Analyzing Call Detail Records to Support Tourism Statistics in Saudi Arabia: An Exploratory Study

Zeyad Alawwad, Wail Alkowaileet, Sattam Alsubaiee and Anas Alfaris
Center for Complex Engineering Systems at KACST and MIT

Munthir M. Alansari
Saudi Commission for Tourism and National Heritage, Saudi Arabia

Ahmad Muhammad Ragab
Faculty of Tourism and Hotels, Minia University, Egypt

Abstract:
Following the dramatic rise in data collection in recent years, considerable attention has been directed towards adopting innovative methods to compile tourism statistics from non-traditional sources. Many initiatives have emerged globally that adopt big data sources for tourism statistics (e.g. mobile positioning data, booking and ticketing data, web searches, social media, etc). The purpose of this paper is to explore the potential use of call detail records (CDRs) as auxiliary sources for reinforcing and improving the quality of tourism statistics in Saudi Arabia. Using CDR data, the Center for Complex Engineering Systems at KACST and MIT has identified the home locations of anonymous users and analyzed their movement patterns to analyze the destinations of domestic tourists from various cities throughout the country. We use this to determine the flow of tourists between administrative regions and compare the results to statistics collected from traditional face-to-face surveys. Additionally, we explore popular destinations for domestic tourists and demonstrate how the sources of tourists can be used to study attractions that are commonly visited by people who live farther away and compare them to destinations that typically receive visitors from nearby cities. Finally, the positions of tourists can be examined in detail to reveal the precise locations of tourist attractions. The possibility of using these passive mobile positioning data for supporting traditional domestic tourism surveys is discussed with the intention of serving the Tourism Statistics System (STS) in Saudi Arabia.

Keywords:
Tourism Statistics, Big Data, Call Detail Records, Tourism in Saudi Arabia.

Introduction
Over the decades, it is quantitatively proved that tourism is one of the fastest growing economic sectors nationally as well as at the global level. A record total of around 1.2 billion cross-border tourists travelled worldwide in 2015 and international tourist flows grew by 4.4% (UNWTO, 2016a). Many developed and developing countries depend largely on the tourism sector as a major catalyst for social and economic development. The UNWTO (2016b) reported that tourism represented 7% of the world’s exports in 2015, making up to 10% of global Gross Domestic Product (GDP), and accounting for 9.5% of total worldwide employment. The aforementioned figures reflect only the significance of international tourism (overnights), although when
considering same-day visits as well as domestic tourism, between 5 to 6 billion domestic visitors (UNWTO, 2016b), the total value of tourism would be significantly higher.

Owing to these substantial impacts of tourism, having reliable and accurate data on tourism is critical for policymakers to develop robust tourism strategies and to make effective decisions. Dwyer and Pham (2013) stated that “Tourism data tend not to be well incorporated in the complex system of official statistics, and often do not receive the full attention they deserve. Within most existing statistical systems, it has been extremely difficult to adequately document the full scale and scope of tourism-related economic activities”. In response to the need for improved tourism statistics, many countries and international organizations have sought to develop a standard statistical framework for tourism with comparable principles, concepts and definitions, and classifications. This lead to the establishment of international statistical standards for tourism as well as the development of Tourism Satellite Accounts (TSA), a methodological framework for economic statistics of tourism (UNWTO, 2010a; 2010b). Over 70 countries or territories around the world have adopted the TSA framework with different levels of implementation (Marion, Massieu, Meis & 2006; Aydin, 2008; UNWTO, 2010c, UNWTO, 2011). Many countries go beyond the standard TSA by articulating ad hoc analytical extensions that provide further understandings of other tourism aspects in destinations as well as to analyze tourism performance at both national and subnational levels (Ragab & Meis, 2016). Additionally, many countries work on the electronic modernization of data collection methods as a key element of a tourism statistics system (Sabato, Torrice, & Morricone, 2012; Tremblay, Chahdi, & Denis, 2014). Tourism surveys were conducted by means of the PAPI technique, yet most countries are transforming to use electronic techniques such as CATI, CAWI, and CAPI (EUROSTAT, 2014).

