Tourism Event Analytics with Mobile Phone Data

YAN LENG, The University of Texas at Austin, USA
ALEJANDRO NORIEGA and ALEX PENTLAND, MIT, USA

Tourism has been an increasingly significant contributor to the economy, society, and environment. Policy-making and research on tourism traditionally rely on surveys and economic datasets, which are based on small samples and depict tourism dynamics at a low granularity. Anonymous call detail record (CDR) is a novel source of data with enormous potential in areas of high societal value: epidemics, poverty, and urban development. This study demonstrates the added value of CDR in event tourism, especially for the analysis and evaluation of marketing strategies, event operations, and the externalities at the local and national levels. To achieve this aim, we formalize 14 indicators in high spatial and temporal resolutions to measure both the positive and the negative impacts of the touristic events. We exemplify the use of these indicators in a tourism country, Andorra, on 22 high-impact events including sports competitions, cultural performances, and music festivals. We analyze these touristic events using the large-scale CDR data across 2 years. Our approach serves as a prescriptive and a diagnostic tool with mobile phone data and opens up future directions for tourism analytics.

CCS Concepts: • **Applied computing** → **Economics**; **Computer-aided design**; • **Information systems** → **Information extraction**; **Business intelligence;**

Additional Key Words and Phrases: Tourism planning, mobile phone data, data mining, events analytics, performance indicators

ACM Reference format:
Yan Leng, Alejandro Noriega, and Alex Pentland. 2021. Tourism Event Analytics with Mobile Phone Data. *ACM/IMS Trans. Data Sci.* 2, 3, Article 25 (August 2021), 22 pages.
https://doi.org/10.1145/3479975

1 INTRODUCTION

Tourism has been an increasingly significant factor in the global economy, society, and environment, accounting for a significant share of GDP and labor force [25, 28]. Planners and policy-makers have long been attempting to understand the contribution and impact of tourism to the economy [4, 26]. The success of tourism events expands its potential to step beyond narrow leisure-based tourism [15]. The growing importance of event tourism, an applied field devoted to understanding and improving tourism through events, has resulted in substantial global competition and generated values for both the public and private sectors [14]. Due to the increasing of event

Authors' addresses: Y. Leng, McCombs School of Business, the University of Texas at Austin, 2110 Speedway, Austin, TX 78705, United States; email: yan.leng@mccombs.utexas.edu; A. Noriega and A. Pentland, MIT Media Lab, 75 Amherst Street, MA 02139, United States; emails: {noreiga, pentland}@mit.edu.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2021 Association for Computing Machinery.
2577-3224/2021/08-ART25 $15.00
https://doi.org/10.1145/3479975

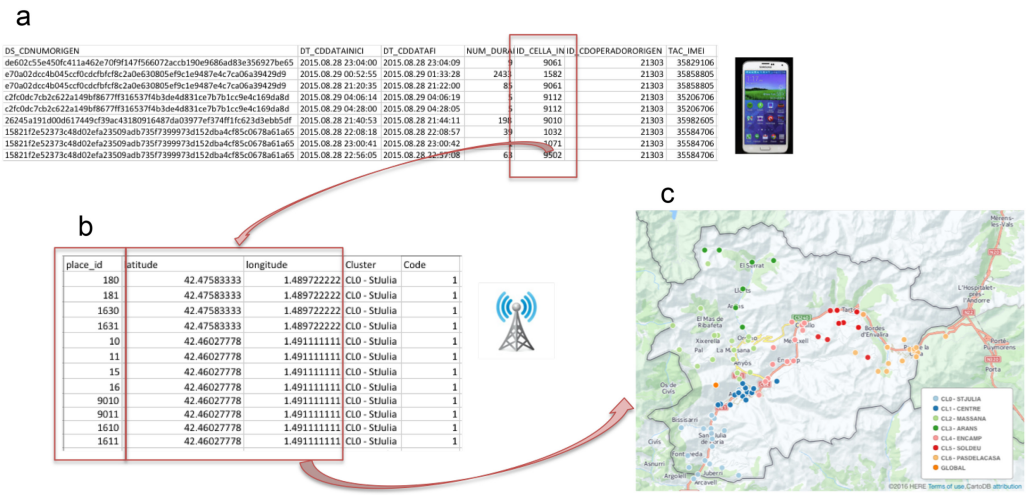
tourism, research on planning, operation, and marketing of special events has grown exponentially since 2008 [15].

Most tourism and events analytics literature is based on surveys, interviews, observations, or focus group studies [4, 10, 12, 15]. Pettersson et al. [2009] studied the spatial and temporal nature of event experiences through interviews, participant observation, and photography at a major sporting event in Sweden, contributing to a better understanding of how visitors interact with the event and with each other. This study also builds theories on what social factors contribute to tourist experiences [24]. Li et al. [2008] conducted a two-phase online survey with a six-month interval to analyze several indicators, including first-time and revisit visitors via many indicators, including demographics, travel behavior, pre- and post-trip congruency in travel activity preferences, and post-trip evaluation on first-time and revisit visitors [23]. These studies reveal insights on event performances and tourist experiences and can significantly benefit from using large-scale behavioral data to validate the theories on tourist behaviors on a wider scale.

In recent years, mobile sensors and social media generate incredible opportunities to provide answers for questions related to traveler behaviors and experiences [3]. Large-scale behavioral data have some unique advantages, since they are dynamic, large-scale, and have a high spatio-temporal resolution and therefore can reveal more insights on tourist behaviors. First, the data are more reliable than self-reported data in surveys, since the reliability of the latter may be questionable due to revealed preference bias, memory error, and non-response biases [16]. Second, researchers can cross-reference large-scale data with experiments or other geo-located and large-scale data. For example, Dong et al. combine and aggregate bank transactions and twitter data into the same geographic regions and observational period to understand the urban segregation patterns [11]. For example, Birenboik et al. tracked 25 students with a mobile application for passive GPS and pops up three micro-surveys an hour for subjective evaluation of crowdedness and sense of security [7]. Wood et al. [2016] combined field investigation and Flickr data, showing that Flickr data can reveal what elements of nature attracted tourists and how this would alter visiting behaviors [30]. In summary, researchers and practitioners can obtain more knowledge about the target market from high-resolution behavioral data.

Large-scale geo-located data, such as **call detail records (CDR)** collected from mobile phones, enabled new ways to understand the connections among travelers, social events, and urban infrastructures [5]. Mobile carriers collect CDR for billing purposes; hence, it is among the most widely existing sources of information to measure human mobility and urban dynamics [18, 20, 29]. CDR has been applied to various fields for social good, such as mobility modeling [18], social influence [19], epidemics [13], and disaster control [8]. CDR is a standardized dataset with information on geolocation and timestamp, and it is commonly used in the data science and transportation literature. Therefore, analysis and methods developed on one CDR datum can be applied to CDRs collected in other cities/countries given its standard format. However, there is a lack of research in using large-scale and longitudinal data in tourism analytics.

In this article, we formalize the performance indicators for tourism events and demonstrate its relevancy for policy-making in tourism. Specifically, we demonstrate how to mobile phone data to evaluate tourism marketing strategies, understand tourist experiences, compare the approximate income of tourists, and estimate externalities. With higher spatial-temporal resolutions compared with self-reported surveys, CDR reveals more insights and more in-depth knowledge into various aspects of the tourism industry. We define a series of informative indicators to evaluate the special events, including tourist flows, dynamics of new tourist flows of revisits, tourist externalities due to congestion, spatial distribution, economic impacts, and profiling of tourist interests. These indicators not only shed light on tourism strategies but also help to enhance the services of other industries (e.g., hospitality, transportation). Policy-makers and managers can easily apply



DS_CDNUMORIGEN	DT_CDDATAINICI	DT_CDDATAFI	NUM_DURA	ID_CELLA_IN	ID_CDOPERADORORIGEN	TAC_IMEI
de602c55e450fc411a462e70f9f147f566072accb190e9686ad83e356927be65	2015.08.28 23:04:00	2015.08.28 23:04:09	9	9061	21303	35829106
e70a02dcc4b045ccf0cdcfbfcf8c2a0e630805ef9c1e9487e4c7ca06a39429d9	2015.08.29 00:52:55	2015.08.29 01:33:28	2433	1582	21303	35858805
e70a02dcc4b045ccf0cdcfbfcf8c2a0e630805ef9c1e9487e4c7ca06a39429d9	2015.08.28 21:20:35	2015.08.28 21:22:00	85	9061	21303	35858805
c2fc0dc7cb2c622a149bf8677ff316537f4b3de4d831ce7b7b1cc9e4c169da8d	2015.08.29 04:06:14	2015.08.29 04:06:19	5	9112	21303	35206706
c2fc0dc7cb2c622a149bf8677ff316537f4b3de4d831ce7b7b1cc9e4c169da8d	2015.08.29 04:28:00	2015.08.29 04:28:05	5	9112	21303	35206706
26245a191d00d617449cf39ac43180916487da03977ef374ff1fc623d3eb05df	2015.08.28 21:40:53	2015.08.28 21:44:11	198	9010	21303	35982605
15821f2e52373c48d02efa23509adb735f7399973d152dba4cf85c0678a61a65	2015.08.28 22:08:18	2015.08.28 22:08:57	39	1032	21303	35584706
15821f2e52373c48d02efa23509adb735f7399973d152dba4cf85c0678a61a65	2015.08.28 23:00:41	2015.08.28 23:00:42	1	1071	21303	35584706
15821f2e52373c48d02efa23509adb735f7399973d152dba4cf85c0678a61a65	2015.08.28 22:56:05	2015.08.28 22:57:08	63	9502	21303	35584706

place_id	latitude	longitude	Cluster	Code
180	42.47583333	1.489722222	CL0 - StJulia	1
181	42.47583333	1.489722222	CL0 - StJulia	1
1630	42.47583333	1.489722222	CL0 - StJulia	1
1631	42.47583333	1.489722222	CL0 - StJulia	1
10	42.46027778	1.491111111	CL0 - StJulia	1
11	42.46027778	1.491111111	CL0 - StJulia	1
15	42.46027778	1.491111111	CL0 - StJulia	1
16	42.46027778	1.491111111	CL0 - StJulia	1
9010	42.46027778	1.491111111	CL0 - StJulia	1
9011	42.46027778	1.491111111	CL0 - StJulia	1
1610	42.46027778	1.491111111	CL0 - StJulia	1
1611	42.46027778	1.491111111	CL0 - StJulia	1

Fig. 1. Call detail records. (a) presents a snapshot of mobile phone data. CDR contains the ID of the user, the timestamp, the cell tower ID, the country that the phone is registered, and the phone type. As demonstrated in (b), we map the cell tower ID in this dataset to another database for the spatial distribution. (c) presents the spatial distribution of cell towers in Andorra.

our method and replicate the study in other regions and sectors for more evidence-based decision-making.

