A Work Project, presented as part of the requirements for the Award of a Master Degree in Economics / Finance / Management from the NOVA – School of Business and Economics.

EVENT MANAGEMENT IN THE ERA OF BIG DATA AND ADVANCED ANALYTICS: AN EMPIRICAL INVESTIGATION IN PORTUGAL

Lukas Jakob Mader

33507

A Project carried out on the Master in (Economics/Finance/Management) Program, under the supervision of:

Leid Zejnilović

03.01.2020

Abstract

This work project can be considered as an example of how big data and advanced analytics can be utilized to inform event management. Knowing what event visitors do, where they go and what they spend before and after their visit to a special event is highly important for event managers. With this in mind, this investigation analyzed and studied mobile call detail record data of tourists visiting a certain music event in Portugal in August 2017. The results successfully showed practical insights on how tourists moved and behaved. These results were then translated into how they can inform event management.

Keywords

Event Management, Human Mobility, Data Analysis, Economic Effects

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

In addition, I acknowledge the help from the Data Science Knowledge Center, that provided the data and the topic for this thesis.

1. Introduction

Both the event management and tourism industry have proven themselves to be important factors for the economy of many countries (Hodur & Leistritz, 2007). The intersection of the two, the so-called event tourism, has become increasingly valuable to both hosting countries and event organizers alike (Getz, Svensson, Peterssen, & Gunnervall, 2012). Mega-events such as the Olympic Games or the FIFA World Cup not only contain visitors from the host country, but attract visitors from all over the world, resulting in possibilities to accomplish economic, social and political agendas. These events and their effects have become the essence of many research papers, mostly focusing on the economic effects, specific marketing, sponsorship and emerging trends (Formica, 1998; Getz, 2000). Furthermore, on the demand side of events tourism, it is of great interest what customers do besides visiting the event, where they go and what they spend (Getz, 2008). Therefore, it becomes evident, that the analysis of special events and especially of the event visitors is of high interest and can influence the way event managers think and make decisions.

With the rise and accessibility of big data, the act of data-based analysis and decision making has become more and more used in many different forms and domains (McAfee et al., 2012). One of these is the field of human mobility. The study of human mobility patterns on both a micro- as well as macro-level has increased in recent years, due to the rise of mobile devices (Naboulsi et al., 2016). Using mobile data to track human mobility is not the only way to do so, but a very effective one, as many studies have shown and it has proven its purpose in a wide array of areas such as sociology, biology or city planning (González et al., 2008). As understanding how people move and behave is also critical for event and tourism management, the impact of big data analysis in general and especially of human mobility analysis on these areas is an interesting field to study.

By defining event attendees of a certain event within a population of tourists and analyzing their mobility patterns and behavior outside of the event, this work project will investigate how big data analysis can be utilized in order to make informed decisions in event management. The work paper will commence by introducing the field of study by defining and displaying the fields of event management and human mobility in detail through a literature review. Afterwards, the data that was used as well as the methodology will be described. After presenting the results, they will be interpreted and discussed to see how they can be utilized to inform event management.

2. Literature Review

This work project is an empirical investigation on the use of big data for event management, in the context of tourism. Therefore, the following section reviews the academic literature about event management and individual mobility, with a focus on the tourism industry.

2.1. Event Management

In order to further delve into the field of event management, a general understanding of what an event is considered to be should be established. Goldblatt (2005) describes events as "a unique moment in time, celebrated with ceremony and ritual to satisfy specific needs." A thematic literature review conducted by Getz (2008) displays a similar definition. According to his findings, events, in the context of event management, are considered to be planned phenomena, which are defined by time and space. Furthermore, each event is unique, due to an ever-changing mix of interactions, people, settings and management systems, such as design elements and the event program. This uniqueness also accounts for its appeal, as you have to be there in order to experience it. Events can vary drastically in size and scope as well as in type of event, ranging from small local events like city festivals to global mega events such as the Olympic Games. Following Ritchie (1984) as well as Hall (1992), Getz

(2008) proposes different types of event classes, with respect to their content and purpose (see Appendix A, Figure 1).

Concerning event size and scope Getz's (2008) distinction between four classifications of events is used heavily throughout the literature of event studies. Based on the factors of tourist demand and value, he makes a distinction between the following types of events: (1) Mega events, (2) Hallmark events, (3) Regional events and (4) Local events. With and increase in size and scope, the demand for tourists and the value increase along with it (see Appendix A, Figure 2).

Event Management as its own professional field is still a rather new concept with a boom in published literature and the emergence of specialized academic degrees in the 1990's (Getz, 2008). Event Management as a profession entails everything within the design, production and management of planned events, regardless of size, domain or objectives (Getz & Page, 2016). Therefore, it is based on a variety of other professions and fields of studies such as business administration, law, marketing, human resources, facility management and design (Getz, 2000). The most researched areas within the field of event management are those of economic effects of events, event specific marketing, sponsorship and emerging trends in the event management field (Formica, 1998; Getz, 2000). The diverse foundation and research areas of event management are also the root for an ongoing debate on whether it should be considered its own field of study or industry.

A field that has been linked more and more with event management and that is also the core of this paper, is the tourism industry. The term *Event Tourism* was first introduced in the late 1980's and manifested through an article by Getz in 1989 in which he developed a framework for planning *events tourism*. Getz (2008) argues, that event tourism should not be considered its own field of studies, as it relies on both tourism and event studies to make sense of it (see Appendix A, Figure 3).

Event management can utilize tourists as new markets that can be targeted and that can help an event grow, as seen in mega-, hallmark- and major-events. From a tourism and destination management side events can be utilized as a tool to drive tourism and shape the image of a certain destination (Getz, Svensson, Peterssen & Gunnervall, 2012). On the demand side it is of interest to study who and why people attend certain events while travelling, what they do besides visiting certain events and what they spend. Another part is the assessment of the impact events have in promoting a positive destination image, place marketing and destination co-branding (Getz, 2008). The supply side should inform on how events can be used to attract tourists, serve as a catalyst for increasing infrastructure, promote a positive destination image and contribute to general place marketing (Getz, 2008).

Especially the economic impact of events has been well documented in the literature (Getz, 2000). Since visitor spending can contribute to the local economy, many communities seek to enhance tourism and visitor-oriented activities. As a result, estimates of the economic impact of event tourism are of interest to a wide variety of parties (Hodur & Leistritz, 2007). A number of papers have investigated the economic impact of specific events, for example the work of Lee et al. (2017) on the economic effects of the Expo 2012 Yeosu Korea on the host and neighboring regions or the measurement of economic effects of the Formula One Grand Prix by Kim et al. (2017). Already in the early years of event tourism research, Dwyer and colleagues (2000) proposed *a framework for evaluating and forecasting the impacts of special events*. By considering the number of visitors, the different types of visitors and the length of stay, estimates of the economic value of tourists were made.

The empirical investigation conducted within this paper will focus on the consumer perspective of event tourism, as it will analyze and categorize the behavior of foreign event attendees of a certain music festival in Portugal, including estimates of the economic impact of these tourists.

2.2. Human Mobility

Studying events is closely associated with an understanding of its attendees and their behavior. A branch of science that studies human movements is known as human mobility.