Innovative methods for compiling tourism statistics from non-traditional sources have gained considerable attention over the last decade. Many initiatives have emerged globally that adopt big data sources for tourism statistics (e.g. mobile positioning data, booking and ticketing data, web searches, social media, etc.). This current paper aims at exploring the potential use of call detail records (CDRs) as auxiliary sources for reinforcing and improving the quality of tourism statistics in Saudi Arabia. A number of research papers have used mobile phone data as innovative data sources for tourism statistics (Ahas et al., 2007; Ahas, 2008; Eurostat, 2014; Heerschap et al., 2014; Aizawa, Henzan & Odawara 2014). One of the first attempts to study the use of mobile phone records for data collection in tourism was initiated in Estonia where Ahas et al. (2007) studied the seasonality of foreign tourist flows within Estonia along with their geographical movement patterns. That study was conducted by analyzing 9.2 million roaming activities of 720,000 different foreign mobile phone IDs, originated from 70 countries, in 224 Estonian municipalities during 387 days; from 1 April 2004 to 21 April 2005. The results were used to supplement tourism statistics by identifying the spatiotemporal patterns of different foreign tourists in Estonia. Although the results showed a relatively good correlation with the accommodation statistics database in Estonia, the authors noted several limitations. Since tourists do not always use roaming services and 25 municipalities lacked mobile phone towers, a portion of tourist trips could not be captured. Additionally, the authors identified tourists based
In 2014, EUROSTAT published the findings of a comprehensive study aimed at exploring the feasibility of using mobile positioning data for tourism statistics. The study aimed to evaluate the possibility of obtaining mobile phone data in EU countries and assess the practicality to utilize such data for tourism measurements. The authors found that access to mobile positioning data often poses severe constraints due to regulatory confines. However, that study stated that the use of mobile positioning data provided numerous insights into tourism behavior. In an analysis of domestic tourism statistics, the study found that mobile positioning data resulted in the highest absolute number of domestic trips compared to traditional data sources such as demand surveys and accommodation statistics. This could be ascribed to the sample size, coverage, and response rates.

Through the literature review, it can be determined that the usage of data mobile positioning techniques as a data source for tourism statistics was is still in its early stages and requires more in-depth analytical models and calculations than that of customary data collection and investigation processes. As mentioned previously, the central objective of this current paper is to use CDR data as a source for domestic tourism data by identifying the home locations of anonymous mobile phone users and analyzing their observed movement patterns. The following study outlines the methodology used to extract tourism statistics from call detail records and compares the results to figures collected from traditional domestic tourism surveys. The results were used to identify areas that received a high number of tourist visits and analyze spatial distribution of the visitors’ home locations to observe which sites are capable of attracting tourists from farther away. Additionally, by examining individual locations that a tourist visits throughout their trip, it is possible to identify the locations of specific tourist attractions.

**Context of the study: The Kingdom of Saudi Arabia (KSA)**

Tourism has been identified as a major driver of growth and change in the national social and economic development plans of the Kingdom of Saudi Arabia. In 2015, the country received...
approximately 1 million inbound tourist trips in addition to an estimated 1 million domestic tourist trips. Recent figures show that tourism appears to play a significant role in the current Saudi economy, accounting for approximately 3.5% of Saudi Arabia’s GDP (MAS, 2016). To study the patterns and effects of tourism in Saudi Arabia, the National System of Tourism Statistics in the KSA has been established and managed by the Centre for Tourism Information and Research (MAS) within the Saudi Commission for Tourism and National Heritage (SCTH). The MAS center conducts annual face-to-face households survey to measure the domestic tourism in the KSA. The target monthly sample size is around 2000 households. The “usual environment” criteria applied in the KSA context to differentiate between visitors and non-visitors are defined as an area of 80 km radius around the person’s place of permanent residence or those places that are visited more frequently than four times a month. Any trip undertaken within this area or more than this frequency limits is said to be within the person’s usual environment and does not constitute domestic tourism (MAS, 2013).

