UNIVERSIDADE FEDERAL DE MINAS GERAIS

Departamento de Ciência da Computação

Programa de Pós-graduação em Ciência da Computação

Amir Hassan Khatibi Moghadam

FINE-GRAINED TOURISM DEMAND PREDICTION: CHALLENGES AND NOVEL SOLUTIONS

Belo Horizonte

2021

AMIR KHATIBI

PREVISÃO DE DEMANDA DE TURISMO EM GRÃO-FINO: DESAFIOS E NOVAS SOLUÇÕES

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

ORIENTADOR: MARCOS ANDRÉ GONÇALVES
COORIENTADORA: ANA PAULA COUTO DA SILVA

Belo Horizonte - MG
Maio de 2021

AMIR KHATIBI

FINE-GRAINED TOURISM DEMAND PREDICTION: CHALLENGES AND NOVEL SOLUTIONS

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

ADVISOR: MARCOS ANDRÉ GONÇALVES
CO-ADVISOR: ANA PAULA COUTO DA SILVA

Belo Horizonte - MG
May 2021

Khatibi, Amir.

K45f Fine-Grained tourism demand prediction [manuscrito]: challenges and novel solutions / Amir Khatibi – 2021.
xxi, 104 f. il.

Orientador: Marcos André Gonçalves.
Coorientadora: Ana Paula Couto da Silva
Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f.98-104.

1. Computação – Teses. 2. Análise de redes sociais – Teses. 3. Turismo - Predicão – Teses. 4. Redes sociais on-line.– Teses. I. Gonçalves, Marcos André. II. Silva, Ana Paula Couto da. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. III.Título.

CDU 519.6*04(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6ª Região nº 1510

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Fine-Grained Tourism Demand Prediction: Challenges and Novel Solutions

AMIR HASSAN KHATIBI MOGHADAM

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

Marcos André Gonçalves

Prof. Marcos André Gonçalves - Orientador
Departamento de Ciência da Computação - UFMG

Ana Paula Couto da Silva

Profa. Ana Paula Couto da Silva - Coorientadora
Departamento de Ciência da Computação - UFMG

Flávio Vinicius Diniz de Figueiredo

Prof. Flávio Vinícius Diniz de Figueiredo
Departamento de Ciência da Computação - UFMG

Prof. Pedro Olmo Stancioli Vaz de Melo
Departamento de Ciência da Computação - UFMG

Daniel Sadoc Menasche

Prof. Daniel Sadoc Menasche
Departamento de Ciência da Computação - UFRJ

Ricardo Torres

Prof. Ricardo da Silva Torres
Department of ICT and Natural Sciences - NTNU

Belo Horizonte, 28 de Maio de 2021.

To my patient, beautiful wife, Bruna,
my lovely, smiling son, Navid,
and my dear family
who always inspire me to be better

Acknowledgments

I would like to express my special appreciation to my outstanding supervisors Prof. Marcos Andre Gonçalves and Prof. Ana Paula Couto da Silva, not only for their great supervising during my PhD studies but also for their support during my stay here at Brazil.

I also want to give my special thanks to Prof. Jussara Marques Almeida for her outstanding contribution to my work.

I wish to thank my parents and my great brother and lovely sister for all their goodness and kindness they always have with me.

Finally, I acknowledge Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for their financial support.

“In God We Trust: All Others Bring Data”

(William Edwards Deming)

