Data in brief 28 (2020) 104906

Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib

Data Article

FISETIO: A FIne-grained, Structured and Enriched Tourism Dataset for Indoor and Outdoor attractions

Amir Khatibi*, Ana Paula Couto da Silva, Jussara M. Almeida, Marcos A. Gonçalves

Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

ARTICLE INFO

Article history:
Received 20 June 2019
Received in revised form 21 November 2019
Accepted 21 November 2019
Available online 2 December 2019

Keywords:
Tourism prediction baselines
Indoor attractions
U.K national museums and galleries
Outdoor attractions
U.S national parks
Trip advisor social media
Climate features

ABSTRACT

This paper aims to introduce our publicly available datasets in the area of tourism demand prediction for future experiments and comparisons. Most of the previous works in the area of tourism demand forecasting are based on coarse-grained analysis (level of countries or regions) and there are very few works and consequently datasets available for fine-grained tourism analysis (level of attractions and points of interest). In this article, we present our fine-grained enriched datasets for two types of attractions – (I) indoor attractions (27 Museums and Galleries in U.K.) and (II) outdoor attractions (76 U.S. National Parks) enriched with official number of visits, social media reviews and environmental data for each of them. In addition, the complete analysis of prediction results, methodology and exploited models, features' performance analysis, anomalies, etc, are available in our original paper, "Fine-grained tourism prediction: Impact of social and environmental features"[2].

DOI of original article: https://doi.org/10.1016/j.ipm.2019.102057.
* Corresponding author. Department of Computer Science, Universidade Federal de Minas Gerais, Avenida Presidente Antônio Carlos 6627, Pampulha, Belo Horizonte, MG, 31270-901, Brazil.
E-mail addresses: amirkm@dcc.ufmg.br, amir.khatibi.m@gmail.com (A. Khatibi).

https://doi.org/10.1016/j.dib.2019.104906

Specifications Table

Subject area	Computer Science Applications - Tourism
More specific subject area	Social Media Data Analysis – Machine Learning Applications On Predicting Tourism Demand
Type of data	Tables Figures Raw and pre-processes data CSV files
How data were acquired	Official visitation data collected from governmental websites: • U.S. National Park Service website, https://irma.nps.gov/Stats/ • Official monthly total numbers of visitors of museums and galleries in the United Kingdom, https://www.gov.uk/government/statistical-data-sets/museums-and-galleries-monthly-visits Climate Data has been collected from governmental websites: • U.S. National Climate Data Center, https://www.ncdc.noaa.gov/cag/time-series/us/ • United Kingdom's national weather service (Met Office), https://www.metoffice.gov.uk/ Social Media Data is crawled from the biggest travel listing website • TripAdvisor, https://www.tripadvisor.com/
Data format	Raw Pre-processed, structured, crossed-over and cleaned dataset Filtered into two categories of indoor (Museum and Galleries in U.K) and outdoor (National Parks in U.S) attractions
Parameters for data collection	All available data after the year 2000
Description of data collection	For official visitation and climate data, we automated downloading using the selenium tools from the official corresponding websites using the collection parameters We collected social media data, crawling the graph of TripAdvisor pages, starting from the page of U.S. national parks. We obtained the reviews and ratings for those U.S. national parks with an available travel contents page
Experimental features	Official data: monthly number of visitors for each attraction Social Media features: monthly number of reviews, average rating Environmental features: monthly minimum, average and maximum temperatures (in Celsius degree), precipitation and rainfall, sunny hours and days of air frost
Data source location	27 Museum and Galleries in U.K 76 National Parks in U.S
Data accessibility	Data is included with this article Also publicly available at [1]: • Repository name: Mendeley • https://doi.org/10.17632/t7bfhtzhxg.1 • Direct URL to data: https://data.mendeley.com/datasets/t7bfhtzhxg/1
Related research article [2]	Khatibi, Amir; Belém, Fabiano; Couto da Silva, Ana Paula; Almeida, Jussara; Gonçalves, Marcos Andre. (2019), "Fine-grained tourism prediction: Impact of social and environmental features", Information Processing & Management, 102057. DOI: https://doi.org/10.1016/j.ipm.2019.102057

Open source IDE executing record and playback test for the web https://www.seleniumhq.org/selenium-ide/.