Methodology

Call Detail Records

The call detail records that were used in this analysis provide information on the mobile phone activity of millions of anonymous users throughout the Kingdom of Saudi Arabia. Whenever a user interacted with their mobile service provider, the time of activity and cell phone tower used were logged in the records along with an anonymous user ID and additional details of the activity performed. The methodology outlined below was used to approximate the locations of users throughout Saudi Arabia for the duration of the data set (December 2015).

Based on the mobile phone usage of each user, a trajectory was constructed to define all of their detected movements throughout the data set. A user’s trajectory consists of a sequence of visited locations and the corresponding time stamps, effectively representing a time series of that individual’s location. The trajectory was used to define users’ home locations whenever possible and, if enough data was available, the time spent at each location was estimated.

Stay Point Algorithm

One of the limitations of cell phone data is that a user’s location is often only defined by the coordinates of the connecting mobile phone tower. Based on network traffic and the distance of the surrounding towers, a user might switch between multiple towers without moving significantly from their original location. Additionally, users in motion may access towers as they pass through an area even if they do not remain there. A stay point algorithm was used to distinguish genuine tourist trips from travelers who are passing by and changes in cell towers. It identified changes that occurred over a small enough time window (within 15 minutes) or distance (less than 500 meters) to filter out transitory locations and merge nearby locations into a single stay point to reduce noise in the user trajectories (fig 1).
Once a user’s trajectory was defined, their home location was detected by iterating over the user’s stay points and identifying the most frequent stay point between 7 pm and 6 am. Since this algorithm requires on a sufficient number of stay points to reliably detect a user’s home locations, approximately 5% of users did not have an identifiable home and were excluded from any additional analysis. Additionally, the use of stay points improved the accuracy of users’ locations by using the coordinates of multiple towers to approximate their position. This technique provides more accurate home locations and helped refine the high-resolution analysis of attraction sites that is presented in the final section of this paper. Further details regarding the stay point algorithm employed in this study can be found in Toole et al (2015).

Filtering Tourists

We defined tourists based on the following guidelines established by the Saudi Commission for Tourism and National Heritage:

1. The destination must be at least 80 km away from the user’s home location
2. The user must have spent at least one night at the visited destination
3. The trips cannot occur more frequently than once a week

The guidelines listed above were followed to maintain consistency with the database that has been established by the Centre for Tourism Information and Research and directly compare the results of from our methodology to statistics collected from traditional surveys.
Figure 2: Clustering a user’s tourist destinations based on location and time window. This particular user made three trips and visited ten individual sites as a tourist.

Once a user leaves, sequences of destinations within a trajectory were clustered into a single trip when a series of locations remained within 80 km of their centroid. If the user stayed within that area for at least one night and not return to the area more than once a week, the trips from that cluster were classified as valid tourist trips and included in the analysis. To analyze tourist trips on an aggregate level, each cluster of destinations was counted as a single trip represented by the cluster’s centroid. These trips were further grouped together based on their locations into the 13 administrative regions within Saudi Arabia. To analyze the most popular attraction sites for tourists at a higher resolution, the individual locations that users had visited were compiled into a heat map that shows the concentration of tourist destinations within a discretized grid.

Results and Discussion

Destination-Origin Matrices

To recreate the Destination-Origin matrices using mobile phone data (fig 3), the aggregated clusters of destinations to define each trip. The administrative region of the destination was determined using the centroid of the corresponding cluster and the origin of each tourist was identified from the user’s detected home location. The results are compared to survey-based statistics (fig 4) to reveal a number of similarities along with several notable differences.
The results for the most popular origin-destination pairs tend to exhibit consistent patterns between the two data sets. In particular, top destinations for tend to be consistent for most origins across both data sets. For example, the two most popular destinations identified for Riyadh residents using the CDR data were Makkah and the Eastern Region. Similarly, the top three destinations for residents of Makkah and Jizan both matched the results of the survey.