As a case study, we apply the proposed indicators to an especially interesting case study in the country of Andorra. Specifically, we use CDR to measure tourist volumes, new tourists, revisit tourists, the spatial distribution of tourists, spending habits of tourists, and the positive and negative externalities of the special events. We demonstrate the use of these indicators for the planning and evaluation of high impact touristic events in Andorra in the year 2015, including cultural festivals and sports competitions. We compare the marketing and organization strategies of several special events and evaluate their relative performances. This article contributes to tourism planning and event analytics using novel indicators extracted from mobile phone records.

The rest of the article is structured as follows. Section 2 introduces the context of Andorra. Then we describe the primary data used in this study. Section 3 defines and describes several informative indicators extracted from mobile phone records. In Section 4, we use two case studies to illustrate how to utilize the defined indicators to assess the performances of several summer and winter events in Andorra. Finally, Section 5 concludes this study and discusses future works.

2 SETTING: ANDORRA AND THE MOBILE PHONE DATA

We briefly introduce the context of a European country of Andorra. We then describe the high-resolution and other infrastructure-related data used to conduct the study.

2.1 Andorra

Andorra is a small European country situated between France and Spain. The economy of this country heavily relies on tourism [1]. The population of Andorra is only 85,000; meanwhile, it attracts an estimated 10.2 million visitors annually [9]. Andorra is famous for skiing, sports, and duty-free shopping. The skiing resorts cover an area of over 175 km^2, accounting for about 37% of the whole area (468 km^2). The sports brings in over seven million visitors annually and an estimated 340 million euros per year. All of these characteristics make Andorra an especially interesting case study for tourism analysis. We display the map of Andorra in Figure 1(c).

2.2 Data

Call Detail Records. The passively collected behavioral data used in this study is the CDR. The mobile carriers initially collect CDR for billing purposes, and hence these data widely exist in almost every country in the world. Though widely available, its application to tourism analytics is new. The coverage of the CDR data used in our study is 100%. This means that every traveler in Andorra who uses their phones to make phone calls, send text messages, or use cellular data for internet connections (e.g., for web browsing) will be included in our data.

The data have a high spatial and temporal resolution, compared with traditional surveys, and have the most substantial penetration rate among all passively-collected data. The personal data are recorded whenever there are phone calls, Short Message Service, and internet data services. It contains information on user ID, the timestamp, the ID of the cell tower, the registry country, and the phone type, as shown in Figure 1(a). We map the cell tower ID to another database for the latitude and longitude. The CDR used in this study was collected for two years, from June 2014 to May 2016, by Andorra Telecom. Andorra Telecom is the only mobile carrier in Andorra; hence, the coverage of mobile phone data in our study is 100%, meaning that we have all the mobile phone data on individuals who visited Andorra. On top of the map in Figure 1(c), we display the distribution of cell towers.

In addition to the spatial and temporal information, CDR stores useful information to infer the socio-demographics of the individuals. For example, researchers can infer the price of the phone based on information recorded in the phone type. Accurately, we map the phone type code to another database for the brand, the vendor, the model, and the system of the phones. Besides, we can infer the nationality of the mobile phone users based on the registry country of the SIM card. Inferring these characteristics, as well as other behavioral segmentation of users, is beneficial for a variety of applications, including marketing campaigns, event recommendations, and coupon distributions. We discuss more on how to use this information in Section 3.

The CDR data are anonymized using the hashing method, meaning that we assign each mobile phone an random ID so that the individual cannot be re-identified.

Road network. We complement the CDR with the road network, which enables us to map the cell-tower-based tourist flows on road links. We acquired the geographic information of the road network from the website of the government of Andorra.[1] We collected the attributes of each road link, including name, the number of lanes, capacities, and free-flow travel time, which is usually available in the government-hosted Geographic Information System database. We discuss how to use the road network information more specifically in Section 3.6.

Traffic Counts. Traffic counts are collected at key locations by the Andorra government to monitor internal mobility. Such traffic counts are only available at critical intersections collected by cameras. Only the monthly traffic counts are publicly available. These aggregated traffic counts can be used as the ground truth to scale up CDR-based traffic flows to the actual traffic flows. We demonstrate how to use this information in estimating congestion in Section 3.6.

3 INDICATORS FROM CDR

To evaluate tourism events from different dimensions, we define measurable indicators from the mobile phone data. The objective of the indicators is to offer business and tourism departments tactical and strategic policy designs. Specifically, we cover marketing, travel experiences, economic impacts, and negative externalities (e.g., traffic congestion). We take advantage of the high spatial

[1]The data are available on Govern D'Andorra.

and temporal granularity of the data and design indicators based on the aggregated behavioral information from tourists to reveal insights that are not available through traditional surveys.

3.1 Tourism Flow: The Popularity of The Events

The first and foremost measure for organizers and planners to evaluate event performances is the tourism flow, which is a commonly used indicator in the tourism industry. This measure is traditionally estimated from surveys. With CDR, we can measure tourist flows with more information about the types of tourists, and at a higher spatial resolution and on a dynamic basis. We can measure the flow by either tourist volumes in Equation (9) or tourist days in Equation (2). Next, we discuss precise formulations.

Let $t^e_{\min}$ and $t^e_{\max}$ be the start and end date for a particular event e. Since tourists may stay for an extra days prior to and after and events, we can add adjust $t_{\min}$ and $t_{\max}$ accordingly. Let $n \in N$ be the individuals who visited Andorra in our observational period. The cell towers individuals connect to when they visit the event venue is $\{L_e\}$. The set of individuals who visit the event e is N_e. We denote $\{t_n\}$ as the set of time instances where an individual n visited the country. We let $\{l_n\}$ be the set of cell towers individuals connect. For an individual n, we extract a series of location-time pairs with information on where and where (i.e., which cell tower) individual n visited Andorra, $\mathcal{G}_n = \{(l_1, t_1), (l_2, t_2), \ldots, (l_n, t_n)\}$.

For an event e, the volume of tourists ($\mathcal{V}_e$), the total number of tourists who visited Andorra during an event period, is defined as,

$$\mathcal{V}_e = \sum_n 1_{n \in N_e}, \tag{1}$$

where $\mathbf{1}_{\{\text{criteria}\}}$ is the indicator function, which equals one if the criteria are satisfied, and zero, otherwise.

Similarly, for an event e, tourist days ($\mathcal{D}_e$) is defined as

$$\mathcal{D}_e = \sum_{n \in N_e} \sum_{t_k \in \{t_n\}} 1_{t^e_{\min} \le t_k \le t^e_{\max}}. \tag{2}$$

Moreover, some complementary information in the CDR helps to segment tourist volume based on the nationalities, highlighting one of the advantages of cell phone data. As shown in Figure 2, the majority of tourists come from Spain, France, Russia, Belgium, Portugal, and the Netherlands. Visitors from different countries exhibit diverse travel patterns, such as gathering in different regions, various stay lengths, and different interests in events. For example, the Spanish and the French take up around 90% of the total number of tourists, as they are geographically closer to Andorra. The Spanish and the French usually visit Andorra for shopping and tend to stay for a shorter period, comparing with tourists who travel from farther countries.

The Spanish usually visit the city of Santa Julia, while the French typically visits the city of Andorra La Vella. Moreover, the Russians usually travel for skiing and tend to stay longer, compared with tourists from other countries in Europe. Understanding visitation patterns from different nations enables tourism departments and business to adapt marketing strategies to attract more tourists, and design operation strategies to improve customer experiences [19, 22]. For example, the transportation department can collaborate with the tourism department to offer frequent shuttles between Santa Julia and the Spanish border with signs and advertisements in Spanish; while providing shuttles between French borders and Andorra La Vella with services in French. Moreover, the tourism department can work with the hospitality industry to offer coupons, distribute tour guides, and offer day trips to skiing resorts targeting Russians.

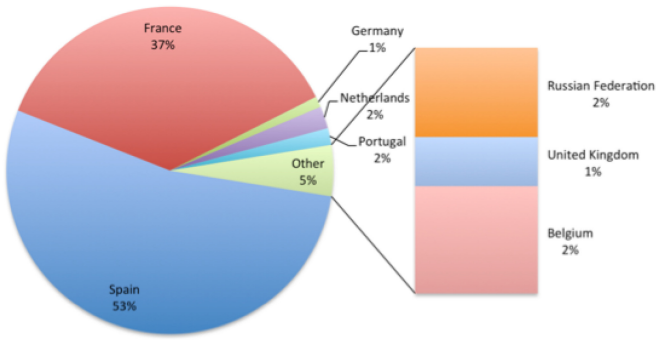


Fig. 2. Percentage of tourists from different countries. The main origins of tourists come from Spain, France, the Netherlands, Portugal, Russia, UK, and Belgium.