Value of the Data

- Low granularity data at the level of attractions
- Inclusion of a large range of attractions (103 attractions) in two categories: indoors (museums and galleries) and outdoors (parks).
- Inclusion of official data as ground-truth.
- Enriched dataset with social media and environmental data, all crossed-over for each type of attraction, serving as a publicly available dataset to be used in the development of further experiments in the area of tourism demand prediction.
- Possibility of studying seasonality, data recency and performance of prediction models in different types of attractions.
- Useful for researchers in areas of Computer Science Applications and Tourism Analysis.

1. Data description

We collected two datasets of touristic attractions split into two categories: (1) outdoor and (2) indoors.

1.1. Outdoor dataset

The outdoor dataset consists of climate, social media and official data for 76 national parks in the United States. For each park, the values of the variables (features) are reported aggregated by month. Each data record includes the number of social media reviews, average rating, average minimum temperature, average temperature, average maximum temperature, average precipitation, and the official number of visits. The period of data spans from January 2011 until September 2016.

1.2. Indoor dataset

The indoor dataset contains climate, social media and official data for 27 museums and galleries in the United Kingdom, covering monthly aggregated values in the period of August 2001 up to August 2018. Each record of the dataset includes the official number of visits, number of social media reviews, average rating, average minimum temperature, average maximum temperature, number of sunny hours, number of frost days and finally average rainfall.

In addition to the above cleaned datasets, we have also included the raw collected data in the folder named "pre-data_crossing". This folder contains all the above mentioned data before data aggregation and dataset integration. One can find official data, social media data and climate data each in a separate folder – with the corresponding names.

1.3. Dataset feature distributions

In this part, we give a general view of the distributions of feature values. We show the Complementary Cumulative Distribution Function (CCDF) of each feature for both datasets in Figs. 1 and 2. The y-axis shows the probability of the feature value *exceeding* the x-axis ($P(X > x)$).

In Fig. 1, we have the CCDF plots for total number of reviews and visits in plots (a) and (b); mean average temperature and mean temperature difference (difference between minimum and maximum temperature in Celsius) in (c) and (d) and mean ratings and mean average precipitation in (e) and (f). Each point represents an individual national park in the outdoor dataset. For instance, in Fig. 1(c) we can see that for about 70% of the parks, the mean average temperature is over 10, whereas almost all parks have the average temperature higher than 5 Celsius degree.

Similarly, Fig. 2 presents CCDF plots for the total number of reviews, total number of visits, mean average temperature (in Celsius), mean number of sunny hours, mean rating and mean raining (in mm) in plots (a), (b), (c), (d), (e) and (f), respectively. As before, each point represents one individual museum/gallery. For instance in Fig. 2(a), the CCDF-plot shows that for only 10% of the museums the total number of reviews was over 2 thousands.

1.4. Dataset feature statistics

In this part, firstly, we present the basic statistics for each of the variables in the dataset. Next, we show the pearson correlations within variables in each dataset.

In the following, Tables 1 and 2 present basic statistics for each category of attractions and corresponding variables/features.

Tables 3 (Parks) and 4 (Museums) provide the pearson correlation analysis [3] of different features extracted from the environmental, social media and official datasets. Although some results are not that surprising – for instance, the positive correlations between temperature and number of visitations for park attractions and the positive correlations between sunny hours and minimum temperature for museum attractions – we have some interesting results such as the high correlations between the number of visits with #Reviews on both types of attractions. This simple analysis indicates a new