2.2.1. Definition of Human Mobility

The research field of human mobility deals with collecting, analyzing and interpreting data on human motion patterns, gathering insights on where people are, where they are going and how they move. Understanding human movement patterns finds itself useful in tackling a number of different issues (González et al., 2008), some of them being urban planning, traffic forecasting, and the spread of biological and mobile viruses. Furthermore, it has enriched various fields of research (Naboulsi et al., 2016), such as Physics, Sociology, Epidemiology, Transportation and Networking. Human mobility research can be conducted in a wide variety of research models, regarding their duration and geographical scope. It can span from analyzing the mobility patterns of people within a timeframe of just a few days within the setting of a conference (Hui et al., 2005) up to analyzing data collected over several months within a whole country (Ahas et al., 2007a).

The boom of mobile services, the accompanied emergence of mobile traffic data and an increasing willingness on the side of mobile operators to share their data were main drivers for the formation of this field of study with its beginnings in 2005 and a high increase in recent years (Naboulsi et al., 2016). According to Naboulsi et al. (2008) mobile traffic data is highly suitable for analyzing human trajectories, as a vast fraction of the total population uses mobile devices.

Smoreda et al. (2013) mention, that other types of data have been used to analyze human mobility as well. Global Positioning Systems (GPS) data is both very precise and has a high recording frequency. Although aligning on these advantages, Calabrese et al. (2010) point out that the logistical efforts in conducting a study using GPS data have to be considered, as special

tracking devices might have to be bought and installed, which is cost intensive and requires explicit consents of the user. Also, GPS coverage is limited in urban areas. Therefore, the trade-off between accuracy of measures and feasibility as well as effort between different measurement styles has to be kept in mind when conducting a research study.

2.2.2. Measurement techniques of mobile traffic data

As stated in the previous section, the foundation and majority of human mobility studies lies within the analysis of data gathered through mobile devices. Within the field of mobile traffic analysis, there are multiple ways to collect and identify data, each having certain advantages as well as disadvantages. The following section will break down the most commonly used types of data and their characteristics.

Smoreda et al. (2013) propose that there are two main categories in which mobile related data can be distinguished in regard to human mobility data collection: *Active* and *passive* localization.

In *active* localization a predefined group of people allow their localization data to be tracked over a certain amount of time for the sole purpose of a study. Localization data can either be collected through network information or via mobility-aware software integrated in the device. In many cases additional data is being provided by the participants. This type of data collection allows for very high granularity of data and information density, but has the disadvantage of being easily restricted in terms of scope of duration and number of participants.

Passive localization consists of using data that is automatically generated and collected by network providers for purposes other than research (e.g. technical or billing purposes) and therefore offers easy access to enormous user populations and the possibility to collect data over almost infinite time spans. Some disadvantages regarding data frequency and granularity become evident when further looking into the three types of passive localization data defined by Iovan et al. (2013): (1) Call Detail Records (CDR), (2) WIFI and (3) Probe Data. CDR data

is based on cell phone billing records and only contains data collected during activity of a user (call, SMS or internet activity) (Smoreda et al., 2013). WIFI data is gathered as soon as a user is connected to a WIFI network (Song et al, 2004). Probe data is gathered through a mobile network probe installed in the device, which collects data on mobile activity (call, SMS or internet activity), as well as localization changes such as handovers (HO) and localization updates (LAU) (Iovan et al., 2013). HOs are localization updates noted when a device's active communication is transferred from one cell of the mobile network to another. LAUs are collected when a device changes the location area, consisting of multiple cells, even when not actively being used. Therefore, data based on these collecting methods is very restricted in space and time.

Furthermore, another issue which all methods that rely on mobile network data share is the issue of inaccurate localization information, due to the way the cellular network is set up (Smoreda et al., 2013). The cellular network consists of multiple location areas, which consist of multiple cells, each covering a certain area (Chen et al., 2016). Each cell contains one antenna that the mobile device can connect to, meaning the location given at any time is not necessarily the exact location of the mobile device, but the one from the nearest antenna. According to Jiang et al. (2013) location precision in urban is on average within 300 meters, whereas in rural area it can go up to several kilometers. This difference has to be taken into account.

Despite the limitations in recording frequency and location precision, that come with the use of mobile network data, the wide use of mobile phones throughout the world make it a highly valuable source (Smoreda et al., 2013).

2.2.3. Human mobility analysis using mobile traffic data in tourism

The use of human mobility analysis finds many applications in the field of tourism and event management. As mentioned in the previous section on event tourism, it is of interest for event

managers to know what event attendees do besides visiting the actual event (Getz, Svensson, Peterssen, & Gunnervall, 2012). Furthermore, the results of such analyses can be used to improve infrastructure, promote a positive destination image and contribute to general place marketing (Getz, 2008). It can also help understand the relationship between points of interest and the points of arrival and departure of tourists (Ahas et al., 2007a).

Human mobility analysis is also a subject of interest in tourism management. Mamei and Colonna (2018) utilized CDR data from a total of 1.2 million tourists to cluster them as tourists, residents, commuters, people in transit or excursionists based on their movement patterns. Hu et al. (2018) used CDR data to discover the mode of transportation of tourists. As it is also useful to combine multiple data sources to make informed decisions, Oliver and Enrique (2014) combined roaming data from 680.000 tourists with electronic payments of 169.000 credit cards to showcase the impact big data can have on city planning.

Furthermore, Ahas et al. (2015) identified some aspects of what makes mobile traffic data so interesting to use in tourism mobility analyses, some of them being that it provides information on tourists staying at unpaid/non-registered accommodations, it makes the identification of phenomena such as repeated visits and frequency of visits possible and it provides detailed insights in regards to the regions and the country of origin of the tourists.

The following investigation will expand the literature in this field of study by providing insights on how the analysis of mobile traffic data can inform event management in a tourism setting.

3. Data and Methodology

3.1. Data

The data used in this paper is an augmented sample of the data that was provided by a European telecommunications provider operating in Portugal. It consists of structured tourist call detail record data gathered between the first until the 30th of August in the year of 2017

all throughout Portugal, including the Portuguese islands. It contains data about 20.243 unique tourists with a total of 3.025.541 rows. The operator provided the time stamp, anonymized client identifier, country where the sim card is registered, and the latitude and longitude of the centroid (mean location or the barycenter of a cell) of a cell that a base station (to which a mobile devise is connected) covers. The remaining data about the administrative units (district, county, and municipality) and the points of interest within the cell are obtained from the external sources, as elaborated in the work by Fonseca (2019). In total, the dataset is comprised from 14 columns containing the following information:

Variable Name	Description	Value Type
event_date	A timestamp (dd-mm-yyyy hh:mm:ss).	Object
day_of_month	Value of the day within the timestamp.	Numeric
hour_of_day	Value of the hour within the timestamp.	Numeric
minute_of_hour	Value of the minute within the timestamp.	Numeric
client_id	Unique number identifying the tourist.	Numeric
origin	Country of origin of the tourist.	Object
cell_ref	Unique number identifying the cell tower the phone is connected to.	Object
label	Variable identifying specific attractions or facilities. Only contains values in district of Lisbon.	Object
centroide_latitude	Latitude value of the cell tower the phone is connected to.	Numeric
centroide_longitude	Longitude value of the cell tower the phone is connected to.	Numeric
distrito	Name of the district the connected cell tower is located in.	Object
concelho	Name of the county the connected cell tower is located in.	Object
municipio	Name of the city the connected cell tower is located in.	Object
poi	Name of the point of interest closest to the connected cell tower.	Object

Figure 1: List of Variables in Original Data Set

3.1.1. Data Preprocessing

Even though the data provided was already structured, it had to be curated to a certain extent to make it useful for the following analysis.