The clearest differences between the two sets of results tend involve shorter trips, particularly those occurring within the same region or between neighboring provinces such as Riyadh and Qassim regions or Al-Jouf and the Arar. In the statistics extracted from call detail records, these trips often represent the largest share of the flow to several administrative regions. However, the results from the surveys show few, if any, trips between the same provinces. This may be due to the limitations in the survey’s sample size, a consequence of omissions in self-reporting or because some mobile phone users who meet the above criteria for domestic tourism do not
Additionally, since we cannot identify the reasons behind users’ trips from passive positioning data, the differences in the results may be caused by mobile phone users who are traveling for non-tourism purposes but meet the criteria for domestic tourism.

Additionally, the sparseness of the survey-based results suggests that the majority of provinces did not have any trips between them throughout the month of December. Using mobile phone data, it was always possible to detect some between any two provinces. However, in most cases the results that were missed by surveys also represented a small percentage (typically under 5%) of the total number of visitors to any particular destination, with the exception of the the results between neighboring provinces that were discussed in the preceding paragraph.

The call detail records were also used to extract tourism statistics that were not available in the surveys conducted by the Saudi Commission for Tourism and National Heritage. Based on the the time spent away from the user’s usual environment (within 80 km of their home location), it was possible to estimate the total number of nights associated with tourist activity throughout the duration of the data (fig 5). Since several expenses such as food and accommodation tend to grow grow in proportion to the duration of tourists’ trip, these results can serve as a valuable proxy for the total tourist expenditure in a region.

The call detail records were also used to extract tourism statistics that were not available in the surveys conducted by the Saudi Commission for Tourism and National Heritage. Based on the the time spent away from the user’s usual environment (within 80 km of their home location), it was possible to estimate the total number of nights associated with tourist activity throughout the duration of the data (fig 5). Since several expenses such as food and accommodation tend to grow in proportion to the duration of tourists’ trip, these results can serve as a valuable proxy for the total tourist expenditure in a region.

![Fig 5. The Destination-Origin matrix between the administrative regions of Saudi Arabia based on the number of nights spent by tourists to each destination in December 2012. The data was compiled from the call detail records of mobile phone users who were identified as tourists.](image)

**Common Tourist Destinations**

Once the trips of domestic tourists are identified using the criteria defined above, it is also possible to use the individual destinations within users’ trajectories to study their trips in greater detail. These trips involve millions of unique stay points that are grouped together into a grid with cell dimensions of 0.01° latitude and 0.01° longitude. If a user has at least one stay point within a cell and it satisfies the conditions for domestic tourism, they contribute an additional trip to that particular cell. We can identify popular tourist destinations by examining the cells with the greatest number of visitors.