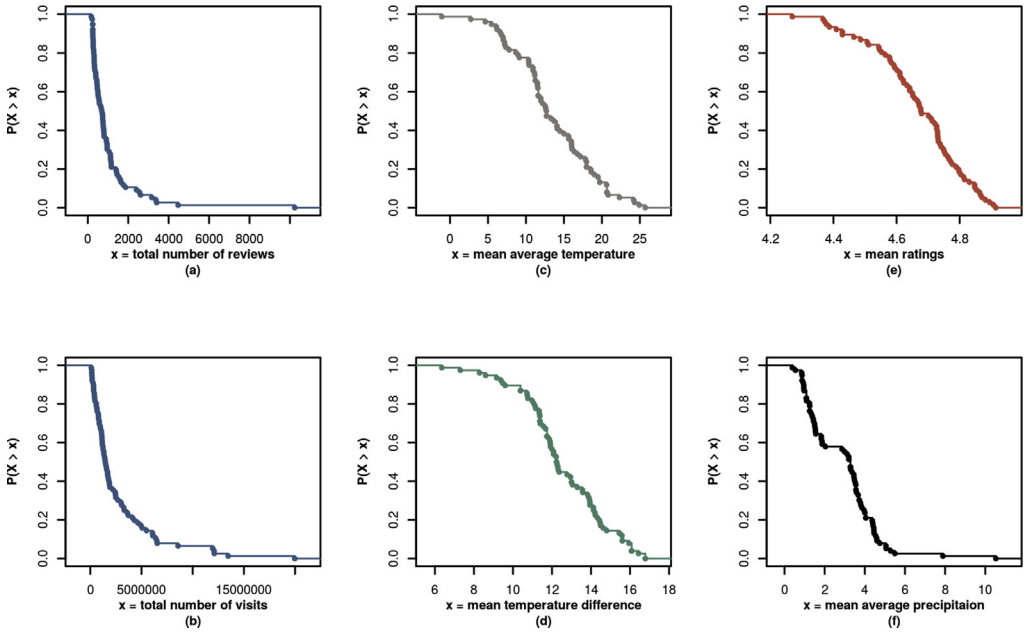

Fig. 1. Complementary Cumulative Distribution Function (CCDF) plots for features in the outdoor dataset.

interesting perspective for predicting the visitation counts at specific touristic places: both social media data and environmental features should be considered to create more accurate tourism prediction models.

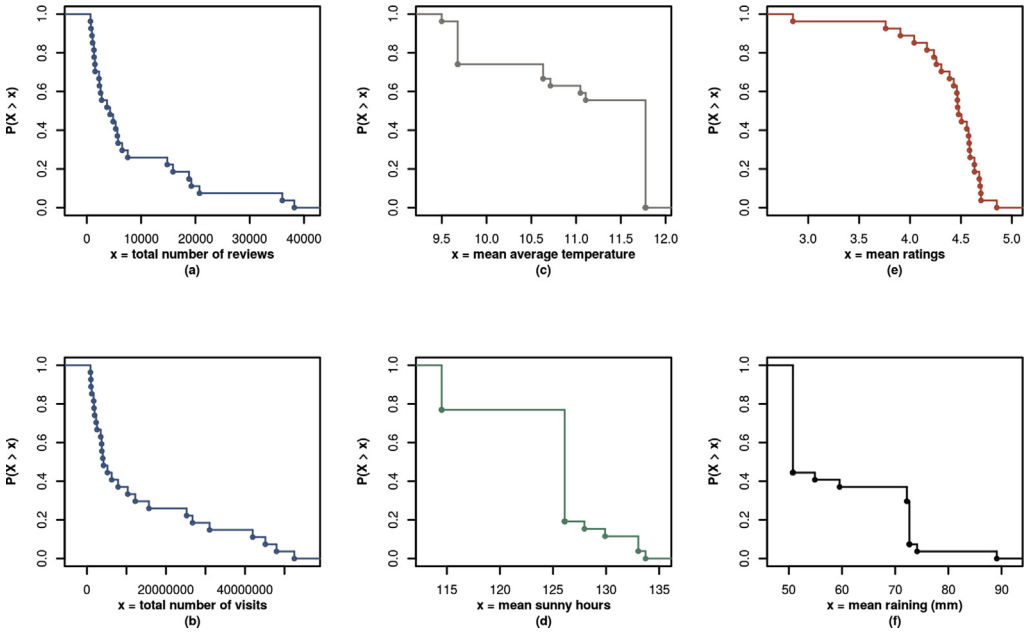

Fig. 2. Complementary Cumulative Distribution Function (CCDF) plots for features in the indoor dataset.