3.2 New Tourists as Evaluations of the Marketing Strategy

Andorra, similar to other tourism cities and countries, release tourism advertisement campaigns in different countries to attract travelers. There are two types of advertisement campaigns. Some advertisements aim to attract travelers to specific events; while other advertisements aim to create people's awareness around Andorra, where they may not have considered traveling to before. This strategy is more general than merely targeting some specific events. The critical question is to evaluate how successful the campaigns are? For a specific event, we propose to use the number of new tourists to evaluate the marketing campaign for the event. The second type of campaign is harder to measure, and we need a more extended period to measure its effect. For example, we can use the number of increased Russian tourists over a year to measure the effect of a campaign in Russia. Altogether, the evaluations of marketing strategies help to provide diagnostic insights into the previous strategy and, in turn, improve future policies.

In this article, we focus on measuring the effect of the first marketing strategy. New tourists volumes and new tourists days for an event e, $\mathcal{V}_e^{\text{new}}$, and $\mathcal{D}_e^{\text{new}}$, are considered as the tourists that have not visited Andorra prior to the event period. Specifically,

$$\mathcal{V}_e^{\text{new}} = \sum_{n \in N_e} \mathbf{1}_{\min(\{t_n\}) \geq t_{\min}}, \tag{3}$$

$$\mathcal{D}_e^{\text{new}} = \sum_{n \in N_e} \sum_{t_k \in \{t_n\}} \mathbf{1}_{\min(\{t_n\}) \geq t_{\min} \text{ and } t_{\min}^e \leq t_k \leq t_{\max}^e}, \tag{4}$$

where $\min(\cdot)$ is the operator returning the minimum of a set.

Other than the aggregated counts, we can use this metric to evaluate personalized interventions. For example, the hotels can distribute coupons to some targeted individuals and investigate whether this strategy effectively attracts the targeted tourists to visit Andorra. We leave this for future study.

3.3 Revisit Tourists as Evaluations of Tourist Experiences during the Events

Understanding tourist experiences is critical in evaluating the performance of events. Traditionally, the tourism department uses questionnaires to collect this information. However, information collected by surveys is labor intensive and may suffer from different sources of biases, such as memory error, stated preference bias and non-responsive error [16]. Whether tourists will revisit Andorra or not reflect tourist experiences during the event and can be used as the revealed

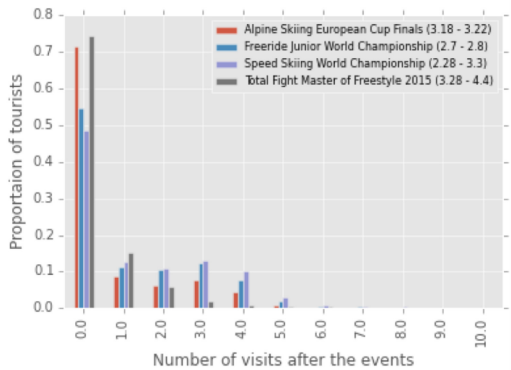


Fig. 3. Number of visits after attending a event. The x-axis and y-axis correspond to the number of visits after the events and the corresponding percentage, respectively. Each color represents a separate event.

preference to avoid the human-generated biases as mentioned earlier [17, 27]. Since CDR is collected continuously, we can easily extract the revisit tourists from these data.

Revisit tourists are the tourists who revisit Andorra after the event. Specifically, number of revisit tourist ($\mathcal{V}_e^{\text{revisit}}$) and revisit days ($\mathcal{D}_e^{\text{revisit}}$) can be computed as

$$\mathcal{V}_e^{\text{revisit}} = \sum_{n \in N_e} \mathbf{1}_{\max(\{t_n\}) \geq t_{\min}}, \tag{5}$$

$$\mathcal{D}_e^{\text{revisit}} = \sum_{n \in N_e} \sum_{t_k \in \{t_n\}} \mathbf{1}_{\max(\{t_n\}) \geq (t_{\min} + t_b) \text{ and } t_{\min}^e \leq t_k \leq t_{\max}^e}, \tag{6}$$

where $\max(\cdot)$ is the operator returning the maximum of the set.

We integrate Equation (3) and Equation (5) to understand whether new tourists revisits potentially becomes loyal visitors (i.e., whether the new tourist revisit or not),

$$\mathcal{V}_e^{\text{new, revisit}} = \sum_{n \in N_e} \mathbf{1}_{\max(\{t_n\}) \geq (t_{\min} + t_b) \text{ and } \min(\{t_n\}) \geq (t_{\min} - t_b)}. \tag{7}$$

Further, organizers can measure the number of revisits after the event. As Figure 3 shows, Freeride Junior World Championship and Speed Skiing World Championship attracted a more significant number of revisits, comparing with the other two events. Freeride Junior World Championship is different from other events, since it is a competition among youth riders aged 15–18. More in-depth studies are needed to understand why these junior events attract a more significant number of tourists. For example, whether events targeting junior athletes or kids attract a higher revisit rate. Moreover, hotels can provide a reward loyal program to attract tourists who tend to stay longer.

The practitioners and policy-makers can perform diagnostic analysis using CDR to understand what factors contribute to more or less revisits. For example, policy-makers can regress the number of revisits on some performance metrics as the independent variables, such as traffic congestion, length of the event, the average price of the hotels. While this may not provide a causal understanding, such an analysis provides a diagnostic tool to identify specific aspects that need to be improved. With this insight, they can use domain knowledge or design experiments. Moreover, policy-makers can study the effect of personalized interventions on revisits. For example, whether specific coupons or ticket bundles improve tourist experiences. The mobile phones provide both the channel to diffuse the intervention, and a medium to collect data to measure such effects.

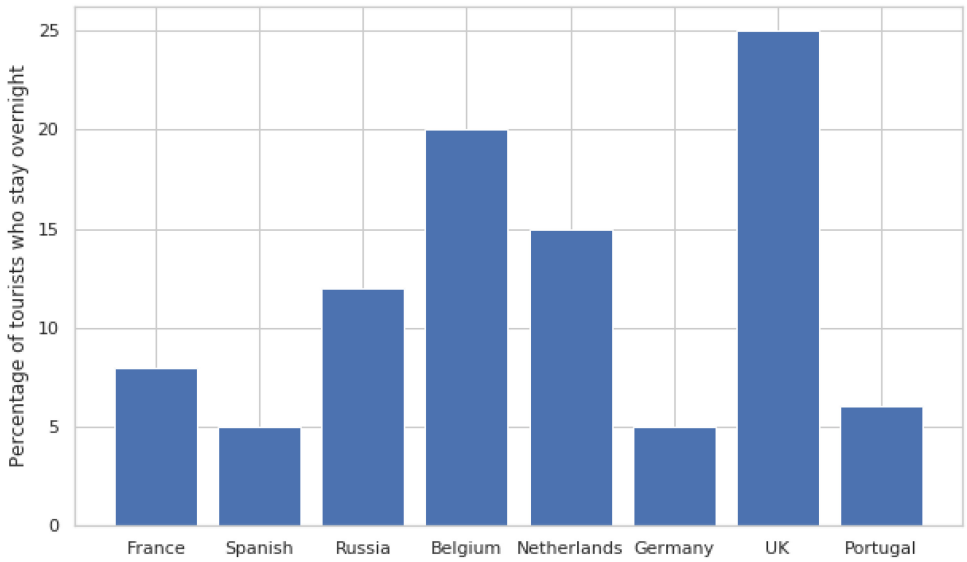


Fig. 4. Percentage of tourists who stay overnights at city La Massana from different nations on a normal day. The x-axis is the country, and the y-axis is the percentage of overnight tourists at La Massana among all tourists from the corresponding nation.

3.4 Overnight Tourists and Average Stay Length: Insights for the Hospitality Industry

We define two measures, overnight tourists and average stay length, to understand how does the hospitality industry benefit from the events. To quantify the overnight tourists, we define two measures: the average number of tourists and the percentage of tourists who stay overnight, denoted as $\mathcal{V}_{\text{night}}$ and $\mathcal{V}^P_{\text{night}}$, respectively.

$$\mathcal{V}_{\text{night, e}} = \sum_{n \in N_e} 1_{\exists t_i \in \{t_n\}, \text{ where } t^{\min}_{\text{night}} \le t_i \le t^{\max}_{\text{night}}}, \tag{8}$$

$$\mathcal{V}^P_{\text{night, e}} = \frac{\mathcal{V}_{\text{night, e}}}{\mathcal{V}_e}, \tag{9}$$

where $t^{\min}_{\text{night}}$ and $t^{\max}_{\text{night}}$ are the earliest and the latest time as the definition for night time. In our study, we set it to be 12 am and 6 am.

This information on overnight tourists facilitates hospitality in designing different marketing strategies. Let us present an example with an analysis of overnight tourists in the city of La Massana. As shown in Figure 4, tourists from the UK, Belgium, and the Netherlands are more likely to stay overnight at La Massana, in contrast with the Spanish and the German. With this result, the hotels in La Massana can design different marketing and operation strategies for various countries. For example, for nations with tourists who stay overnight in Andorra, but mostly in the capital area to be close to the event venue (e.g., the Russians), hotels can offer free shuttles to the event venues or lower the price to appeal to the tourists. Moreover, hotels can partner with the tourism department or tourist agencies for advertising in Spain, Germany, and Portugal. For example, they can offer a sale package together with the event companies or attract new tourists by listing the advertisements with the local tourist office or the tourism website.