<table>
<thead>
<tr>
<th>Tourist Home Location</th>
<th>Riyadh</th>
<th>Qassim</th>
<th>Makkah</th>
<th>Madinah</th>
<th>Eastern</th>
<th>Aseer</th>
<th>Jizan</th>
<th>Najran</th>
<th>Baha</th>
<th>Tabuk</th>
<th>Hall</th>
<th>Jouf</th>
<th>Arar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Riyadh</td>
<td>550832</td>
<td>130289</td>
<td>292438</td>
<td>51700</td>
<td>305474</td>
<td>98700</td>
<td>47871</td>
<td>42499</td>
<td>12871</td>
<td>23634</td>
<td>46952</td>
<td>19525</td>
<td>14825</td>
</tr>
<tr>
<td>Qassim</td>
<td>29160</td>
<td>29004</td>
<td>13949</td>
<td>34811</td>
<td>2241</td>
<td>1022</td>
<td>884</td>
<td>295</td>
<td>5191</td>
<td>15381</td>
<td>2843</td>
<td>2817</td>
<td></td>
</tr>
<tr>
<td>Makkah</td>
<td>424751</td>
<td>57680</td>
<td>425413</td>
<td>208000</td>
<td>144882</td>
<td>147780</td>
<td>72564</td>
<td>39051</td>
<td>65965</td>
<td>44429</td>
<td>22264</td>
<td>7941</td>
<td>5143</td>
</tr>
<tr>
<td>Madinah</td>
<td>18640</td>
<td>19689</td>
<td>216396</td>
<td>93499</td>
<td>34995</td>
<td>10524</td>
<td>4342</td>
<td>4242</td>
<td>1727</td>
<td>32272</td>
<td>11767</td>
<td>2817</td>
<td>1596</td>
</tr>
<tr>
<td>Eastern</td>
<td>282477</td>
<td>28258</td>
<td>102803</td>
<td>27814</td>
<td>468447</td>
<td>28608</td>
<td>15063</td>
<td>14204</td>
<td>4386</td>
<td>7331</td>
<td>14025</td>
<td>6480</td>
<td>14647</td>
</tr>
<tr>
<td>Aseer</td>
<td>109807</td>
<td>2364</td>
<td>121430</td>
<td>6990</td>
<td>35467</td>
<td>106519</td>
<td>66074</td>
<td>34866</td>
<td>8560</td>
<td>10843</td>
<td>969</td>
<td>556</td>
<td>490</td>
</tr>
<tr>
<td>Jizan</td>
<td>61730</td>
<td>1519</td>
<td>81466</td>
<td>4830</td>
<td>22467</td>
<td>87667</td>
<td>12947</td>
<td>16789</td>
<td>2277</td>
<td>11110</td>
<td>563</td>
<td>421</td>
<td>602</td>
</tr>
<tr>
<td>Najran</td>
<td>43513</td>
<td>1318</td>
<td>33885</td>
<td>4687</td>
<td>18377</td>
<td>35673</td>
<td>12634</td>
<td>25534</td>
<td>1172</td>
<td>1375</td>
<td>381</td>
<td>342</td>
<td>251</td>
</tr>
<tr>
<td>Baha</td>
<td>8600</td>
<td>310</td>
<td>39349</td>
<td>929</td>
<td>3225</td>
<td>6303</td>
<td>1194</td>
<td>873</td>
<td>124</td>
<td>551</td>
<td>137</td>
<td>39</td>
<td>26</td>
</tr>
<tr>
<td>Tabuk</td>
<td>19473</td>
<td>2700</td>
<td>29174</td>
<td>23711</td>
<td>7617</td>
<td>7586</td>
<td>5208</td>
<td>758</td>
<td>809</td>
<td>42688</td>
<td>4207</td>
<td>4796</td>
<td>1733</td>
</tr>
<tr>
<td>Hall</td>
<td>48402</td>
<td>18424</td>
<td>16421</td>
<td>11681</td>
<td>19411</td>
<td>836</td>
<td>402</td>
<td>588</td>
<td>189</td>
<td>7727</td>
<td>41405</td>
<td>5028</td>
<td>2569</td>
</tr>
<tr>
<td>Jouf</td>
<td>15706</td>
<td>2669</td>
<td>5176</td>
<td>2301</td>
<td>6462</td>
<td>853</td>
<td>314</td>
<td>229</td>
<td>54</td>
<td>5533</td>
<td>3215</td>
<td>9960</td>
<td>7430</td>
</tr>
<tr>
<td>Arar</td>
<td>18032</td>
<td>2957</td>
<td>3904</td>
<td>1170</td>
<td>16716</td>
<td>523</td>
<td>356</td>
<td>158</td>
<td>52</td>
<td>1640</td>
<td>2358</td>
<td>7325</td>
<td>11487</td>
</tr>
</tbody>
</table>