First of all, the dataset showed a number of rows which contained empty values in all of the localization related features. These rows were dropped, as these are the features of highest

importance and strategically inserting values (e.g. from the previous column containing values) wouldn't change the level of information provided by the dataset.

Further, as the objective of this investigation is to analyze the mobility and behavior patterns of event tourists outside of a certain event, it is crucial to first identify those tourists. Using the coordinates of the event as well as a buffering technique which creates an area around a specific point, the potential event area could be established. Also using the dates of the event, it was possible to identify those tourists that have been within the event area during the given timeframe at least ones. The level of accuracy can be questioned as (1) the buffering technique won't create the exact event area measures and (2) the overall accuracy of the localization information based on mobile traffic has to be considered, as mentioned in the literature review part of this paper. But due to the scope of this work project and to showcase how advanced analytics can be utilized, this approach is sufficient and leaves us with a number of 2.529 potential event attendees.

3.1.2. Feature Generation

Based on the dataset that was obtained through the aforementioned data preprocessing, feature generation was conducted to create new features based on the existing ones, in order to extend the pool of information on which the segmentation of tourists and analysis will be based on. The following table shows the features generated (see Appendix B, Figure 1 for detailed information on how features were generated).

Feature Name	Desciption	Value-Type
total_stay	Number of days spent in Portugal in August.	Numeric
total_nights	Number of Nights spent in Portugal in August.	Numeric
days_before_event	Number of days the tourist arrived before his/her first day at event.	Numeric
days_after_event	Number of days the tourist stayed after his/her last day at event.	Numeric
days_at_event	Number of days the tourist spent at the event site.	Numeric
mean_of_arrival	Assumption of how the tourist arrived. (Air, land, water or unknown)	Categorical
mean_of_departure	Assumption of how the tourist left. (Air, land, water or unknown)	Categorical
unique_districts_visited	The number of unique districts the tourist has been in during his/her stay.	Numeric
unique_concelhos_visited	The number of unique counties the tourist has been in during his/her stay.	Numeric
unique_pois_visited	The number of unique Points of Interest the tourist has been close to.	Numeric
total_distance_travelled	The number of kilometers the tourist travelled from beginning until end of his/her stay.	Numeric

max_distance_from_start	The distance between the tourists point of arrival and the point furthest away from that (air line).	Numeric
daily_distance	Average number of kilometers travelled per day.	Numeric
economic_value_stay	Estimated value of what a tourist has spent on accommodation.	Numeric
economic_value_food	Estimated value of what a tourist has spent on food.	Numeric
total_economic_value	Sum of economic value of accommodation and economic value of food.	Numeric
days_low_budget_area, days_mid_budget_area, days_high_budget_area	Number of days spent in areas categorized by level of expensiveness.	Numeric

Figure 2: List of Features Generated

3.1.3. Economic Value Related Variables

Unlike the remaining features, the features regarding economic value are not solely based on information provided by the dataset, but also contain external information. Therefore, their generation will be explained more thoroughly in this section.

For the economic value of night stays by tourists (*economic_value_stay*) a more accurate estimate than simply multiplying the number of nights stayed by an estimate average for accommodation was desired. Therefore, the average hotel price for a number of 248 concelhos was manually researched via Google Search and implemented in the analysis. Holding this information and assuming that the location of the last entry of the day was the location where tourists would spend the night, a more precise average for each night could be calculated. The sum of each overnight value per tourist makes up their economic value in overnight stays.

Under the assumption that the level of average prices for hotel stays within a concelho is in line with the overall level of expenditures/expensiveness of that concelho, each concelho was put into bins (low_budget_area < 40eur/night, 41eur/night < mid_budget_area < 65eur/night and 64eur/night < high_budget_area) considering the level of expensiveness. The values for the bin boundaries were chosen intuitively considering the minimum and maximum values of stays and the frequency of values. Now each day spent for each tourist was assigned to one of the bins, based on the median budget area stayed in. The median budget area was based on the most entries of concelho budget areas during a day. Due to the way the data was

collected, most entries do not necessarily equal most time spent at. But due to the low level of data granularity this method was preferred. Knowing in what kind of area the tourist spent each day, the days spent in each area were calculated, resulting in the variables *days_low_budget_area, days_mid_budget_area* and *days_high_budget_area.*

For calculating the economic value spent on food (*economic_value_food*) by tourists, these variables were used in combination with according estimates of how much a tourist would spend in one of these areas. The values used were based off of a website that estimates daily tourist expenditures within three categories: low-level spending, mid-level spending and high-level spending (budgetyourtrip, 2019). The given values for food were then matched with the respective budget area. The economic value spent on food for each tourist, was then calculated by summing up the values of days spent in a certain area multiplied by its respective value of food spent per day.

The *total_economic_value* for each tourist was then calculated by adding *economic_value_food* and *economic_value_stay* together.

3.2. The Event

The event, whose possible attendees will be analyzed in this paper, was located in the south of Portugal (37.5527 ° N, 8.731517 ° W) from 1st August until 6th August in the year of 2017.

3.3. Methodology

To study the behavior, type, movements and the economic value generated of tourists who may have attended the event, first an exploratory data analysis will be conducted. After obtaining general understanding of the data, an unsupervised machine learning technique will be applied on the dataset to generate clusters (segments) of tourists that may more effectively inform event management.

3.3.1. Exploratory Data Analysis

Exploratory Data Analysis was conducted to summarize the main characteristics of the dataset used. First, a summary table is generated, describing the information about each feature: range, mean, and standard deviation for numerical features, and percentages for the categorical features (where appropriate). For some of the categorical features of interest for event management, frequency distributions are shown as histograms.

3.3.2. Clustering

In large datasets, it is hard to act upon the individual behaviors of many tourists. Therefore, a common method to obtain actionable information is to reduce the dataset to a few groups of tourists by applying clustering algorithms. These analyze statistical information and group individuals based on their common features and statistical patterns.

The first step in clustering is to generate a correlation matrix to avoid using highly correlated features. Next, a suitable algorithm needs to be selected for the clustering. In this case, k-means was selected, as it is one of the most popular clustering methods, due to its simplicity and efficiency (Bholowalia & Kumar, 2014). Using k means, a data set is split into *k* groups, where *k* is fixed a priori (Wagstaff et al., 2001). In order to place the points within a dataset in one of the *k* clusters, a number of *k* distinct centroids is being defined. An iterative process of (1) assigning each point in the dataset to the closest cluster centroid and (2) updating each cluster center to the mean of its constituent instances, makes up the clustering process. Most commonly, for numeric features, a Euclidean distance metric is used, whereas for symbolic features, the Hamming distance is used.

To establish the number of clusters (k) in k-means, the so-called elbow technique was used (Kodinariya & Makwana, 2013). Within this technique the cost of training is calculated for different numbers of clusters, starting with k=2, and visualized as cost function (x=number of clusters, y=cost of training). Ideally the graph will show a value k where the cost drops

drastically and afterwards reaches a plateau. This is the ideal number of k, as increasing the number of clusters only has a small effect on the difference between them. Visually, at this point the graph will be bent and look similar to an elbow, hence the name.