Following overnight tourists, we propose to measure the average stay length, which tends to correlate with the revenues of the hospitality industry from the events positively. We use Figure 5

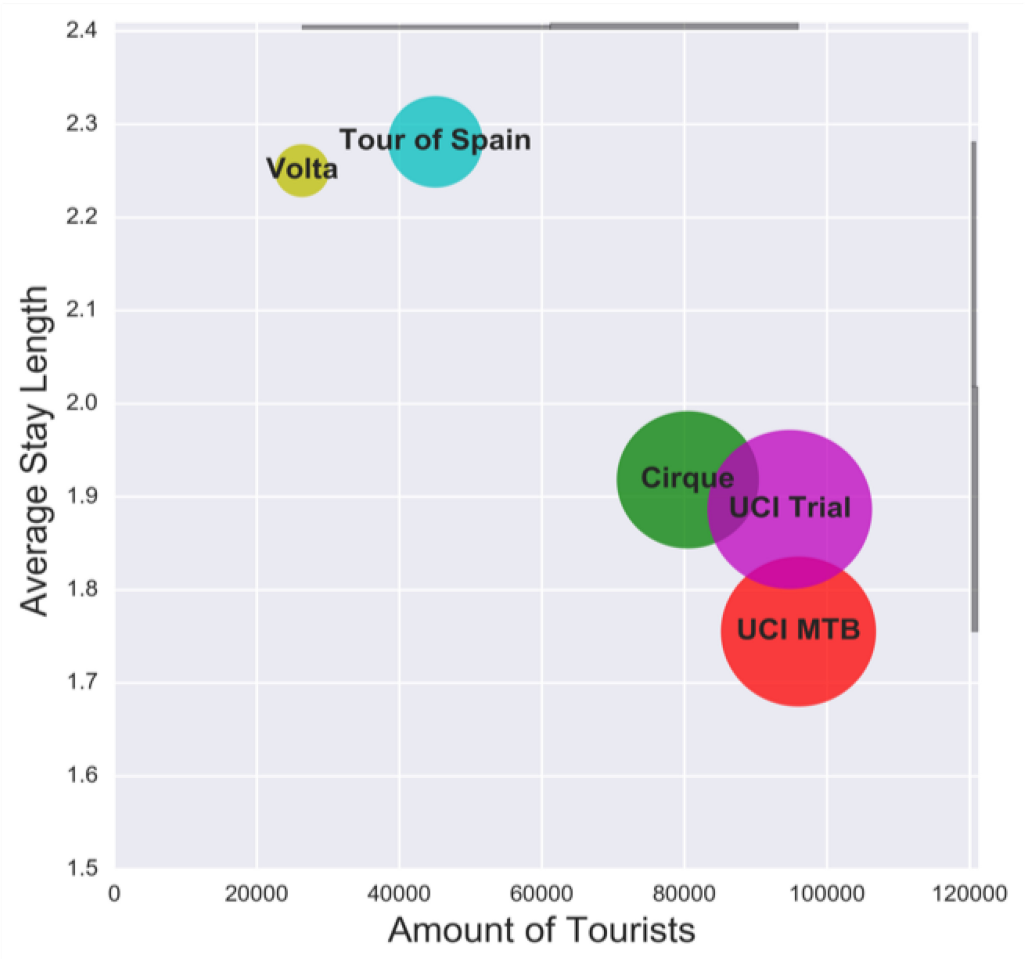


Fig. 5. Length of stay for different events. The x-axis and y-axis correspond to the amounts of tourists and the average stay length. Each circle corresponds to an event, and the radius is proportional to the amount of tourists.

to demonstrate its importance. Hotel revenues result both from a large number of tourists and the longer stays, which corresponds to the upper-right diagonal of the figure. As we can see, even though Cirque De Soleil, UCI Trial, and UCI MTB attracted a more significant number of tourists, the tourists stayed for about a half-day shorter than Volta and Tour of Spain. This figure suggests the organizers of Volta and Tour of Spain boost the popularities through better marketing, and the organizers of Cirq, UCI Trial, and UCI MTB to attract longer stays with better services or lower prices of hotels. This type of analysis helps hotels understand their markets, and also inform the marketing strategies on attracting longer stays for the Cirq, UCI Trial, and UCI MTB. More analysis on these events can be found in Section 4.2.

3.5 Spatial Distribution as a Measure for Positive Externality

Next, we define a measure of the spatial distribution of tourists during the event. More distributed tourists generate two benefits. On the one hand, events attract tourists to visit Andorra, and an

associated positive externality is that tourists may explore other regions of Andorra. If visitors explore more regions, then more businesses, other than the organizations that host the event, can benefit from the events due to the increasing number of visitors. Therefore, other parts of the country can also increase revenues as a positive externality of such an event. On the other hand, the tourism department can identify the "hotspots" where the number of tourists exceeds the capacities of the road infrastructures or the **Point of Interests (POIs)** from the spatial distributions. This situation creates negative externalities on tourist experiences, residents' life quality, energy consumption, and environment. That is, distributed tourists will help reduce the crowdedness of specific areas and reduce congestion. With these two benefits, the tourism department needs to collaborate with other industries (e.g., hospitality, transportation) to attract and balance tourists in different regions.

To understand the spatial distribution of tourists, we aggregate the total number of tourists who have visited cell tower c during the event period $N_{c,e}$, where c is the index of the cell tower, C is the set of cell towers, and $c \in C$,

$$N_{c,e} = \sum_{n \in N_e} \mathbf{1}_{\exists l_k = c \text{ and } t_{\min} \le t_k \le t_{\max} \text{ for } (l_k, t_k) \in \mathcal{G}_n}. \tag{10}$$

Tourist flows exhibit high temporal variability. Thus, it is beneficial to measure the spatial distribution of tourists during a specific period for evidence-based campaigns. For example, traffic flows during the daytime are more relevant for event venues and tourist attractions. Measuring the tourist flows during noon and dinner time is more instructive for restaurants. The tourist flow at night is more relevant to the hospitality industry. Specifically, we can measure the spatial distribution during the period τ as

$$N_{c,e,\tau} = \sum_{n \in N_e} \mathbf{1}_{\exists l_k = c \text{ and } t_{\min} \le t_k \le t_{\max} \text{ and } t_k \in \tau, \text{ for } (l_k, t_k) \in \mathcal{G}_n}, \tag{11}$$

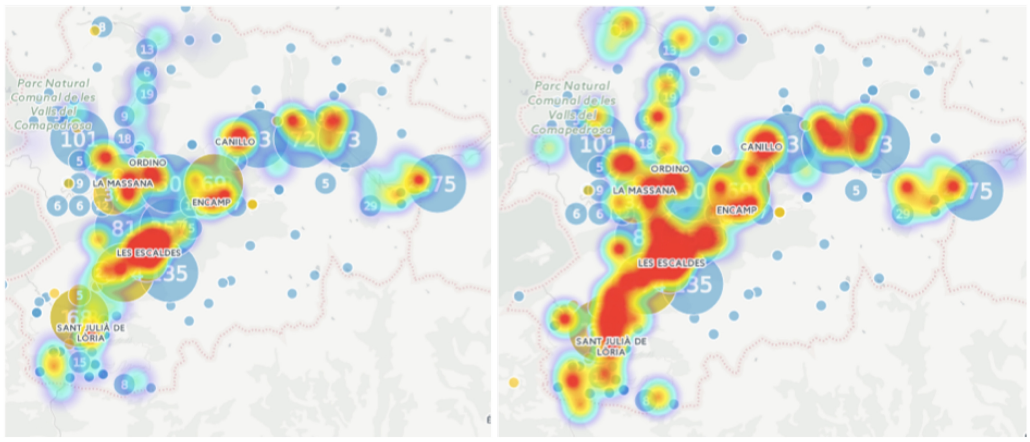
where τ indexes a certain time period of interest.

The policy-makers can build a visualization platform to monitor traffic flows and derive actionable insights. Feeding the visualization with the streaming data is compelling for real-time monitoring to respond to emergencies and unexpected events. This type of visualization makes the big data accessible and digestible to narrate, engage, and educate a larger audience and citizens. For example, Figure 6 visualizes the spatial distribution of tourists from 3 am to 4 am and from 5 pm to 6 pm on a non-event day.[2] We can see that most tourists centered at the capital, Les Escaldes, before sunrise and in the late afternoon. There were significantly fewer tourists in other cities at night, especially on the periphery of the country, compared with the capital area.

3.6 Congestion: The Negative Externalities from the Events

Tourists may cause congestion, which is the externality by tourism that adversely impacts local communities, traveler experiences, and environmental systems. Popular events, seemingly successful in attractiveness, are at a higher risk of congestion, which, however, becomes a disadvantage. Severe congestion during the event indicates the need to improve tourist demand management or provide more transportation services on the supply side [21]. Mobile phone data provides dynamic monitoring of the performance of the transportation system. The ability to monitor traffic flows at a more granular spatio-temporal scale, compared with that shared by the government, provides a deeper and timely understanding of tourist travel patterns, and enables more demand-responsive urban traffic planning and operations management [18, 20].

[2]The visualization is built with the CARTO, which can be accessed through this link.



(a) 3 am – 4 am (b) 5 pm – 6 pm

Fig. 6. Spatial distribution of tourists on 3–4 am and 5–6 pm on a non-event day. The blue to red color correspond to sparse to dense population.

Piezoelectric sensors or cameras provide direct measures on traffic congestion and traffic flows. However, due to the limited resources, they are not installed on every road links. In this situation, CDR provides a cost-efficient approach to infer the information from mobile phone data.