Table 1
Basic statistics for 76 National Parks in U.S.

	Min.	1stQuart.	Median	Mean	3rdQuart.	Max.	Skewness	Std. Dev
#visits	0	11970	36710	82900	96030	1001000	3.16	127202.9
Avg temp.	−18.56	7001	14.45	13.73	21.67	31.61	−0.41	9.54
Max temp.	−14.5	13.28	21.22	20.03	28.17	39	−0.53	9.84
Min temp.	−22.61	0.7223	7945	7427	14.95	26.67	−0.24	9.43
Precipit.	0	1.01	2.34	2931	4.07	25.03	2.33	2.67
#Reviews	1	7	15	30.38	34	615	5.54	50.51
Avg. rating	3	4533	4714	4669	4848	5	−1.42	0.26

Table 2
Basic statistics for 27 Museums and Galleries in U.K.

	Min.	1stQuart.	Median	Mean	3rdQuart.	Max.	Skewness	Std. Dev
#visits	0	18470	42440	131300	200300	811200	1.44	164726.3
Max temp.	2.3	10	14.8	14.75	19.4	27	0.08	5.47
Min temp.	−3.7	3.7	7.5	7348	11	15.2	0.02	4.27
#Air frost days	0	0	0	2.52	4	27	2.06	4.33
Rainfall	0.4	32.8	52.8	59.98	82.25	254.2	1.03	37.43
Sunny hours	18.5	68.5	122.1	124.5	173.4	311.4	0.25	61.21
#Reviews	1	8	27	84.5	92.75	1114	3.16	142.24
Avg. rating	1	4.25	4.56	4385	4.69	5	−2.52	0.56

2. Experimental design, materials and methods

We used five different sources for our data collection, namely (1) the U.S. National Park Service, (2) TripAdvisor web page, (3) U.S. national climate data center, (4) the Department for Digital, Culture, Media and Sport of England, and (5) the U.K. national weather service (Met Office). After data collection, we gathered, cleaned and merged all data into two categories of attractions, namely, (1) outdoors and (2) indoors. In the following we elaborate on the data sources for each of the attraction types.

2.1. Outdoor dataset

We accessed the U.S. National Park Service website (https://irma.nps.gov/Stats/) to download the monthly total number of visitors for each national park in the period of January 1996 to February 2016. We consider this dataset as ground truth for possible tourism analysis.

We collected social media data from TripAdvisor - the largest travel website with more than 570 million reviews and 455 million average monthly unique visitors (http://www.tripadvisor.com/). Specifically, we conducted a crawling on the graph of TripAdvisor pages, starting from the page of U. S. national parks. We obtained the reviews and ratings for those U.S. national parks with a travel contents page and then aggregated the results in a monthly fashion to make it comparable with the

Table 3
Pearson correlation results for 76 National Parks in U.S.

corr.	#visits	Avg temp.	Max temp.	Min temp.	Precipit.	#Reviews	Avg rating
#visits	1						
Avg temp.	0.249	1					
Max temp.	0.24	0.99	1				
Min temp.	0.252	0.989	0.959	1			
Precipit.	−0.002	0.154	0.074	0.235	1		
#Reviews	0.329	0.202	0.188	0.213	0.064	1	
Avg. rating	−0.07	−0.114	−0.124	−0.102	0.003	0.006	1

Table 4
Pearson correlation results for 27 Museum and Galleries in U.K.

corr.	#visits	Max temp.	Min temp.	#Air frost days	Rainfall	Sunny hours	#Reviews	Avg rating
#visits	1							
Max temp.	0.145	1						
Min temp.	0.121	0.958	1					
#Air frost days	−0.087	−0.716	−0.737	1				
Rainfall	−0.13	−0.205	−0.079	0.023	1			
Sunny hours	0.082	0.77	0.637	−0.536	−0.373	1		
#Reviews	0.445	0.13	0.113	−0.083	−0.077	0.055	1	
Avg. rating	0.042	0.037	−0.038	0.062	−0.059	−0.036	0.196	1

ground-truth dataset. For each national park, the monthly aggregated number of reviews alongside the average rating scores of reviewers were collected for the period of January 2011 until September 2016.

