Data article

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

**Authors:** Amir Khatibi, Ana Paula Couto da Silva, Jussara M. Almeida, Marcos André Gonçalves

**Affiliations:** Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

**Contact email:** amirkm@dcc.ufmg.br

**Abstract**

This paper introduces our publicly available datasets in the area of tourism demand prediction for future experiments and comparisons. Most 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 datasets available for fine-grained tourism analysis as well (level of attractions and points of interest). In this article, we present our fine-grained 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 [1].

**Specifications Table**

|  |  |
| --- | --- |
| Subject area | Tourism – Computer Science |
| More specific subject area | Social Media Data Analysis – Machine Learning Applications On Predicting Tourism Demand |
| Type of data | CSV files |
| How data was 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 https://www.tripadvisor.com/ |
| Data format | Structured, crossed-over, cleaned dataset filtered into two categories of indoor (Museum and Galleries in U.K) and outdoor (Nation Parks in U.S) attractions. |
| Experimental factors | For each attraction, data features are aggregated monthly |
| Experimental features | Official data: monthly number of visitors for each attraction  Social Media features: monthly number of reviews, average rating  Environmental features: monthly min, average and max temperatures (in Celsius degree), precipitation and rainfall, sunny hours and days of air frost. |
| Data source location | 27 Museum and Galleries in U.K in indoor dataset.  76 National Parks in U.S in outdoor dataset. |
| Data accessibility | All the data is available with this article. |

**Value of the data**

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

**Data**

We have two datasets with collected information for several indoor and outdoor attractions.

(i) In the outdoor dataset, we have 76 national parks in United States. For each park the history of monthly aggregated values is reported. Each data record includes the number of social media reviews and average rating of month; monthly average, minimum and maximum temperatures and precipitation; and the official reported monthly number of visits as well. The historic data spans from January 2011 until September 2016.

(ii) For the indoor dataset we have the same data structure 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 monthly number of visits; the number of social media reviews and average rating of month; monthly average, minimum and maximum temperatures, number of sunny hours and days of frost in the month; and the average rainfall.

In addition to the cleaned datasets, we have also included the raw collected data in the folder named “pre-data crossing” containing all the above mentioned data before data grouping and datasets integration. One can find official data, social media data and climate data each in separated folders -- with the corresponding name, respectively.

**Experimental Design, Materials and Methods**

We collected datasets from five different sources, namely: (1) the U.S. National Park Service; (2) *TripAdvisor*; (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). We gathered, cleaned and merged all data into two categories of attractions, namely, (I) outdoors and (II) indoors.

(I) Outdoor Dataset: We downloaded from the U.S. National Park Service website (https://irma.nps.gov/Stats/), the monthly total number of visitors for various national parks from January 1996 to February 2016. These numbers could serve as ground truth for prediction tasks.

We also collected social media data from *TripAdvisor* - the largest travel website with more than 570 million reviews and 455 million monthly average unique visitors (http://www.tripadvisor.com/). Specifically, we conducted a crawling on the graph of *TripAdvisor* pages, starting from the pages of the same 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 monthly to make it comparable with our ground-truth dataset. At the end, we gathered monthly number of reviews along with the average rating scores of reviewers for the period of January 2011 till September 2016 over a number of parks.

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 climate divisions in the U.S. states and regions. 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 from January 2000 to November 2016.

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

(II) Indoor Dataset: We downloaded the official monthly total numbers of visitors for a number of museums and galleries in the United Kingdom, from April 2004 to July 2018 (https://www.gov.uk/government/statistical-data-sets/museums-and-galleries-monthly-visits).

Likewise the outdoor attractions, we collected from *TripAdvisor* the reviews and ratings for those museums and galleries with a travel contents page and aggregated the results monthly to make it comparable with our ground truth dataset, i.e., the dataset of official visits. At the end, we achieved monthly number of reviews along with the average rating scores of reviewers during the period of August 2001 till August 2018 for a number of museums and galleries.