We propose the following three-step method to estimate traffic flow. First, we compute the tower-to-tower **Origin-Destination (O-D)** matrix by aggregating the individual movements between cell towers c_1 and c_2 during the period h. It is vital to estimate traffic flows by each hour of the day due to the high temporal variability of traffic flows. The tower-to-tower traffic flow $F_{c_1,c_2,h}$ is computed as

$$F_{c_1,c_2,h} = \sum_{n \in N_e} \sum_{(l_k,t_k),(l_k,t_{k+1}) \in \mathcal{G}_n} \mathbf{1}_{l_k = c_1 \text{ and } l_{k+1} = c_2, (t_k + t_{k+1})/2 \in h}. \tag{12}$$

Second, we assign the O-D flows to road links. Let $\mathcal{M}(\cdot)$ be a mapping from O-D pairs to road links, defined as, $\mathcal{M} : (c_1, c_2) \rightarrow \{v\}$. The amount of traffic flows on road link i during the time period h is, $V_{i,h}$, which can be computed as

$$V_{i,h} = \sum_{c_1 \in C, c_2 \in C} F_{c_1,c_2,h} \times \mathbf{1}_{i \in \mathcal{M}(c_1,c_2)}. \tag{13}$$

The third step is to scale the aggregated tower-to-tower traffic flow to actual vehicle trips using traffic counts as the ground truth. This step is critical, since individuals may travel in the same vehicle, hence do not directly translate into vehicle trips that contribute directly to congestion. Moreover, though the penetration rate of mobile phone data is high, the event-driven characteristics of the data make it highly possible that some of the trips are missing in the data. Let SV_i be the scaled traffic counts on road link i,

$$SV_{i,h} = V_{i,h} \times \beta_i, \tag{14}$$

where β_i is the scaling factor for road link i. We can estimate β_i as

$$\beta_i = \frac{TC_{i,\tilde{h}}}{\sum_{h \in \tilde{h}} V_{i,h}}, \tag{15}$$

where $TC_{i,\tilde{h}}$ is the traffic flow on road link i during period $\tilde{h}$ collected by the transportation department. We use $\tilde{h}$ to denote the time period the traffic counts are collected by the transportation department. The higher temporal resolution of the traffic counts collected by the transportation department improves the accuracy of the scaling factor.

After we obtain the traffic flow, we can estimate the externality. In transportation, practitioners and planners estimate travel time as a function of the traffic flow and characteristics of the road infrastructure. The most widely used approach is the Bureau of Public Road function [2], which models travel time (t_{estimate}) as a function of the ratio between actual traffic volume and road maximum flow capacity, volume-over-capacity [6], as shown in Equation (16). Specifically,

$$t_{\text{estimate}} = t_{\text{free flow}} \times (1 + \alpha(V/\text{C})^{\gamma}), \tag{16}$$

where $t_{\text{free flow}}$ is the maximum traffic flow at which vehicles can move in a reasonable order a lane of the road. Δt is the delay caused by congestion. Free flow travel time depends on the number of lanes per direction and the capacity per lane. α and γ are parameters that are used to characterize the non-linear relationship between V/C and t_{estimate}. According to the Bureau of Public Roads, the default values are $\alpha = 0.15$ and $\gamma = 4$. The congestion time (Δt^{+}) is the time difference between the estimated travel time with the estimated flow and the free flow travel time,

$$\Delta t^{+} = t_{\text{estimate}} - t_{\text{free flow}}. \tag{17}$$

In this study, we use the congestion time as a negative externality. Since congestion generates economic, productivity, environmental, and energy loss, policy-makers can further measure the loss in different aspects.

3.7 Approximate Spending Habits of Tourists

Next, we aim to understand the spending habits of tourists, which has several benefits for decision-makers. First, understanding the spending habits at the individual and aggregated level enables businesses to promote offers to particular customer segmentation. Specifically, approximating the spending habits of tourists helps businesses identify the potential markets that are more relevant for their products or services. For example, wine tasting events with high entrance fees may be more appealing to high-income tourists, while street food tasting events with free access may be more attractive to lower-income tourists. Moreover, revenues from the special events are assumed to correlate with the disposable income and spending habits of tourists, which can, to some degree, be proxied by the price of the phone positively. If we compare the average phone price across events, then we can approximately understand and compare the income distribution of tourists.

Since CDR does not contain information directly about the income or spending habits of tourists, we propose an indicator to approximate tourist income. The average price of the phones of tourists may reveal their spending habits and the disposable income of the tourists. Note that the phone price is a weak proxy for revenues or spending in Andorra. The accurate revenues and individual spending generated from events can be measured by other passively collected data, such as banking transactions, if available. However, the average price of the phone is also informative as a relative measure across different events. In this way, this indicator shows that some events attracted tourists with higher disposable income than others. For an event targeting higher-income tourists, the event companies can distribute the advertisement to individuals with pricey phones.

We scrape the price of the phone from Amazon API based on the brand, model, and the operation system, recorded in the CDR data.[3] With this information, we can define a mapping function $\mathcal{B}$

[3]We use the Amazon price API, which is available through this link.

from individual n to the price of the phone Pr, where $\mathcal{B} : n_e \rightarrow Pr$. The average phone price of tourists attending a particular event for an event e is

$$B_e = \frac{1}{N_e} \sum_{n \in N_e} \mathcal{B}(n). \tag{18}$$

3.8 Tourist Interests from Historical Behavioral Profiles

The interests of the tourists can advise businesses and tourism departments on designing tactical and strategic policies. We can use the historical behavioral patterns as a proxy, which real three types of interests, i.e., events, activities, and stores. This knowledge enables businesses to promote advertisements to attract new tourists and retain existing visitors. For example, policy-makers can ask questions on how to target a particular event to specific market segmentation and how to design marketing strategies for different countries.

We infer tourist interests in two steps. We first label the cell towers by integrating the geographic locations of the cell towers and the POIs surrounding the cell towers. The POIs can reveal potential activities that tourists can engage in nearby the cell towers. Specifically, we define a mapping function $\mathcal{S}$ from cell tower c to a type of activity a: where $\mathcal{A} : c \rightarrow a$ and $a \in A$, where A is a set of activities, including shopping, nature, wellness, leisure, gastronomic, culture, event venue, and others.

Second, we label the interests of each individual, based on the cell towers they connected. Specifically, we categorize each cell tower based on the surrounding Points of Interest. Aggregating the cell towers, we obtain a distribution for activities individuals are interested in to generate a behavioral profile of each individual. This behavioral profile of individual n is a distribution of time spends at each type of location, represented as, $\{(a, \mathbb{P}_n(a))\}_{a \in A}$,

$$\mathbb{P}_n(a) = \frac{\sum_{l_k \in \{\tilde{l}_n\}} \mathbf{1}_{\mathcal{S}(l_k)=a}}{|\{\tilde{l}_n\}|}, \tag{19}$$

where $\{\tilde{l}_n\}$ is a list of all the locations individual n visited. Note that this is different from $\{l_n\}$, which is the set of locations individual n visited. The difference between the two is $\{\tilde{l}_n\}$ captures the multiple occurrences of one cell tower. $|\cdot|$ is the operator counting all of the elements in the list.

The tourist interests of nationalities are informative in targeted marketing. As shown in Figure 7,[4] the Spanish and the French are more interested in shopping and natural activities, which make up two thirds and one-fourth of their total interest, respectively. One-third of the Dutch traveling to Andorra is interested in natural and shopping activities, which suggests a strategy to target the Netherlands with sales and skiing coupons. Two-thirds of Russians are interested in natural activities, with another one-sixth interested in cultural activities. Therefore, apart from advertising shopping and natural activities, the tourism department could focus their marketing strategies on advertising for museums and cultural events. For example, the tourism department can work with the ticket offices of the tourist attractions to provide annual membership for skiing, or bundling tickets for skiing and museums. As many Dutch and Russians are interested in cultural activities, the museum managers can provide brochures in the corresponding languages, and design production strategy with the predictions on tourists from different nations.

Moreover, we can use other mobility indicators for tourist segmentation. As an example, we use the percentage of overnight tourists, the number of cell towers visited in 80% of the time, the radius

[4]This figure can be found in our preliminary study on a non-archival conference proceeding via this link.

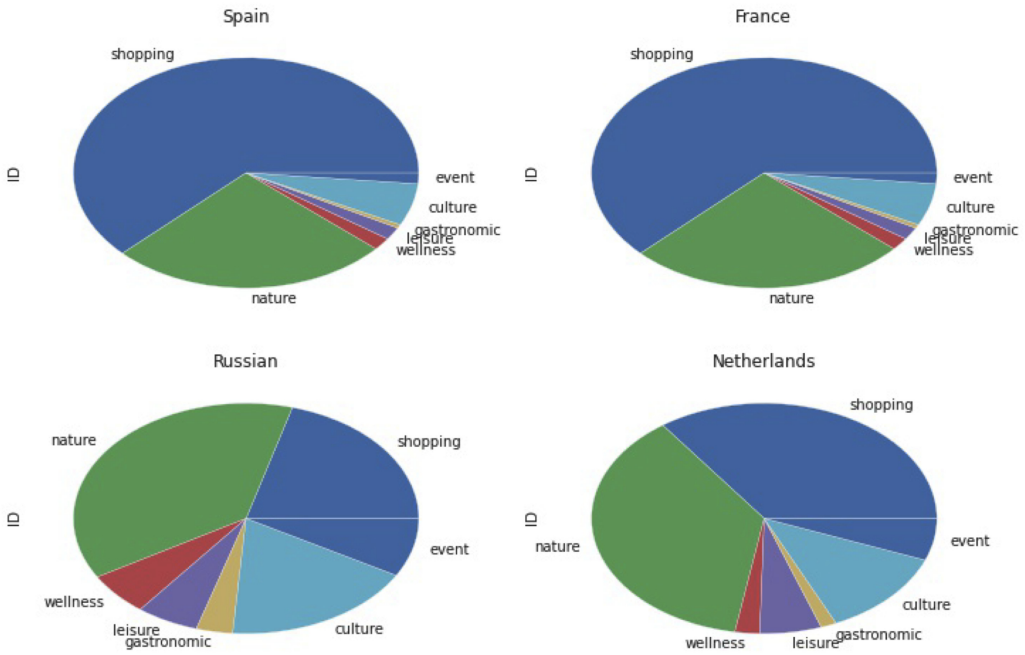


Fig. 7. Behavioral profile of tourists from four countries. The four pie plots correspond to the average distribution of interest for the Spanish, the French, Russians, and Dutch.