Finally, given that there are many dimensions used to generate clusters, the t-sne visualization technique is used to represent the clusters (Maaten & Hinton, 2008). In addition, descriptive statistics of the key features for each of the clusters is provided, to offer an evidence that could lead to a comprehensive interpretation of the clusters.

4. Results

4.1. Exploratory Data Analysis of the Event Attendees

	day_of_month	hour_of_day	minute_of_hour	client_id	centroide_latitude	centroide_longitude
count	303293.000000	303293.000000	303293.000000	3.032930e+05	303293.000000	303293.000000
mean	10.340766	13.794140	29.612856	3.825761e+06	39.123039	-8.537498
std	7.228648	5.921948	17.352190	2.153658e+06	1.736818	0.672160
min	1.000000	0.000000	0.000000	1.327000e+03	32.636024	-25.402031
25%	4.000000	10.000000	15.000000	2.221087e+06	37.511501	-8.794341
50%	9.000000	14.000000	30.000000	3.454547e+06	38.722301	-8.678287
75%	15.000000	19.000000	45.000000	5.761524e+06	41.053894	-8.281872
max	30.000000	23.000000	59.000000	7.595182e+06	42.122757	-6.255675

Figure 3: Descriptive Statistics of Original Data Frame with all entries of Event Attendees.

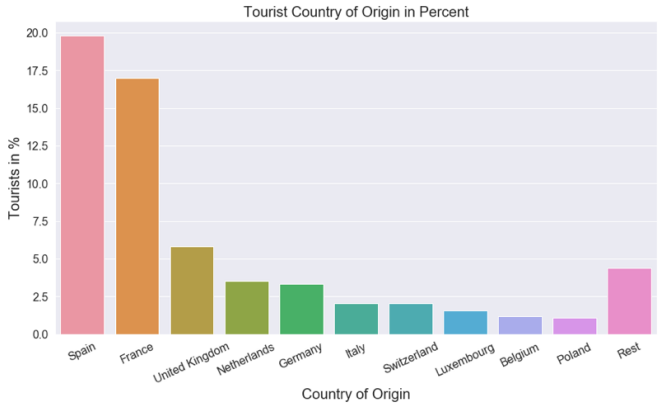
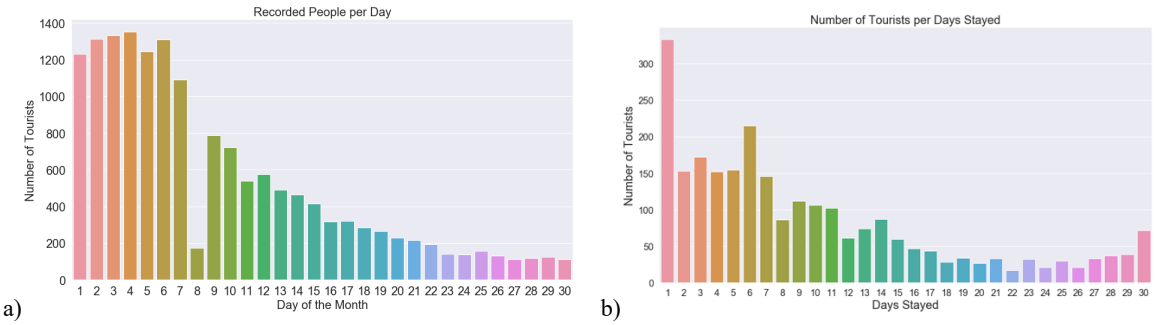


Figure 4: Tourist Country of Origin in Percent

Figure 4 displays the distribution of countries within the total population of event attendees in percent. It becomes evident that Spain and France represent the largest fraction of the tourists

with 19.7% and 17%, respectively. With a large difference, the UK, the Netherlands and Germany follow with 5.8%, 3.5%, 3.3%, respectively.



a) b)

Figure 5: a) Recorded Tourists per Day and b) Tourists per Days Stayed

The number of unique visitors recorded per day (see 5a) is fairly stable throughout the timespan of the music event (1st August – 6th August). From then on, a continuous decrease can be identified, which is expected, as only people who attended the event in the beginning of the month are considered. A very low number of people on the 8th can be seen, which is probably due to missing values in the dataset.

Figure 5b) shows, that the biggest fraction of tourists only appears in the dataset for a short period of time with the largest amount of people only appearing for a single day. This graph also explains the distribution displayed in Figure 5a) as all tourists appear at the beginning of the month and a decline of people with increasing total stay can be seen. A slight increase can be noticed towards at the end.

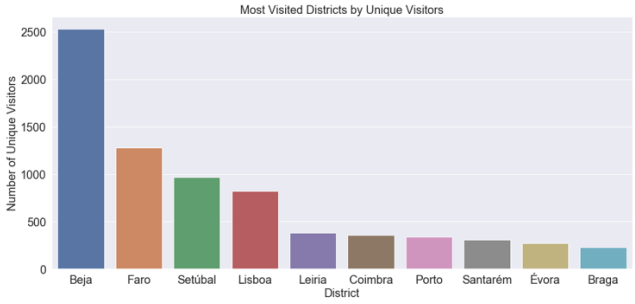


Figure 6: Districts visited by Unique Visitors

Figure 6 gives an overview about what other areas tourists have been in as well. Besides Beja, where the festival was held and therefore every single person in the dataset has been to, Faro, Setubal and Lisbon were the most visited districts.

4.2. Clustering

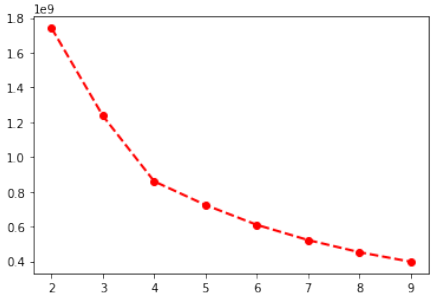


Figure 7: Visualization of elbow method used for clustering

Even though not as striking as preferable, a bent can be noticed at k=4. Therefore, this value was used for the segmentation of tourists within this analysis.

4.3. Data Analysis of different Segments

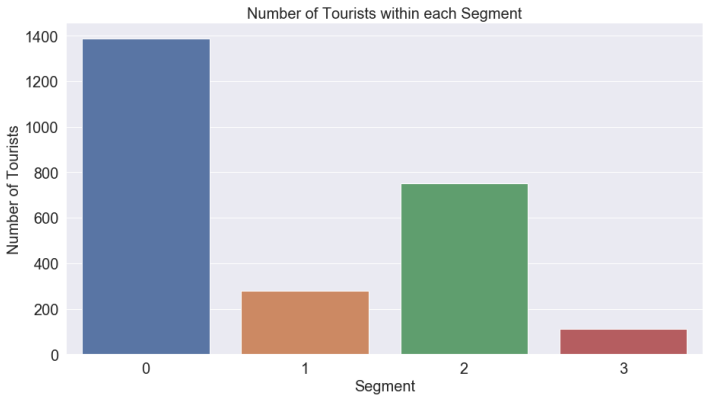


Figure 8: Number of Tourists within each Cluster

Segment 0 contains by far the most tourists with a fraction of 54.8% and 1387 tourists in total. Segment 2 follows with 29.6% and a total of 751 tourists. Segment 1 contains 279 people, which accounts for 11%. Segment 3 is the smallest segment with a total number of only 112 people and a share of 4.4%.

