. 26, 2018). *Outdoor Advertising Braced for its Programmatic Moment Digiday*. Accessed: Feb. 6, 2019. [Online]. Available: https://digiday.com/marketing/outdoor-advertising-braced-programmatic-moment/
[16] R. Church and C. ReVelle, "The maximal covering location problem," *Papers Regional Sci. Assoc.*, vol. 32, no. 1, pp. 101–118, 1974.
[17] Y. Xu, S.-L. Shaw, Z. Fang, and L. Yin, "Estimating potential demand of bicycle trips from mobile phone data—An anchor-point based approach," *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 8, p. 131, 2016.
[18] A. T. Murray, D. Tong, and T. H. Grubesic, "Spatial optimization: Expanding emergency services to address regional growth and development," *Studies Appl. Geograph. Spatial Anal.*, 2012, pp. 109–122.
[19] E. S. Hougland and N. T. Stephens, "Air pollutant monitor siting by analytical techniques," *J. Air Pollut. Control Assoc.*, vol. 26, no. 1, pp. 51–53, 1976.
[20] R. Wei and A. T. Murray, "Continuous space maximal coverage: Insights, advances and challenges," *Comput. Oper. Res.*, vol. 62, pp. 325–336, Oct. 2015.
[21] R. G. Cromley, J. Lin, and D. A. Merwin, "Evaluating representation and scale error in the maximal covering location problem using GIS and intelligent areal interpolation," *Int. J. Geograph. Inf. Sci.*, vol. 26, no. 3, pp. 495–517, 2012.
[22] D. Tong and A. T. Murray, "Maximising coverage of spatial demand for service*," *Papers Regional Sci.*, vol. 88, no. 1, pp. 85–97, 2009.
[23] A. T. Murray, D. Tong, and K. Kim, "Enhancing classic coverage location models," *Int. Regional Sci. Rev.*, vol. 33, no. 2, pp. 115–133, 2010.
[24] G. Alexandris and I. Giannikos, "A new model for maximal coverage exploiting GIS capabilities," *Eur. J. Oper. Res.*, vol. 202, pp. 328–338, Apr. 2010.
[25] R. Wei, "Coverage location models: Alternatives, approximation, and uncertainty," *Int. Regional Sci. Rev.*, vol. 39, no. 1, pp. 48–76, 2016.
[26] A. T. Murray and D. Tong, "Coverage optimization in continuous space facility siting," *Int. J. Geograph. Inf. Sci.*, vol. 21, no. 7, pp. 757–776, 2007.
[27] *Introduction of Wuxue, Hubei Province*. Accessed: Nov. 11, 2018. [Online]. Available: http://www.wuxue.gov.cn/art/2018/4/24/art_9_55.html/
[28] K. Li, X. Xu, and M. N. S. Swamy, "Modelling and analysis of regional service behavior properties of mobile internet applications," *IEEE Access*, vol. 5, pp. 4795–4807, 2017.
[29] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.
[30] H. Elsawy, W. Dai, M. Alouini, and M. Z. Win, "Base station ordering for emergency call localization in ultra-dense cellular networks," *IEEE Access*, vol. 6, pp. 301–315, 2018.
[31] J. Xiang, Z. Zhou, L. Shu, T. Rahman, and Q. Wang, "A mechanism filling sensing holes for detecting the boundary of continuous objects in hybrid sparse wireless sensor networks," *IEEE Access*, vol. 5, pp. 7922–7935, 2017.
[32] M. Zhang, P. Yang, C. Tian, and S. Tang, "You can act locally with efficiency: Influential user identification in mobile social networks," *IEEE Access*, vol. 5, pp. 136–146, 2017.
[33] Y. Yue, Y. Zhuang, A. G. O. Yeh, J.-Y. Xie, C.-L. Ma, and Q.-Q. Li, "Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy," *Int. J. Geograph. Inf. Syst.*, vol. 31, no. 4, pp. 658–675, 2017.
[34] X. Yang *et al.*, "Understanding spatiotemporal patterns of human convergence and divergence using mobile phone location data," *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 10, p. 177, 2016.
[35] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in rome," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 141–151, Mar. 2011.
[36] Y. Lu, Z. Zhao, B. Zhang, L. Ma, Y. Huo, and G. Jing, "A context-aware budget-constrained targeted advertising system for vehicular networks," *IEEE Access*, vol. 6, pp. 8704–8713, 2018.
[37] B. Page, Z. Anesbury, S. Moshakis, and A. Grasby, "Measuring audience reach of outdoor advertisements," *J. Advertising Res.*, vol. 58, no. 4, pp. 456–463, 2018.
[38] F. Jiang, K. Thilakarathna, M. Hassan, Y. Ji, and A. Seneviratne, "Efficient Content Distribution in DOOH Advertising Networks Exploiting Urban Geo-Social Connectivity," in *Proc. 26th Int. Conf. World Wide Web Companion*, Apr. 2017, pp. 1363–1370.
[39] Z. Li, G. Xie, J. Lin, Y. Jin, M.-A. Kaafar, and K. Salamatian, "On the geographic patterns of a large-scale mobile video-on-demand system," in *Proc. IEEE Conf. Comput. Commun.*, Toronto, ON, Canada, Apr./May 2014, pp. 397–405.
[40] Z. Fang, X. Yang, Y. Xu, S.-L. Shaw, and L. Yin, "Spatiotemporal model for assessing the stability of urban human convergence and divergence patterns," *Int. J. Geograph. Inf. Sci.*, vol. 31, no. 11, pp. 2119–2141, 2017.
[41] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng, "Characterizing user behavior in mobile internet," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 1, pp. 95–106, Mar. 2015.
[42] K. Han, J. Park, and M. Y. Yi, "Adaptive and multiple interest-aware user profiles for personalized search in folksonomy: A simple but effective graph-based profiling model," in *Proc. Int. Conf. Big Data Smart Comput. (BIGCOMP)*, Jeju, South Korea, Feb. 2015, pp. 225–231.

MENG HUANG is currently pursuing the Ph.D. degree with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. Her research interest includes human mobility models and applications.

ZHIXIANG FANG received the M.Sc. and Ph.D. degrees from Wuhan University, in 2002 and 2005, respectively, where he is currently a Professor with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing. His research interests include spatiotemporal modeling of human behavior, space-time GIS for transport, and intelligent navigation.

SHILI XIONG received the M.S. degree in advertising from the University of Illinois at Urbana-Champaign, in 2015, where she is currently pursuing the Ph.D. degree with the Institute of Communications Research. Her research interests include information processing and attention in new media and audience analyses.



TAO ZHANG received the Ph.D. degree from the State Key Laboratory of Software Engineering, Wuhan University. She is currently with China Mobile Group Hubei Company Limited as a Senior Engineer of communication system. Her research interests include the service-oriented architecture and web service.