Table 1. Characteristics of Each Cluster of French Tourists

	Group 1	Group 2	Group 3	Group 4
Percentage of overnight tourists (%)	8	11	9	40
Number of cell towers visited in 80% of time	2.6	4.1	3.1	4.1
Radius of gyration	3.33	7.5	4.8	7.5
Length of stay (days)	1.10	1.03	1.07	1.03
Inter-visit time (days)	15	15	101	26

of gyration, length of stay, and inter-visit time to understand the characteristics of French visitors. We compute these mobility indicators via the Python Bandicoot package.[5] We first perform k-means clustering on these tourists, using the features extracted from January 2015 to July 2015. To allow for the analysis, we only include tourists with more than two records. Then, we analyze the typical characteristics of each cluster, as shown in Table 1. Finally, we describe their characteristics in Table 2. Groups 1 and 2 are frequent visitors to Andorra. Group 1 consists of day-trippers and may be a profitable market for shopping sales. Group 2 are mountain lovers and are active across a large area; hence, they might be attracted by bicycling and skiing activities and events. Group 3 prefers to visit the suburban areas and visit Andorra less frequently. The tourism department can target this group of tourists for natural activities in the mountain areas. Group 4 is active at night; hence, advertisements on bars and hotels may be more appealing to them.

[5]The Python Bandicoot package is developed for analyzing mobile phone data, and can be accessed through this link.

Table 2. French Tourists Segmentation

Group	Descriptions	Percentage
1	Frequent and night inactivate	42%
2	Frequent and visit large areas of Andorra	30%
3	Suburban lover and infrequent	15%
4	Active at night, mainly visit the downtown of Andorra	12%

4 CASE STUDY ON SPECIAL EVENTS IN ANDORRA

In this section, we use the developed indicators in Section 3 to showcase a comprehensive assessment of the high-impact events in Andorra based on CDR. The goal is to show how to monitor event performance and reveals insights that are not available through traditional tourism surveys. We present three case studies. In the first case study, we evaluate different indicators related to tourists counts to understand how to break down and understand the numbers, specifically through new and revisit tourists. The capability to analyze new and revisit tourists highlights the advantage of extracting behavioral information through longitudinal information.

In the second and third case studies, we present the performances of a series of events in a radar plot, which provides a succinct visualization of the performance of events via different indicators.

4.1 In-Depth Analysis of the Number of Tourists

We start by suggesting new ways to dig into the tourist volume. We evaluate the performances of four winter events in the year of 2015, including:

- Free-ride Junior World Championship: 02/07/2015–02/08/2015.
- Speed Skiing World Championship: 2015/02/28/2015–03/03/2015.
- Alpine Skiing European Cup Finals: 03/18/2015–3/22/2015.
- Total Fight Master of Freestyle: 03/26/2015–04/04/2015.

In the winter events, we focus on tourists, tourist days, new tourists, and whether the new tourists will revisit. It is common to measure tourist flows during events to assess the popularity of the events. However, more information about the new tourists and whether they revisit Andorra or not to are needed to evaluate (1) the marketing strategy before the events; and (2) tourist experiences during the events.

Tourists and tourist days represent the popularity of the event. However, are popular events successful in attracting new tourists? More importantly, will they turn new tourists into loyal tourists? As shown in Figure 8,[6] Free-ride Junior World Championship, though less popular than most other events, attracted a large number of new tourists and revisit visitors. This observation suggests the event's success in not only the marketing strategy before the events but also the operation and organization of the events. Contrarily, though Total Flight Master of Freestyle has a high popularity, the lowest revisit rate among all winter events implies the most impoverished travel experiences compared with other events. The performance on these indicators shows that the tourism department should investigate how to improve this event to retain tourists. For example, the tourism department can perform a regression analysis to see which factors contribute to the low revisit rate, as discussed in Section 3.3.

[6]This figure can be found in our preliminary study on a non-archival conference proceeding via this link.

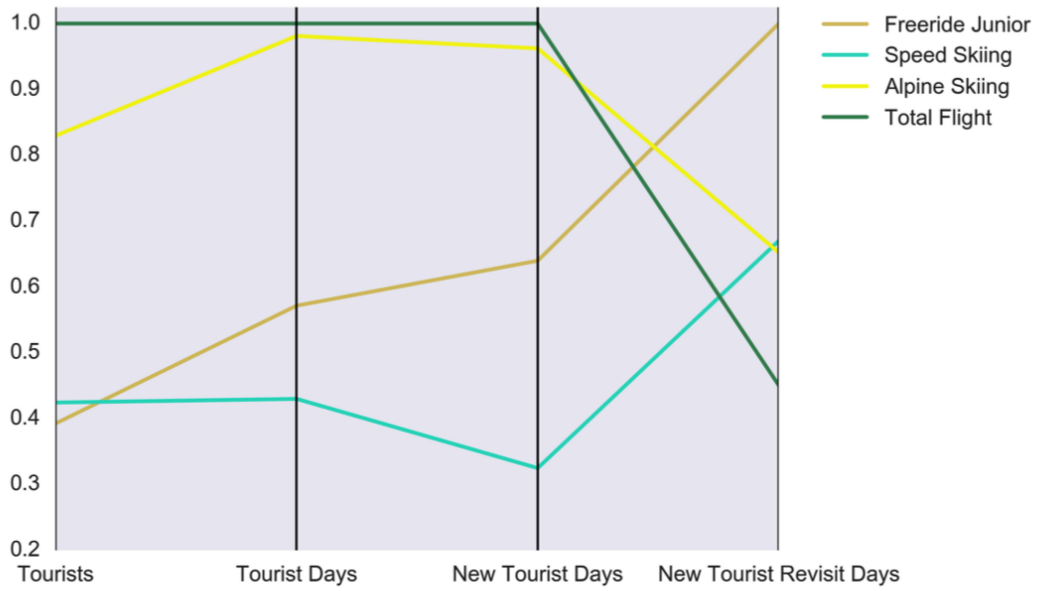


Fig. 8. Winter events analytics. The x-axis shows four indicators. The y-axis shows the relative performance of each event, compared with the best event in that indicator. That is to say, one on the y-axis is the best each event can perform in each dimension.

4.2 Ranking Events by Their Relative Performances

Next, we analyze the events by ranking them with their relative performances on more indicators. We start from eight indicators on five Summer events in 2015, followed by 10 indicators on 13 events in 2016.

4.2.1 Summer Events in 2015. Every summer, the Andorra government host different sports (mainly bicycling) and cultural events to attract more customers. Figure 9[7] shows the performances of five summer events on a radar chart, based on tourist days, new tourists, spatial distribution, approximate spending habits (average price of the phone), and peak congestion. We rank these summer events based on their performances on different indicators. The farther out the event locates on the radar plot, the higher it ranks among all indicators. Precisely, for all indicators, except for "peak congestion," more superior performances receive higher ranks. We select the following five high-impact summer events:

- Storia by Cirque du Soleil: 07/15/2015–07/20/2015.
- Volta als Ports d'Andorra: 07/12/2015.
- MTB Masters World Championships: 08/23/2015–08/27/2015.
- Tour of Spain: 08/31/2015–09/03/2015.
- UCI Trial Masters World Championships: 09/01/2015–09/06/2015.

From Figure 9, we obtain the following insights on the events. MTB World Master Championships attracted a large number of tourists and the most significant number of new tourists. These statistics highlight its success in the marketing campaign. The results also suggest more efforts are required to improve on the following aspects. First, the tourism department and event organizers

[7]This figure can be found in our preliminary study on a non-archival conference proceeding via this link.

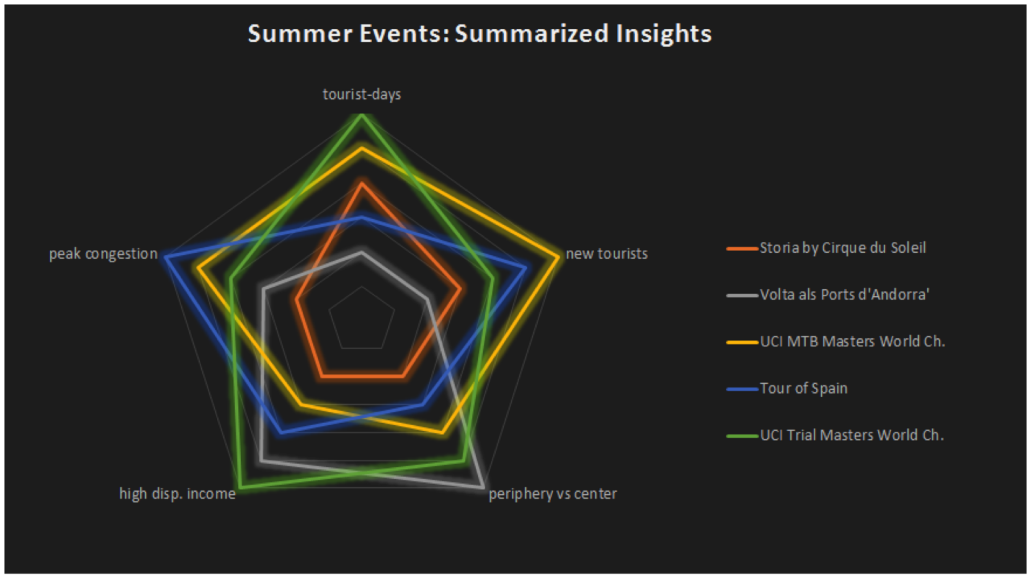


Fig. 9. Summer events analytics. The relative position on the radar plot indicates the relative performance in the five events. Each pentagon represents each indicator, and each color corresponds to a single event.

should design marketing strategies to attract more tourists to the suburban areas. Second, planners should organize events to develop new businesses in attracting higher-income visitors. Third, the transportation and tourism department should jointly manage travel demand and provide more transit supply to reduce congestion [21].