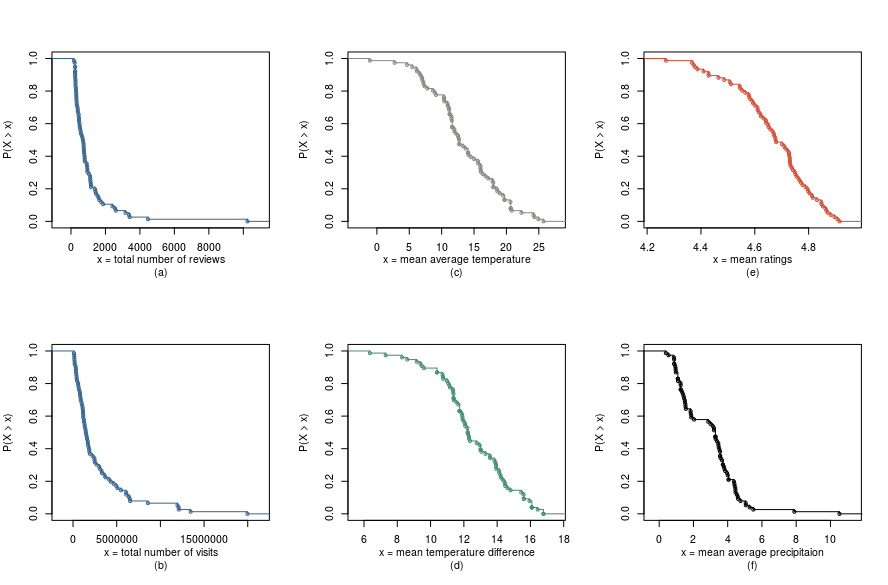
United Kingdom's national weather service (https://www.metoffice.gov.uk/) was our source for gathering climate (environmental) data. They basically provide climate data in 37 climate divisions in U.K. In our collection, for each gallery or museum, we used the climate data associated with the closest climate station considering the earth curvature distance between them. Specifically, we collected the monthly average temperatures, days of air frost[[1]](#footnote-2), total sunshine duration besides the total monthly rainfall. Our climate data covers the period of January 1880 to August 2018.

We initially selected 38 museums and galleries in England with social media data, environmental data and monthly official visitation census available in our datasets. In a further step, like 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.

In order to give a general view of the features distribution along the attractions, we have plotted the Complementary Cumulative Distribution Function (CCDF) plots for each of features for our two datasets outdoors and indoors (Figures 1 and 2, respectively). The axis-X shows the total/average of a feature for different attractions while axis-Y represents the probability of that feature having a value greater than the values in axis-X in the dataset.

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

Similarly, Figure 2 presents CCDF plots for the total number reviews, total number of visits, mean average temperature (in Celsius), mean number of sunny hours, mean rating and mean raining in parts (a), (b), (c), (d), (e) and (f), respectively. As before, each point represents one individual museum/gallery. In Figure 2(b), for instance, the CCDF-plot shows that for only 20% of the museums the total number of visits was over 20 millions.

Figure 1 – Complementary Cumulative Distribution Function (CCDF) plots of features in outdoor dataset

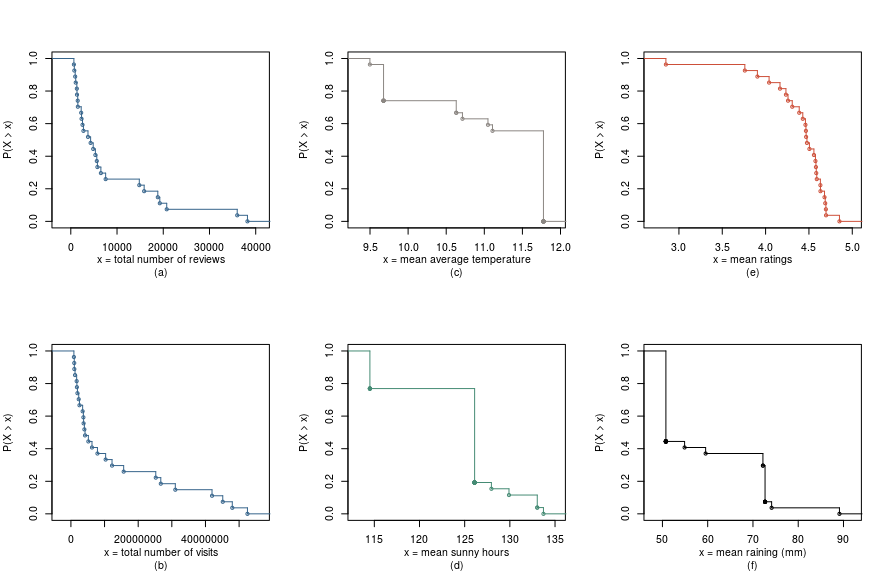


Figure 2 – Complementary Cumulative Distribution Function (CCDF) plots of features in indoor dataset

**Acknowledgements**

This work is partially supported by CNPq, CAPES, and Fapemig.

**References**

[1] Amir Khatibi, Fabiano Belem, Ana Paula Couto da Silva, Jussara M. Almeida, Marcos A. Gonçalves, Fine-Grained Tourism Prediction: Impact of Social and Environmental Features, Information Processing and Management, in Press

1. . 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 freezing point of water at a height of at least one meter above the ground. [↑](#footnote-ref-2)