Resumo

A previsão é de extrema importância para a Indústria do Turismo. O desenvolvimento de modelos para prever a demanda de visitação a locais específicos é essencial para formular planos e políticas de desenvolvimento turístico adequados. também é essencial reduzir os impactos e custos negativos. Normalmente, as cidades e os países investem uma grande quantidade de dinheiro no planejamento e na preparação para receber (e lucrar) os turistas. O sucesso de muitos negócios depende em grande parte ou totalmente do estado da demanda turística. A estimativa da demanda turística pode ser útil para planejadores de negócios na redução do risco de decisões sobre o futuro, uma vez que os produtos turísticos são, em geral, perecíveis (desaparecem se não forem usados). No entanto, há um conjunto de desafios a superar, por exemplo, a maioria dos estudos anteriores neste domínio enfoca a previsão para um país inteiro e não para áreas de granulação fina dentro de um país (por exemplo, atrações turísticas específicas), principalmente por causa da falta de censo e dados disponíveis. Em outras palavras, apenas um número limitado de trabalhos e baselines estão disponíveis para lidar com o difícil problema de previsão de demanda turística de granulação fina (por atração). O outro desafio é a alta incerteza da demanda turística devido à interferência de fatores como taxa de câmbio, preço do combustível, mudanças climáticas, crises financeiras locais e globais e até epidemias e pandemias sobre comportamento cíclico e/ou tendencia de visitações em que poderiam causar desvios dramáticos nas previsões de demanda turística, se não forem devidamente consideradas.

Por outro lado, com o rápido crescimento da popularidade dos aplicativos de mídia social, a cada ano mais pessoas interagem nos recursos online para planejar e comentar suas viagens. Motivados por tal observação, sugerimos aqui que os dados acessíveis em redes sociais online ou sites de viagens, além dos dados ambientais, podem ser usados para apoiar a inferência da contagem de visitação para atrações turísticas internas ou externas.

Além disso, argumentamos que três requisitos-chave de previsão de turismo de granulação fina devem ser atendidos: (i) recência - os modelos de previsão devem considerar o impacto de eventos recentes; (ii) sazonalidade - o comportamento do turismo é inerentemente sazonal; e (iii) especialização do modelo - atrações individuais podem ter padrões idiossincráticos de visitação muito específicos que devem ser levados em consideração. Argumentamos que esses três requisitos principais devem ser considerados explicitamente e em conjunto para fazer avançar o estado da arte em modelos de previsão de turismo.

Nossa solução para os desafios na previsão do turismo de granulação fina é uma nova arquitetura que usa em conjunto dados de mídia social e recursos ambientais, adaptável a diferentes cenários de demanda turística, enquanto também propomos a inclusão conjunta de três requisitos principais do turismo - recência, sazonalidade e a especialização de modelos de previsão não apenas para captar os aspectos sazonais da demanda turística, mas também acompanhar as tendências recentes devido às mudanças locais/globais.

Em nossos experimentos, analisamos contagens de visitação, características ambientais e dados de mídia social relacionados a 27 museus e galerias no Reino Unido, bem como a 76 parques nacionais nos Estados Unidos. Nossos resultados experimentais revelam altos níveis de precisão para prever a demanda turística enquanto quantificamos o efeito de cada um tipo desses recursos. Também mostramos que a incorporação explícita de requisitos de turismo como recursos nos modelos pode melhorar a taxa de previsões altamente precisas em mais de 320% em comparação com o estado da arte atual. Além disso, eles também ajudam a resolver casos de previsão muito difíceis, anteriormente insolúveis pelos modelos atuais. Também fornecemos análises aprofundadas sobre o desempenho dos modelos nos cenários (simulados) em que é impossível cumprir todos os três requisitos - por exemplo, quando não temos dados históricos suficientes para uma atração para capturar sazonalidade. Finalmente, outra contribuição do nosso artigo é uma quantificação do impacto de cada um dos três fatores nos modelos aprendidos. Nossos resultados mostram que os mais importantes são, de fato, a especialização do modelo e a sazonalidade, mas a recência é muito eficaz quando não há dados históricos suficientes sobre uma atração específica.

keywords: Previsão de demanda de turismo, previsão detalhada, análise de séries temporais, dados de mídia social, dados ambientais