UCI Trials Master World Championship performed better than other events in most indicators measured in the radar plot. This event was the most popular one of all five events, as it has the most tourist days. The attendees used the most expensive cell phones compared with other events; hence they might have generated more revenue per capita than other events. It also performed well in that individuals distributed across the periphery areas, rather than crowded into the city center. It ranked third in both congestion and new tourists.

The year of 2015 was the first year Andorra organized the Cirque De Soleil. The government spent a large amount of funding to advertise and support this event. For example, the government heavily subsidized and offered a free ticket to this performance. As expected, the Cirque de Soleil attracted tourists with the lowest average disposable income. There was a smaller number of new tourists, suggesting the need for improved marketing. Tourists gathered in the center of the country close to the event venue. Therefore, this event did not attract visitors to explore different cities of Andorra. Fortunately, this event produced the least amount of congestion, since there were not that many tourists and not many tourists traveled across the country. Better marketing, organization, and transportation strategies are needed to improve the performance of Cirque De Soleil. For example, the government can send coupons to tourists to encourage traveling to other cities. Moreover, better marketing strategies are needed to attract new customers. For example, they can spend more effort in the Netherlands and Russia, where there exist strong interests in cultural events, as shown in Figure 7.

Tour of Spain ranks second in its appeal to new tourists. It generates the highest amount of congestion due to the high concentration of tourists at the country center. Comparatively, Volta als Ports d'Andorra is a local event where participants cycle around the county. It distributes visitors

to different cities and generates a small amount of congestion. However, it attracts the least amount of tourist days and new tourists. Event organizers can further investigate the differences in organizational and marketing strategies of these two events and improve accordingly.

4.2.2 Analytics on 13 Events in 2016. The organizers and planners can flexibly choose the indicators that are of interest to them. Let us present another example with more indicators, as shown in Figure 10. We chose 13 events from January to May in 2016. To avoid an overly dense radar plot, we separate them onto two figures based on the months the events were held. The ranking in the radar plot corresponds to the rank among all the chosen events. The outer they are on the radar plot, the higher the rank of the event. In other words, except for congestion, a value of 13 on the radar plot corresponds to the best performance among all events. A value of one corresponds to the best performance in congestion.

Comparing the two radar plots, we can see that events in January and February generally performed better than events from March to May. January and February are the high skiing seasons for Andorra. On the one hand, tourists have more holidays. However, ski slopes are the country's biggest attraction, and these winter months attract more snow sport lovers.

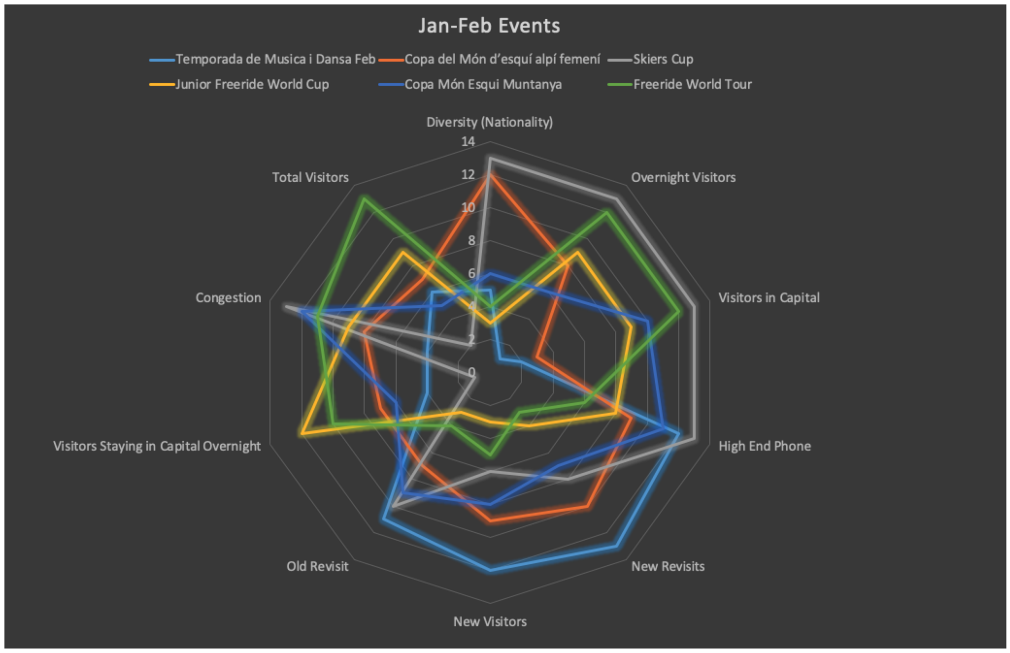
During January and February, the Skiers Cup performed the best across all events in terms of the diversity of nationality, percentage of overnight tourists, and its appeal to tourists with high-end phones. However, a large number of tourists at the capital resulted in severe congestion, even though the total number of tourists was small. Similarly, we observe that Copa Mon Esqui Muntanya, which is the mountain ski competition, generated the second-highest congestion; however, it does not attract as many tourists, compared with other events. These indicators highlight the importance of improving travel demand management to avoid the negative impact of congestion on both tourists and residents [21]. Freeride World Tour performed the best in the popularity and attracting tourists to stay overnight. Tourists centered mostly in the capital both during the day and at night. Therefore, this indicates a need for museums, hotels, and ski resorts in other cities to attract more customers.

There were three music events in the series of Temporada de Musica i Deansa that happened separately in February, March, and May. The music event in February and March performed similarly well, especially in attracting tourists with higher-end phones, new tourists, and attracting them to revisit. Since the music event was appealing to a different tourism market, this suggests the tourism department to partner with event companies to serve a different market other than traditional bicycling and skiing events. The music event in May, Temporada de Musica i Deansa, however, performed differently than the earlier two. It attracted the highest percentage of revisited tourists. Moreover, in this event, a more significant number of tourists visited the capital where the event was hosted, compared with the one in February and March. Tourists who attended this event for the first time might tend to explore other areas of Andorra. Moreover, these first-time tourists were more likely to stay in the capital overnight, which possibly contributed to a higher income for hotels in the capital. This indicator also suggests hotels in other cities to provide better services in the competition, as discussed in Section 3.4.

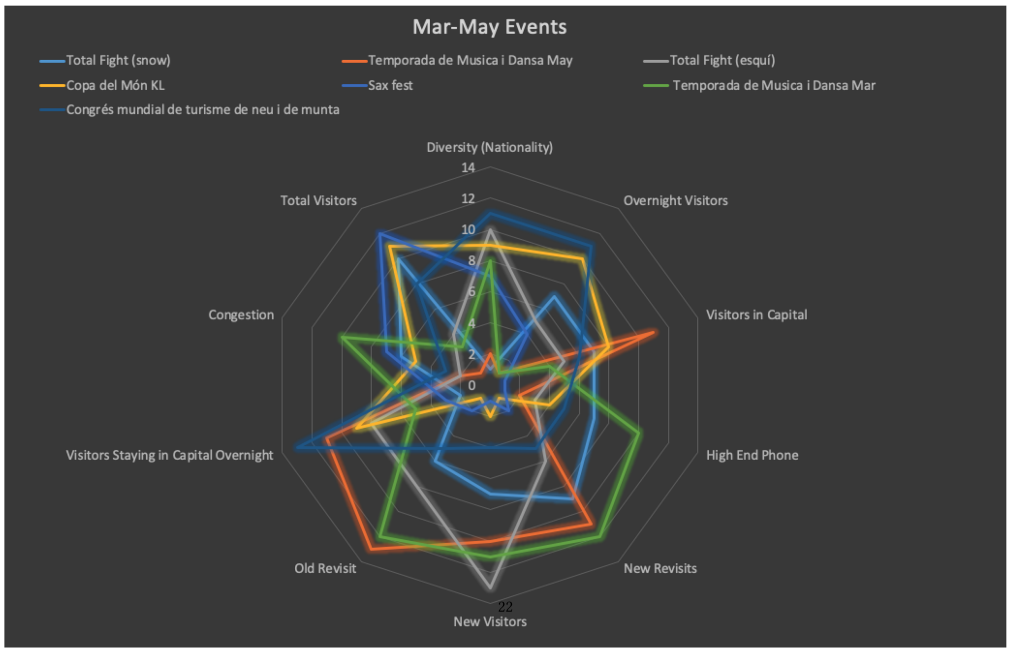
The World Snow and Mountain Tourism Congress, named Congreso mundial de turisme de neu i de munta, attracted the most significant number of tourists to stay overnight in the capital. Since tourists stayed longer in this event, most of whom stayed in the capital, this advises hotels in other cities to provide better service at a lower price to attract tourists, e.g., offer shuttles to the event venue in the capital. Or they can partner with travel agencies to improve marketing.

5 CONCLUDING REMARKS

Tourism accounts for a significant share of GDP and labor force globally. It is increasingly important economically, socially, and environmentally. Traditionally, tourism policies rely on surveys



(a) Analysis on the events events from January to February



(b) Analysis on the events from March to May

Fig. 10. Analysis on the events from January to May in 2016. The relative position on the radar plot indicates the relative performance across all events. Each decagon represents each indicator, and each color corresponds to a single event.

and questionnaires, which are costly and labor-intensive. Anonymous CDR is a new source of data, with enormous potential in areas of high societal value, such as tourism and urban development. In this article, we highlight several advantages of CDR in event analytics and tourism management. First, CDR monitors tourism at a high temporal granularity while it is costly to conduct tourism surveys regularly. Second, CDR is an opportunistic dataset, initially collected for billing purposes. Therefore, it does not incur extra costs to obtain such data. Third, CDR-based tourism analysis enables us to extract more insights than traditional surveys. For example, CDR can inform the tourism department about the number of new visitors for an event, which enables them to evaluate their marketing strategy. CDR can also advise the tourism department about the percentage of tourists who will revisit, which enables them to evaluate tourist experiences during tourism events. CDR also provides behavioral information about travelers, useful for customer segmentation. Last, behavioral information provides a more reliable source of information about tourist experiences, while surveys can suffer from the stated preference bias. The application of CDR in tourism is promising in providing timely and comprehensive insights about tourist experiences and sheds light on what needs to be improved.