Fig 5. The Destination-Origin matrix between the administrative regions of Saudi Arabia based on the number of nights spent by tourists to each destination in December 2012. The data was compiled from the call detail records of mobile phone users who were identified as tourists.
The two most visited tourism sites were the two holiest sites in Islam: Al-Masjid Al-Haram (the Sacred Mosque) in Makkah and Al-Masjid An-Nabawi (the Prophet’s Mosque) in Madinah, both mosques with a long history of religious and cultural significance in Saudi Arabia that act as the primary drivers of tourism in their respective cities. Both of these destinations were able to attract large numbers of domestic tourists throughout the month from all over the country. The distribution of tourists’ home locations is in stark contrast to many other tourist attractions such the port city of Jizan located in the eponymous administrative region (fig 7).
As a mosque of unique importance in Islamic faith, Al-Masjid Al-Haram in Makkah is capable of effectively drawing in visitors from all over the country. A large number of tourists come from distant areas such as Riyadh and the Eastern Region and the distribution of visitors’ home locations is consistent with the country’s population distribution. Although the city of Jizan was able to attract a significant number of tourists, they primarily came from the nearby regions of Aseer and Makkah with some contribution from the city of Riyadh.

Identifying Tourist Attractions
In the previous section, the analysis focused on the total number of tourists who visited a general area to determine the common tourist destinations. To simplify the identification of popular destinations on a national scale, it was beneficial to use relatively large grid cells and combine individual stay points into a single trip. Once these areas are identified, however, it is possible to explore tourist destinations at higher spatial resolution by examining the individual stay points that fell within their respective cells.

Fig 8. Staypoints of tourists (red) who visited the area of Al-Masjid Al-Haram

In regions that have a sufficient density of mobile phone towers, it becomes possible to identify such tourist hubs with a high degree of spatial accuracy. For example, visitors to Al-Masjid Al-Haram in Makkah (fig 8) tended to have the highest presence south of the mosque in an area with a high density of services and amenities for tourists. Their locations were most concentrated along row of buildings along the southern boundary of the mosque. This complex contains several high-profile international hotels as well as numerous shopping centers and restaurants. In contrast, the absence of tourist trips in the area immediately northwest of the mosque was consistent with the ongoing construction projects in that location.
Conclusion

Through the analysis of call detail records, this paper has explored the prospect of supplementing traditional tourism statistics through the analysis of non-traditional sources. By analyzing the movement patterns of mobile phone users that meet the definition of domestic tourists in Saudi Arabia, Destination-Origin matrices were constructed using mobile positioning data and compared to the results of surveys conducted by MAS. Despite broad similarities in between the patterns observed in the two data sets, there were notable differences that may have been due to limitations in the survey methodology (small sample size, coverage, focusing on main destination while neglecting transit-destinations) or because some mobile phone users who meet the criteria for domestic tourism do not travel for tourism purposes. In addition, the analysis of CDR data is capable of providing additional results that were not obtained by surveys such as identifying the popular tourist destinations by examining the cells with the greatest number of visitors as well as identifying most common tourist attractions at sub-region scale levels. However, the one-month reference period places considerable limitations on the current study and availability of a more extensive data (e.g. 12 months) would provide a more comprehensive view of domestic tourism flows.

The findings of this study demonstrate that data extracted from Call Detail Records provide wide-ranging and achievable potential for tourism statistics, particularly in its capacity to provide additional details in terms of location (national, regional, sub-national, specific attraction, etc) that can support tourism planning and decision making processes. As these detailed data are difficult to collect through traditional HHs tourism surveys, the incorporation of CDR datasets can significantly enhance the analytical capacity of traditional tourism data sources. The approach could be further developed to cover other segments of domestic tourism (e.g. same-day trips) and study the effects of special tourism events and festivals in specific locations.
References