	Segment 0	Segment 1	Segment 2	Segment 3
total_stay	4.051911	24.985663	12.589880	22.366071
total_nights	3.051911	23.985663	11.589880	21.366071
days_before_event	0.658976	1.039427	1.532623	1.535714
days_after_event	1.801730	21.838710	9.392810	18.723214
days_at_event	1.591204	2.107527	1.664447	2.107143

Figure 9: Stay related feature averages per segment

Figure 9 displays the averages of the stay related features generated in each cluster. Segment 0 shows a rather short stay for tourists with an average of 4.05 days. Segment 2 shows an

average of 12.58 days and Segment 1 and 3 show long average stay periods with 24.98 and 22.36 days, respectively.

	Segment 0	Segment 1	Segment 2	Segment 3
unique_districts_visited	2.309301	3.903226	4.663116	8.553571
unique_concelhos_visited	3.523432	8.394265	10.135819	26.794643
unique_pois_visited	5.182408	13.899642	17.017310	45.258929
total_distance_travelled	179.140072	719.231980	678.316111	3098.870010
max_distance_from_start	112.479438	209.421800	271.405219	414.884588
daily_distance	46.188419	29.180023	61.490899	147.834577

Figure 10: Movement related feature averages per segment

The segments show an increase in unique places visited with each segment. People within segment 0 tend to only visit few unique places. Segment 1 visits approximately twice as many. Segment 2 shows somewhat similar numbers as segment 1, yet slightly higher. Segment 3 shows by far the highest numbers in terms of places visited with 8.55 districts visited and having been in close proximity to 45.25 different points of interest.

Looking at travel distances, segment 3 shows the highest numbers in all three categories. Segment 1 and 2 show similar total travel distances, but daily distance travelled for segment 2 is more than twice as high, which makes sense considering the overall time spent (see fig. 9). A moderate number of daily distance travelled, 46.18 kilometers can be noticed for segment 0. Also, with each segment the radius in which they moved seems to increase.

	Segment 0	Segment 1	Segment 2	Segment 3
days_low_budget_area	0.438356	0.903226	1.764314	2.714286
days_mid_budget_area	1.397260	11.161290	6.039947	14.437500
days_high_budget_area	1.879596	12.630824	4.026631	5.205357
economic_value_night_stay	214.407586	1479.950602	659.178128	1197.031984
economic_value_food	142.472963	979.928315	408.994674	734.107143
total_economic_value	356.880549	2459.878918	1068.172802	1931.139127

Figure 11: Economic value related features averages per segment

Looking at time spent within certain budget areas during the total stay, it becomes evident that throughout all segment the majority is spent in mid and high budget areas. For almost all segments an almost equal divide between time spent in mid and high budget areas can be noticed. Only segment 3 shows a majority of time spent in mid budget areas (14.44 days

compared to 5.2 days in high budget areas). The economic value is highly dependent on time of stay and areas stayed in. In general, an increase in total economic value can be noticed with the time stayed by each segment. Despite similar time spent by segment 1 and 3 (24.9 and 22.3 days, respectively) a significantly larger economic value can be noticed for segment 1, due to spending more time in high budget areas (12.6), compared to segment 3 (5.2).

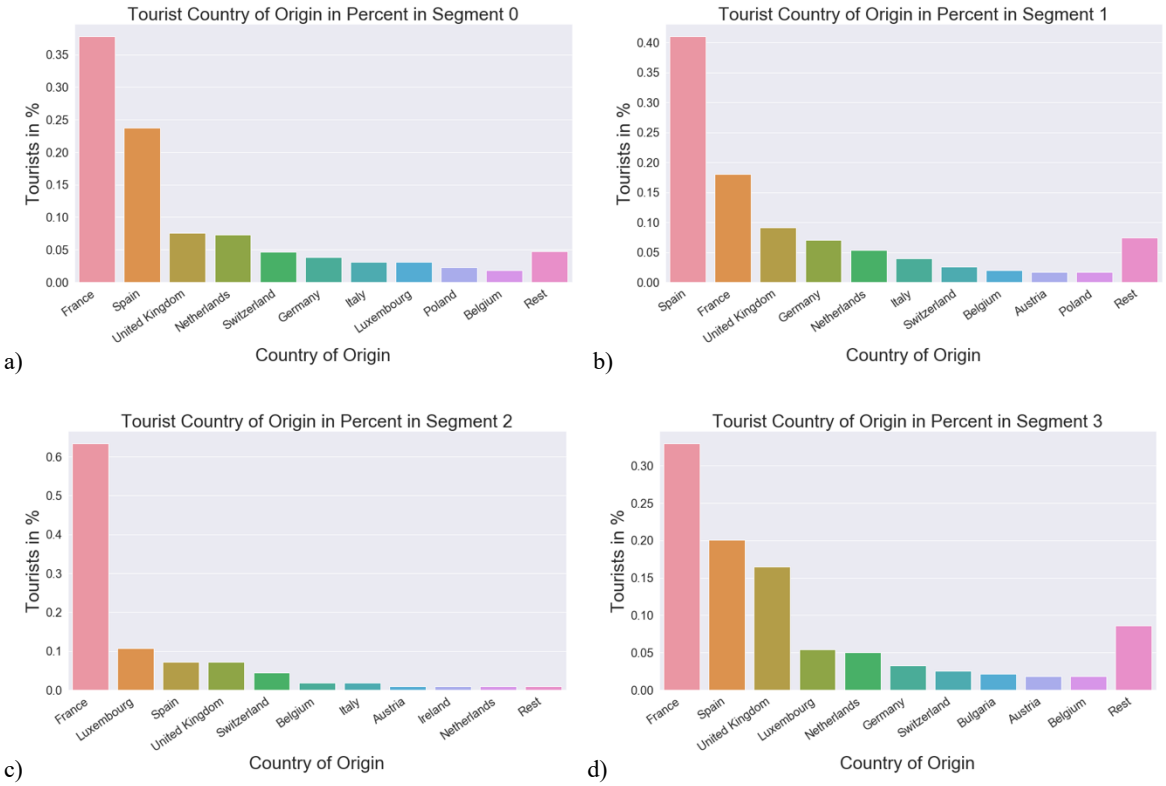


Figure 12: Tourist country of Origin within each segment.

As expected, considering figure 4, France and Spain are overall the most represented countries within the segments. What stands out is that segment 0 is by the biggest fraction represented by Spanish tourists (41%) followed by France with 18%. Segment 1 and 2 contain to the largest part of French tourists with 32.9% and 37.8%, respectively, followed by Spanish tourists with 20% and 23%, respectively. Furthermore, segment 1 shows a respectable number of British tourists with a fraction of 16.4%. Segment 3 consists mostly of French tourists (63.3%), followed by tourists from Luxembourg (10%). Apart from these numbers the segments are distributed fairly equally between other countries which due to the rather low number of tourists in the dataset in general only account for few people.

	Segment 0	Segment 1	Segment 2	Segment 3
Spain	0.701603	0.069051	0.219482	0.009864
France	0.358680	0.131994	0.407461	0.101865
United Kingdom	0.533613	0.193277	0.239496	0.033613
Netherlands	0.513889	0.097222	0.381944	0.006944
Germany	0.720588	0.066176	0.213235	NaN

Figure 13: Distribution of Segments within Nationality

Figure 13 takes a deeper look into how the top 5 represented nationalities are distributed throughout the different segments. Spanish tourists are mainly situated in segment 0 with 70% and another big chunk of 21.9% in segment 2. The distribution of French tourists is more even with the largest fraction in segment 2, followed by segment 0 with 40.7% and 35.8%, respectively. In general, most nationalities have their biggest fraction in segment 0 followed by segment 2. France is the only nationality with a fraction worth mentioning in segment 3 (10.1%).