Abstract

Forecasting is of the utmost importance for the Tourism Industry. The development of models to predict visitation demand to specific places is essential to formulate adequate tourism development plans and policies. It is also essential to reduce negative impacts and costs. Usually, cities and countries invest a huge amount of money for planning and preparation in order to welcome (and profit from) tourists. The success of many businesses depends largely or totally on the state of tourism demand. Estimation of tourism demand can be helpful to business planners in reducing the risk of decisions regarding the future since tourism products are, generally speaking, perishable (gone if not used). However, there are a set of challenges to overcome, for instance most of prior studies in this domain focus on forecasting for a whole country and not for fine-grained areas within a country (e.g., specific tourist attractions) mainly because of lack of official census and available data. In other words, only a limited number of works and baselines are available which deal with the hard problem of fine-grained (per attraction) tourism demand prediction. The other challenge is the high uncertainty of tourism demand due to interference of factors like exchange rate, fuel price, climate changes, local and global financial crises and even epidemics and pandemics over cyclic and/or trending behavior of visitations in where could cause dramatic deviations in tourism demand forecasts, if they are not properly considered.

On the other hand, with the rapid popularity and growth of social media applications, each year more people interact within online resources to plan and comment on their trips. Motivated by such observation, we here suggest that accessible data in online social networks or travel websites, in addition to environmental data, can be used to support the inference of visitation count for either indoor or outdoor tourist attractions.

In addition, we argue that in the context of fine-grained tourism prediction, three specific key requirements should be fulfilled: (i) recency – forecasting models should consider the impact of recent events; (ii) seasonality – tourism behavior is inherently seasonal; and (iii) model specialization – individual attractions may have very specific idiosyncratic patterns of visitations that should be taken into account. We argue that these three key requirements should be considered explicitly and in conjunction to advance the state-of-the-art in tourism prediction models.

Our solution to the challenges in fine-grained tourism prediction is a novel architecture using in jointly social media data and environmental features, adaptive to different scenarios of Tourism demand, while we also propose conjunctive inclusion of three main tourism requirements - recency, seasonality and model specialization in the prediction models not only to be able to capture the seasonal aspects of tourism demand but also follow the recent trends due to local/global changes.

In our experiments, we analyze visitation counts, environmental features and social media data related to 27 museums and galleries in the U.K. as well as 76 national parks in the U.S. Our experimental results reveal high accuracy levels for predicting tourism demand while we quantify the effect of each type of these features. We also show that the explicit incorporation of Tourism requirements as features into the models can improve the rate of highly accurate predictions by more than 320% against the current state-of-the-art. Moreover, they also help to solve very difficult prediction cases, previously unsolvable by the current models. We also provide in depth analysis regarding the performance of the models in the (simulated) scenarios in which it is impossible to fulfill all three requirements – for instance, when we do not have enough historical data for an attraction to capture seasonality. Finally, another contribution of our paper is a quantification of the impact of each of the three factors in the learned models. Our results show that the most important ones are indeed model specialization and seasonality but recency is very effective when there is not enough historical data about a specific attraction.

keywords: Tourism demand prediction, fine-grained prediction, time-series analysis, social media data, environmental data

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Motivation

According to the World Travel and Tourism Council (WTTC), as of 2019 annual research covering 185 countries and economies, the global travel and tourism contribution to the GDP[1] is at 10.4% supporting 319 million jobs. This corresponds to 10% of the global employment. Considering new jobs across the world, the contribution of travel and tourism industry is even higher, achieving 25% of all global new jobs created over the last five years[2]. Thus having estimated values of future tourism demand in the weeks, months, and years ahead can serve as a base for preparing activities necessary for creating comprehensive tourism policies [Chetty, 2011].