Climate (environmental) data was collected from the U.S. National Climate Data Center (https://www.ncdc.noaa.gov/cag/time-series/us/). To that end, we built a specific web crawler, since the climate data is aggregated for each climate division in the U.S. states and regions in a different url. For each U.S. national park, we used the climate data associated with the closest climate division based on the Earth curvature distance between target points. We collected the monthly minimum, maximum and average temperatures as well as the monthly precipitation. Our climate data covers the period of January 2000 to November 2016.

We initially selected 124 national parks in the U.S. with available social media data, environmental data and monthly official visitation in our datasets. In a further step, we filtered out parks with very few reviews in TripAdvisor. Indeed, we discarded all parks with fewer than 200 reviews in the last 3 years (i.e. less than 5 reviews per month, on average). After the filtering process, we retained 76 national parks.

2.2. Indoor dataset

We downloaded the official monthly total number of visitors of museums and galleries in the United Kingdom from April 2004 to July 2018 by accessing the following url: (https://www.gov.uk/government/statistical-data-sets/museums-and-galleries-monthly-visits).

Likewise the outdoor dataset, we collected users reviews and ratings in the period of August 2001 to August 2018 from TripAdvisor for museums and galleries with an available travel content page. Next, we aggregated the results in a monthly manner in order to convert the data to the same granularity of the ground truth dataset, i.e., the dataset of official visits.

United Kingdom's national weather service (Met office at https://www.metoffice.gov.uk/) was the data source for gathering climate (environmental) data. It provides climate data for 37 climate divisions in U.K.. For each gallery or museum, we collected the climate data of the closest climate station considering the earth curvature distance. Specifically, we gathered the monthly average temperature, monthly number of air frost days,[2] monthly sunshine duration and monthly rainfall. The climate data covers the period of January 1980 to August 2018.

After crossing the three aforementioned datasets, we end up with 38 museums and galleries in England with social media data, environmental data and monthly official visitation census available. In an additional step, as performed for the outdoor dataset, we discarded museums and galleries with very few reviews in TripAdvisor. Specifically, we filtered out all attractions with fewer than 250 reviews in the last 5 years (i.e. less than an average of 5 reviews per month). After the data cleaning process, we retained 27 museums and galleries.

[2] An air frost occurs when the air temperature falls to or below the freezing point of water; it is usually defined as the air temperature being below the freezing point of water at a height of at least one meter above the ground.

Acknowledgements

This work is partially supported by CNPq (169823/2017-2), CAPES (422593/2018-4), and Fapemig (00543-17).

Conflict of Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: This work is partially supported by CNPq, CAPES, and Fapemig and the the grant numbers has been provided previously as following in the order 169823/2017-2, 422593/2018-4 and 00543-17.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2019.104906.

References

[1] Amir Khatibi, Couto da Silva, Ana Paula, Jussara Almeida, Marcos Andre Gonçalves, FISETIO: A FIne-Grained, Structured and Enriched Tourism Dataset for Indoor and Outdoor Attractions", Mendeley Data, V1 (Experiment Data Files), 2019, https://doi.org/10.17632/t7bfhtzhxg.1.

[2] Amir Khatibi, Fabiano Belém, Couto da Silva, Ana Paula, Jussara Almeida, Marcos Andre Gonçalves, Fine-grained Tourism Prediction: Impact of Social and Environmental Features, Information Processing & Management, 2019, p. 102057, https://doi.org/10.1016/j.ipm.2019.102057.

[3] Jacob Benesty, Jingdong Chen, Yiteng Huang, Israel Cohen, Pearson correlation coefficient, in: Noise Reduction in Speech Processing, Springer, Berlin, Heidelberg, 2009, pp. 1–4.