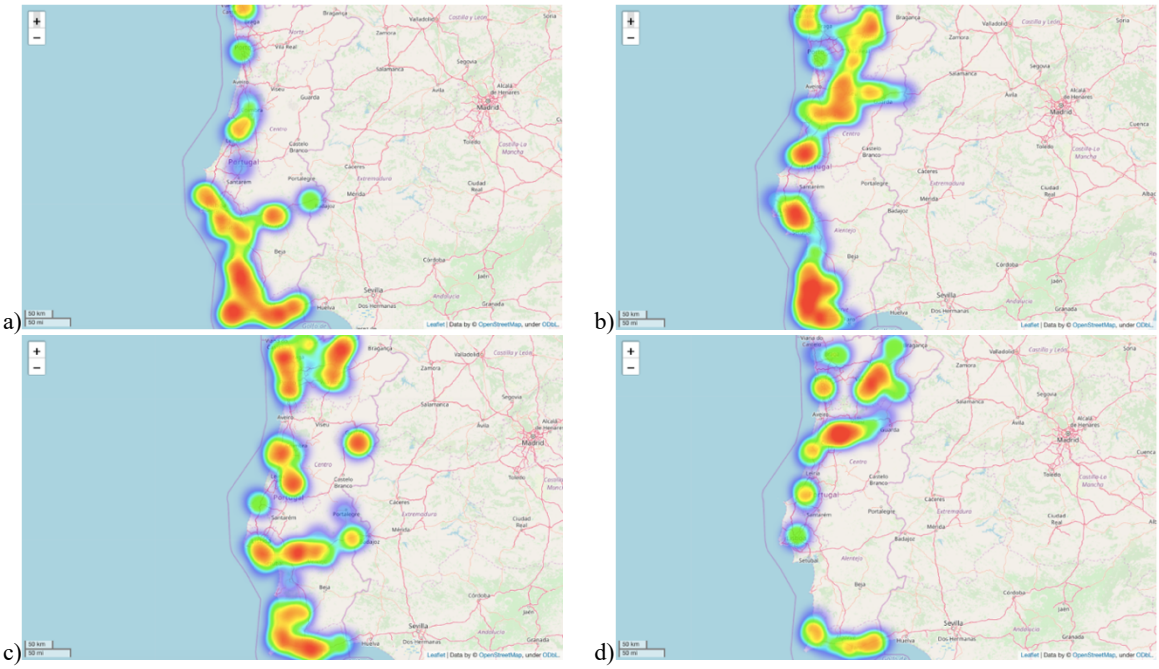


Figure 14: Heat Maps showing most documented areas by each segment.

The heatmaps presented above show where tourists went and according to color how many visits have occurred. This is measured in total visits, so reoccurring visits per tourist are considered. 14a shows that people from segment 0 tend to stay close to the festival area. Segment 0 has a larger spread and also visits more places up north. Similar tendencies can be seen in 14c, plus more activity in the center of Portugal. Segment 3, which supposedly

contains tourists with the most places visited shows a lot of activity in the north of Portugal, plus an even spread along the south. In general, people tend to stay close to the ocean and near major cities.

5. Discussion

5.1. Interpretation of Data

Based on the information obtained through data analysis, the clusters can be classified and the common type of tourist within the clusters can be defined. This can help event managers understand what kind of people attend their events and how they behave outside of the event visit. The following tourist types (personas) can be obtained.

5.1.1. Personas

Segment 0: Short term event tourists

Event visitors within segment 0 tend to only stay for a short amount of time with an average of roughly 4 days. Therefore, it could be concluded, that they come to Portugal for the sole reason of visiting the event. They arrive on the same day of the event or one day before. Apart from visiting the event, they don't do a lot of travelling and the only other districts they visit could be the ones through which they arrive or depart. They spend the majority of their time around the event area, which in this case is a high budget area.

Segment 1: Long term travelers looking for leisure and attractions

These tourists tend to stay in Portugal for long periods of time (24.9 days on average) and spend around 2 days at the event site. Despite their long time spent in Portugal they don't tend to visit too many places with around four districts and eight concelhos. With only 29 kilometers, their daily distance covered is the lowest between the four segments. They divide their time spent rather equally between mid and high budget areas. It looks like they tend to

stay close to the ocean areas and the main cities/attractions. It can be assumed that their main objective is indulgence and fun.

Segment 2: Medium term travelers looking to explore

These tourists tend to stay an average of 12 days. Considering their medium length stay, they spent time at a rather high number of places (4.66 districts and 10 concelhos). They also cover twice as much distance per day than segment 1 and their travel radius is also significantly larger (271km). They spent slightly more time in mid budget areas than in high budget areas. They visit places all throughout Portugal and also visit the more central areas. Their objective could be to discover and experience as much of Portugal as possible in the limited amount of time they are there.

Segment 3: Long term travelers who want to see it all

With an average of 22 days these tourists spend a lot of time in Portugal. They use their time by visiting a large number of places, by averaging 8.5 districts and 26 concelhos. With a daily travel distance of 147.83 kilometers and a travel radius of 414 kilometers, they cover a lot of ground. The majority of their time is spent in mid budget areas. Besides visiting the event in the south of Portugal they seem to spend the majority of time in the north of Portugal also considering the areas away from the coast.

5.2. Applying the information

Now that a lot of knowledge about the visiting tourists in general and the tourists within the clusters has been established, the question of how exactly this information can be utilized to make informed decisions in event management arises. This will be discussed in the following section.

5.2.1. Targeting and Promotion

One of the objectives of event managers to look into tourist data could be to gather insights on how and where to promote an already established event in foreign countries, in order to

use tourists as a new market for growth (Getz, Svensson, Peterssen, & Gunnervall, 2012). Simply looking at the nationalities of tourists travelling Portugal and those attending the event already can give first insights on which countries could be suitable to promote in. Either focusing on tourists from countries that already show interest in the event, in this case France and Spain (see figure 4) or focusing on countries that are underrepresented in event attendance, but show a general potential, for example the UK, the Netherlands and Germany. Besides only finding first indicators on where to promote a certain event, the insights gathered in this analysis can further assist in building a promotional strategy for each country. By clustering and building personas based on the obtained information an understanding of what kind of different tourist types there are has been established, giving first indicators on how they behave and what might interest them. This can be used to create specialized promotions fitted to certain tourist types. Furthermore, knowing the distribution of nationalities within clusters (see figure 12) and the distribution of tourist types within nationalities (see figure 13) gives indication on what kind of promotion to use in which country to attract to achieve the most promising results.

Two examples of how to promote an event in foreign countries could be (1) creating content for advertising purposes and (2) creating pre-planned travel packages, which, besides the festival tickets, could contain means of travel, accommodation or activities fitted to the preferences of certain tourist types.