This study demonstrates a new evidence-based decision-making method for the tourism industry based on the added value of CDR. We focus on the design, analysis, and evaluation of tourism strategies at local and national levels. In Andorra, we use CDR to evaluate marketing strategies, understand tourist experiences, assess revenues, and externalities generated by touristic events. We do this by extracting many indicators at a high spatial and temporal resolution, such as tourist flows, new tourists, revisit patterns, overnight tourists, tourist externalities on transportation congestion, the spatial distribution of tourists, approximate spending habits, profiling of tourist interests and tourist segmentations. Some of the indicators mentioned above are difficult or costly to measure by the traditional surveys. We demonstrate the use of these indicators for the planning and evaluation of 22 touristic events in Andorra, including cultural festivals, music events, and sports competitions, to demonstrate how to use this information for tourism planning and management.

Our study provides insights and has many practical applications for tourism management. The method and metrics we proposed in this article can be used as a prescriptive and diagnostic tool to evaluate the impacts of the events. First, the CDR-derived indicators enable tourism departments and event organizers to monitor the event performances dynamically. The performance of the events on different metrics allows policy-makers and organizers to evaluate the events. Our metric allows practitioners and decision-makers to pinpoint the areas to be improved. Practitioners can easily compare across different events and learn from the policies of more successful ones.

Second, the information on new tourists enables organizers and policy-makers to evaluate their marketing strategies. Moreover, the rich knowledge from CDR helps tourism department design strategies targeting different tourist segmentation. For example, the Andorra government needs to design various marketing strategies for Russia and Spain as they present different behavioral patterns. For instance, most Spanish tourists are day-trippers, while Russia tends to stay longer. Moreover, the Russians exhibit more interest in cultural events than the Spanish. With this information, the hospitality industry, museums, and cultural activities may spend more effort on advertisements in Russia.

Third, some positive effects of tourist events may come with negative externalities that need to be measured together with the benefits. For example, as shown by Figure 9, events with a large number of visitors may result in congestion with unmanaged travel demand and insufficient transit supply. Therefore, tourism and the transportation department should work together to balance the tourist flows. For example, Leng et al. propose a method to make location recommendations while balancing traffic flows [21]. Moreover, the event organizers could design policies to reduce

the crowdedness, such as plan the event across multiple days or distribute the event in different cities.

Our study is not without limitations and therefore points out several future directions. First, CDR might not provide an exact measurement of the event performance. For example, the new tourists in the data may be organic tourists rather than attracted by the advertisement. Moreover, individuals who connect to the cell tower close to the event venue may be pass-bys, rather than real attendees. Therefore, we need to combine high-resolution CDR with traditional surveys and government statistics to reduce the bias in the CDR. Second, future research can integrate the CDR with the contextual information from other social media platforms (e.g., TripAdvisor, Twitter, and Facebook) to combine different sources of information. Combining the contextual and textual information with the passively-recorded CDR can reveal much more insights on tourist behaviors. We can use natural language processing techniques (e.g., sentiment analysis and entity extraction) to better understand user experiences. Third, further studies can perform analysis on longitudinal data to evaluate whether improving certain aspects of events indeed boosts the performances. Moreover, we can estimate the effect of personalized interventions. For example, we expect to see more practical applications of our method and other large-scale behavioral data in generating insights in different cities and countries. An automated open platform will help streamline the tourism analytics and event management process.

REFERENCES

[1] Andorra. 2015. The World Facebook. CIA. Retrieved on October 2019 from https://www.cia.gov/the-world-factbook/countries/andorra/.

[2] Rahmi Akçelik. 1991. Travel time functions for transport planning purposes: Davidson's function, its time dependent form and alternative travel time function. *Austr. Road Res.* 21, 3 (1991).

[3] R. Baggio. 2016. Big data, business intelligence and tourism: A brief analysis of the literature. In *IFITT Workshop on Big Data & Business Intelligence in the Travel and Tourism Domain.*

[4] Jacint Balaguer and Manuel Cantavella-Jorda. 2002. Tourism as a long-run economic growth factor: The Spanish case. *Appl. Econ.* 34, 7 (2002), 877–884.

[5] Aleix Bassolas, Maxime Lenormand, Antònia Tugores, Bruno Gonçalves, and José J. Ramasco. 2016. Touristic site attractiveness seen through Twitter. *EPJ Data Sci.* 5, 1 (2016), 12.

[6] Moshe Ben-Akiva and Michel Bierlaire. 1999. Discrete choice methods and their applications to short term travel decisions. In *Handbook of Transportation Science.* Springer, 5–33.

[7] Amit Birenboim. 2016. New approaches to the study of tourist experiences in time and space. *Tour. Geogr.* 18, 1 (2016), 9–17.

[8] Vincent D. Blondel, Adeline Decuyper, and Gautier Krings. 2015. A survey of results on mobile phone datasets analysis. *EPJ Data Sci.* 4, 1 (2015), 10.

[9] *Andorra.* Wikipedia. Retrieved Aug 2021 from https://en.wikipedia.org/wiki/Andorra.

[10] Phil Crowther, James Bostock, and John Perry. 2015. Review of established methods in event research. *Event Manage.* 19, 1 (2015), 93–107.

[11] Xiaowen Dong, Alfredo J. Morales, Eaman Jahani, Esteban Moro, Bruno Lepri, Burcin Bozkaya, Carlos Sarraute, Yaneer Bar-Yam, and Alex Pentland. 2019. Segregated interactions in urban and online spaces. *EPJ Data Science* 9, 1 (2020), 20.

[12] Alain Dupeyras and Neil MacCallum. 2013. Indicators for measuring competitiveness in tourism. *Indicators for Measuring Competitiveness in Tourism: A Guidance Document*. OECD Tourism Papers, 2013/02. OECD Publishing.

[13] Enrique Frıas-Martınez, Graham Williamson, and Vanessa Frıas-Martınez. 2013. Simulation of epidemic spread using cell-phone call data: H1N1 case study. In *Proceedings of the 3rd Conference on the Analysis of Mobile Phone Datasets (Netmob'13).*

[14] Donald Getz and Stephen Page. 2019. *Event Studies: Theory, Research and Policy for Planned Events.* Routledge.

[15] Donald Getz and Stephen J. Page. 2016. Progress and prospects for event tourism research. *Tour. Manage.* 52 (2016), 593–631.

[16] Kate Kelley, Belinda Clark, Vivienne Brown, and John Sitzia. 2003. Good practice in the conduct and reporting of survey research. *Int. J. Qual. Health Care* 15, 3 (2003), 261–266.

[17] Ji Yeon Lee. 2009. *Investigating the Effect of Festival Visitors' Emotional Experiences on Satisfaction, Psychological Commitment, and Loyalty*. Texas A&M University.
[18] Yan Leng et al. 2016. *Urban Computing Using Call Detail Records: Mobility Pattern Mining, Next-location Prediction and Location Recommendation*. Ph.D. Dissertation. Massachusetts Institute of Technology.
[19] Yan Leng, Xiaowen Dong, Esteban Moro, et al. 2018. The rippling effect of social influence via phone communication network. In *Complex Spreading Phenomena in Social Systems*. Springer, 323–333.
[20] Yan Leng, Jinhua Zhao, and Haris Koutsopoulos. 2021. Leveraging individual and collective regularity to profile and segment user locations from mobile phone data. *ACM Transactions on Management Information Systems (TMIS)* 12, 3 (2021), 1–22.
[21] Yan Leng, Larry Rudolph, Jinhua Zhao, Haris N. Koutsopolous, and Alex 'Sandy' Pentland. 2017. Synergistic data-driven travel demand management based on phone records. In *Proceedings of ACM KDD UrbComp.17, Halifax, Nova Scotia, Canada*. 9 pages.
[22] Yan Leng, Yehonatan Sella, Rodrigo Ruiz, and Alex Pentland. 2018. Contextual centrality: Going beyond network structures. arXiv:1805.12204. Retrieved from https://arxiv.org/abs/1805.12204.
[23] Xiang Robert Li, Chia-Kuen Cheng, Hyounggon Kim, and James F. Petrick. 2008. A systematic comparison of first-time and repeat visitors via a two-phase online survey. *Tour. Manage.* 29, 2 (2008), 278–293.
[24] Robert Pettersson and Donald Getz. 2009. Event experiences in time and space: A study of visitors to the 2007 World Alpine Ski Championships in Åre, Sweden. *Scand. J. Hospit. Tour.* 9, 2–3 (2009), 308–326.
[25] Sara Proença and Elias Soukiazis. 2008. Tourism as an economic growth factor: A case study for Southern European countries. *Tour. Econ.* 14, 4 (2008), 791–806.
[26] Haiyan Song, Larry Dwyer, Gang Li, and Zheng Cao. 2012. Tourism economics research: A review and assessment. *Ann. Tour. Res.* 39, 3 (2012), 1653–1682.
[27] Tun Thiumsak and Athapol Ruangkanjanases. 2016. Factors influencing international visitors to revisit Bangkok, Thailand. 4, 3 (2016), 220–230.
[28] UNTWO. 2015. World tourism barometer, April 2015. Retrieved Oct 2020 from https://www.e-unwto.org/doi/abs/10.18111/wtobarometereng.2015.13.2.1.
[29] Brian Wheatman, Alejandro Noriega, and Alex Pentland. 2016. Electricity demand and population dynamics prediction from mobile phone metadata. In *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 196–205.
[30] Spencer A. Wood, Anne D. Guerry, Jessica M. Silver, and Martin Lacayo. 2013. Using social media to quantify nature-based tourism and recreation. *Sci. Rep.* 3 (2013), 2976.

Received November 2020; revised May 2021; accepted June 2021