Taking into account the Spanish market, looking at figure 13 it becomes evident, that the majority of event attendees (70,16%) belong within segment 0, meaning they only stay for a very short period of time and might only come to Portugal to visit the event. Considering the close proximity of Spain to Portugal, this seems plausible. So, for the sole purpose of attracting more potential visitors, approaching this segment seems reasonable. As for creating content for advertising purposes, this content should mainly focus on the event experience, as

these tourists are not interested in travelling much. Regarding offering specialized packages, these could simply include the tickets, shuttle buses, due to the close proximity, and accommodation possibilities on the event site.

When looking at the French market on the other hand, the majority of the tourists stay within segment 2 (40.74%), which is defined by medium term travelers who like to explore the country. Given the longer distance, compared to the Spanish market, this makes sense. Therefore, this segment should be approached in a different manner. Besides the event experience, the content created for this segment, could show the beautiful scenery of Portugal and its most iconic places, using the country as an additional motivator. Pre-planned travel packages could include flights, a rental car and hotel stays in popular areas or cities.

5.2.2. Area-Scouting

Another objective of event-managers could be to make out potential areas for where a new event could be held at. Looking at data from event attendees of existing events can reveal insights on where the target group likes to go as well. Looking at the heat maps in figure 14a, b and c, one can see that those segments that tend to travel throughout the country (segment 1,2 and 3) all spend time in the north of Portugal. Whether that could be due to interest in the area or because it is located on their homebound is not clear, yet in any way it shows benefits for these tourists. Therefore, potential event areas can be identified by studying the movement patterns of tourists throughout a country.

5.3. Limitations

Even though the work project succeeded in showcasing how data analysis can be used to inform event management, it was limited due to a number of factors.

The dataset only offered limited information in some regards. For one, the overall population within the dataset was rather small to begin with, resulting in a small number of potential event attendees. Furthermore, it showed low granularity regarding the number of entries per

day for some tourists, yet the already small number of attendees made disregarding these tourists questionable. Another factor that should be considered is the precision of the positional information, due to the way the mobile network is set up. Lastly, it only considered data gathered within the month of August, even though tourists could have arrived before or stayed until after. Again, disregarding those tourists was questionable, due to the low number of attendees. Also, the remaining trajectory of these tourists still contains useful information on their mobility and behavior patterns.

Furthermore, only one source of data was used. Adding other data sources, such as credit card data or data on hotel or Airbnb bookings could add another layer of information, which could result in more precise and informative results.

5.4. Future Implications

Future research could be conducted taking into account the limitations of this study and therefore adding to this field of study.

Firstly, data sources other than CRD mobile localization data could be used in order to guarantee higher data granularity.

Furthermore, adding more data sources could be used to obtain more informative and diverse results.

Similar studies could be conducted for different events and different countries to compare the effectiveness of the approach.

6. Conclusion

The objective of this work project was to showcase how big data analysis can be utilized to inform event management. By using mobile localization data, the mobility patterns and behavior of potential event attendees was analyzed and translated into new features describing the tourists. Through clustering, different tourist types were successfully

identified. The paper then showed how that information could potentially be applied to the decision making of event managers by giving examples. The paper successfully displayed how data analysis, especially that of mobility patterns, can be helpful in informing event management by conducting an empirical investigation and providing results.

Bibliography

Ahas, R., Aasa, A., Silm, S., & Tiru, M. (2007a). Mobile positioning data in tourism studies and monitoring: case study in Tartu, Estonia. In: M. Sigala, L. Mich, J. Murphy (Eds.), Information and communication technologies in tourism 2007, (pp. 119-128). Springer, Vienna.

Ahas, R., Armoogum, J., Esko, S., Ilves, M., Karus, E., Madre, J.L., . . . Tiru, M. (2015). Eurostat feasibility study on the use of mobile positioning data for tourism statistics, Report.

Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications, 105(9).

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2010). Real-time urban monitoring using cell phones: A case study in Rome. IEEE Transactions on Intelligent Transportation Systems, 12(1), 141-151. doi.org/10.1109/TITS.2010.2074196

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation research part C: emerging technologies, 68, 285-299.

Dwyer, L., Mellor, R., Mistilis, N., & Mules, T. (2000). A framework for evaluating and forecasting the impacts of special events. Proceedings of event evaluation, research and education, Sydney, 31-45.

Formica, S. (1998). The development of festivals and special events studies. Festival management and event tourism, 5(3), 131-137.

Getz, D. (1989). Special events: Defining the product. Tourism management, 10(2), 125-137.

Getz, D. (2000). Developing a research agenda for the event management field. Events beyond, 10-21.

Getz, D. (2008). Event tourism: Definition, evolution, and research. Tourism management, 29(3), 403-428.

Getz, D., & Page, S. (2016). Event studies: Theory, research and policy for planned events. Routledge.

Getz, D., Svensson, B., Peterssen, R., & Gunnervall, A. (2012). Hallmark events: definition and planning process. International Journal of Event Management Research, 7(1/2), 47-67.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. nature, 453(7196), 779. doi/org/10.1038/nature06958

Goldblatt, J. J. (1990). Special events: the art and science of celebration. Van Nostrand Reinhold.

Goldblatt, J., 2005. Special Events: Event Leadership for a New World. 4th edition. John Wiley & Sons: Hoboken.

Hall, C. M. (1992). Hallmark tourist events: impacts, management and planning. Belhaven Press.

Hodur, N. M., & Leistritz, F. L. (2007, October). Estimating the economic impact of event tourism: A review of issues and methods. In Journal of Convention & Event Tourism (Vol. 8, No. 4, pp. 63-79). Taylor & Francis Group.

Hu, H., Zhu, X., Hu, Z., Wu, J., & Zhang, X. (2018). Discovering Transportation Mode of Tourists Using Low-Sampling-Rate Trajectory of Cellular Data. 2018 5th International Conference on Systems and Informatics (ICSAI), (pp. 1120-1125). doi.org/10.1109/ICSAI.2018.8599469

Hui, P., Chaintreau, A., Scott, J., Gass, R., Crowcroft, J., & Diot, C. (2005, August). Pocket switched networks and human mobility in conference environments. In Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking (pp. 244-251). ACM.

Iovan, C., Olteanu-Raimond, A.-M., Couronne, T., & Smoreda, Z. (2013). Moving and calling: Mobile phone data quality measurements and spaciotemporal uncertainty in human mobility studies. In D. Vandenbroucke, B. Bucher, & J. Crompvoets, Gepgraphic information science at the heart of Europe (pp. 247-265). Cham: Springer. doi.org/10.1007/978-3-319-00615-4_14

Kim, M. K., Kim, S. K., Park, J. A., Carroll, M., Yu, J. G., & Na, K. (2017). Measuring the economic impacts of major sports events: the case of Formula One Grand Prix (F1). Asia pacific journal of tourism research, 22(1), 64-73.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. International Journal, 1(6), 90-95.

Lee, C. K., Mjelde, J. W., & Kwon, Y. J. (2017). Estimating the economic impact of a mega-event on host and neighbouring regions. Leisure Studies, 36(1), 138-152.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), 2579-2605.

Mamei, M., & Colonna, M. (2018). Analysis of tourist classification from cellular network data. Journal of Location Based Services, 12(1), 19-39. doi.org/10.1080/17489725.2018.1463466

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

Naboulsi, D., Fiore, M., Ribot, S., & Stanica, R. (2016). Large-scale Mobile Traffic Analysis: a Survey. IEE Communications Survey & Tutorials, 18(1), 124-161. doi.org/10.1109/COMST.2015.2491361

Oliver, V., & Enrique, G. (2014). Big Data y Turismo: Nuevos Indicadores para la Gestión Turística. Retrieved from http://www.rocasalvatella.com/sites/default/files/big_data_y_turismo-cast-interactivo.pdf

Brent Ritchie, J. R. (1984). Assessing the impact of hallmark events: Conceptual and research issues. Journal of travel research, 23(1), 2-11.

Smoreda, Z., Olteanu-Raimond, A.M., & Couronné, T. (2013). Spatiotemporal data from mobile phones for personal mobility assessment. Transport survey methods: best practice for decision making, 41, 745-767.

Song, L., Kotz, D., Jain, R., & He, X. (2004). Evaluating location predictions with extensive Wi-Fi mobility data. IEEE INFOCOM, 2, 1414-1424. doi.org/ 10.1145/965732.965747

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001, June). Constrained k-means clustering with background knowledge. In Icml (Vol. 1, pp. 577-584).

Zheng, W., Huang, X., & Li, Y. (2017). Understanding the touristic mobility using GPS: Where is the next place? Tourism Management, 59, 267-280. doi.org/10.1016/j.tourman.2016.08.009

Appendix

Appendix A

Figure 1: Different Event Classes by Getz (2008)

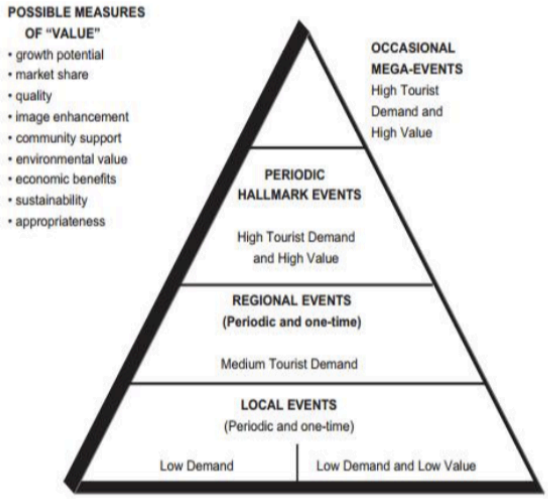


Figure 2: Different Event Types by Getz (2008)

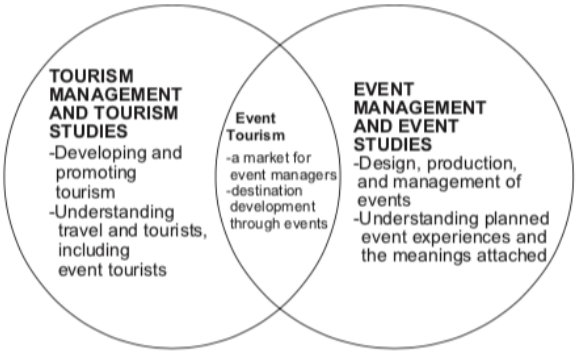


Figure 3: Intersection of Tourism and Event Management by Getz (2008)

Appendix B

Feature-Name	Formula
total_stay	$total\_stay = d_l - d_f$ d_l = Value of last day in Portugal d_f = Value of first day in Portugal
total_nights	$total\_nights = d_l - d_f + 1$ d_l = Value of last day in Portugal d_f = Value of first day in Portugal
days_before_event	$days\_before\_event = d_{ef} - d_f$ d_{ef} = Value of first day at Event d_f = Value of first day in Portugal
days_after_event	$days\_after\_event = d_l - d_{el}$ d_l = Value of last day in Portugal d_{el} = Value of last day at Event
days_at_event	$days\_at\_event = d_{el} - d_{ef}$ d_{el} = Value of last day at Event d_{ef} = Value of fist day at Event
mean_of_arrival	Assuming that the first entry for each tourist displays the place of arrival. Considering airplane, car/train and ship as mean of arrival. L_a = Location of arrival *Air*: Considering the coordinates of the only three airports in Portugal (Lisbon, Port and Faro) *Land*: Considering all districts on the border to Spain. *Water*: Considering the coordinates of all harbors active in touristic transport. If L_a is within the area of an airport: apply the value *air* If L_a is within a bordering district: apply the value *land* If L_a is within the area of a harbor: apply the value *water* If L_a is not within any of these areas: apply the value *unknown*
mean_of_departure	Assuming that the last entry for each tourist displays the place of departure. Considering airplane, car/train and ship as mean of departure. L_d = Location of arrival *Air*: Considering the coordinates of the only three airports in Portugal (Lisbon, Port and Faro) *Land*: Considering all districts on the border to Spain. *Water*: Considering the coordinates of all harbors active in touristic transport. If L_d is within the area of an airport: apply the value *air* If L_d is within a bordering district: apply the value *land* If L_d is within the area of a harbor: apply the value *water* If L_d is not within any of these areas: apply the value *unknown*
unique_districts_visited	Count the number of unique values within the *distritos* column.
unique_concelhos_visited	Count the number of unique values within the *concelho* column.

unique_pois_visited	Count the number of unique values within the *poi* column.
total_distance_travelled	Summing the Geodesic distance of all consecutive locations: $$\text{total\_distance\_travelled} = \sum_{k=0}^{n} 2 \times R \times \arctan\left(\sqrt{a_k}, \sqrt{1} - a_k\right)$$ $$a_k = hav(\Delta\varphi + \cos(\varphi_k) \times \cos(\varphi_{k+1}) \times hav(\Delta\lambda)$$ $$hav(\theta) = sin^2(\frac{\theta}{2})$$ *k* = index starting value of entries *n* = number of entries ϕ = latitude value λ = longitude value *R* = radius of the earth in kilometers (6371km)
max_distance_from_start	Calculating the distance between the location of the first entry and every other entry and defining the highest value as *max_distance_from_start*.
daily_distance	$$\text{daily\_distance} = \frac{total\_distance\_travelled}{total\_stay}$$
'economic_value_night_stay'	Considering the last entry of each day as overnight location. Summing up individual values of all locations stayed at over night. $$economic\_value\_night\_stay = \sum_{k=0}^{n} P_C$$ *k* = index starting value of entries *n* = number of nights P_C = price of average hotel stay in certain concelho *C* *C* = certain concelho at index between *k* and *n* (last entry of the day)
'days_low_budget_area', 'days_mid_budget_area', 'days_high_budget_area'	Considering three budget areas (low, mid and high): *low budget area* < 40eur/night 41eur/night < *mid_budget_area* < 65eur/night high budget area 64eur/night < *high_budget_area* For each day, taking the median value of concelhos visited. Comparing this value with the different budget areas and placing the respected concelho in on of these bins. Counting the days spend at each.
'economic_value_food'	$$economic\_value\_food = \mathrm{n}_l \times P_l + \mathrm{n}_m \times P_m + \mathrm{n}_h \times P_h$$ *n* = number of days spent in certain area *P* = average price of food expenditure for certain area *l* = low budget area *m* = mid budget area *h* = high budget area
'total_economic_value'	$$total\_economic\_value = economic\_value\_night\_stay + economic\_value\_food$$

Figure 1: Detailed information on how features were generated

Appendix C

Figure 1: t-SNE of Clusters Generated

Using the t-SNE method, the clusters can be visualized in a two-dimensional graph, despite their multidimensionality. Yet, it has to be considered, that the distances between points are not in proportion and therefore are not very meaningful.