Thus, decision makers in tourism related industries such as transportation, accommodation facilities, hotels and traveling agencies, all need to have good estimates of future demand in the weeks, months, and even years ahead in their businesses. Without reliable estimates of future demand, it is difficult, if not impossible, to formulate adequate tourism development plans and policies [Chetty, 2011]. In addition, it is a principal tourism policy to ensure that visitors are hosted in a way that maximizes the benefits to them and stakeholders, while minimizing the negative effects, costs, and impacts [Goeldner and Ritchie, 2006]. Generally, tourism products are perishable [Frechtling, 2012]. This is the case for unsold seats in a flight after it has already taken off, unsold tickets to a park when it has closed for the day; empty rooms in a hotel the next day. All such products, and thus the revenue opportunities associated with them, vanish with time. This indicates the importance of not only shaping demand in the short run but also anticipating it in the long run, to avoid unsold

[1] Gross domestic product (GDP) is a monetary measure of the market value of all the final goods and services produced in a specific time period

[2] Global Economic Impact & Trends 2020 at link $https://wttc.org/Research/Economic-Impact$

'inventory' and unfulfilled demand.

In this context, the development of models to predict future visitation demand to specific places and regions can be of great benefit. However, there are a set of challenges to overcome, for instance most of prior studies in this domain focus on forecasting for a whole country and not for fine-grained areas within a country (e.g., specific tourist attractions) mainly because of lack of official census and available data. In other words, only a limited number of works and baselines are available which deal with the hard problem of fine-grained (per attraction) tourism demand prediction. The other challenge is the high uncertainty of tourism demand due to interference of factors like exchange rate [Webber, 2001], fuel price, climate changes [Hengyun Li and Li, 2016], local and global financial crises [Maditinos and Vassiliadis, 2008] and even epidemics and pandemics such as covid/19[3] over cyclic and/or trending behavior of visitations in where could cause dramatic deviations in tourism demand forecasts, if they are not properly considered.

A possible solution to the above challenges is the use of accessible data in online social networks or travel websites since with the rapid popularity and growth of social media applications, each year more people interact within online resources to plan and comment on their trips. As a result, most of these factors are reflected quickly in social media [Asur and Huberman, 2010; Chunara et al., 2012; Vecchio et al., 2018] due to the large number of participating users (from global cities to metropolitans, and even small villages) and vast amounts of content produced and shared on a daily basis.

According to Statista[4], by 2018, 71% of Internet users are social media users (2.62 billion people) and this number is expected to grow even further. Another study of Statista[5] shows that by 2022, more than 90% of Internet users will be social media users; this is almost 4 billion people in the world, i.e. half of the world population. Another study in 2019 [6] reveals that 76% of tourists post vacation photos or comment during their visitations while 40% post Restaurant reviews and 46% post Hotel reviews.

The frequent participation of billions of users in social media websites like TripAdvisor, Instagram, Facebook and Foursquare, and the huge amounts of social media data produced by users [Vecchio et al., 2018] suggest a possibly high correlation between users' interactions in online social networks and their real world activities. Among various online social networks, TripAdvisor, the world's largest travel site[7], with over 630 million reviews

[3] In 2020, Travel & Tourism faced unprecedented challenges and an existential threat (till the time of writing this work at April 2021) from the impact of the COVID-19 virus globally. However, according to the World Travel & Tourism Council, one of the key drivers in the sector's recovery from COVID-19 would be domestic tourism (tourism inside the country).

[4] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[5] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[6] https://visual.ly/community/Infographics/travel/impact-social-media-travel-and-hospitality-industry

[7] comScore Media Metrix for TripAdvisor Sites, worldwide, October 2017

and opinions, is covering the world's largest selection of travel listings worldwide. TripAdvisor has the world's largest travel community, with 490 million average monthly unique visitors[8] with over over 50% of users browse the site on their mobile[9].

In spite of the large amount of user interactions in social media websites, one of the challenges in tourism prediction is official data gathering. Indeed, the number of visits for many touristic points is neither well-documented nor easily available. Moreover, conducting surveys at entrances of major attractions is expensive and provides only limited spatial and temporal coverage. The situation is worse in developing countries and even more complicated regarding remote touristic sites. This is possibly the reason why most prior efforts of touristic activities forecasting proposed solutions where the prediction models are built and tested over a whole country and not for specific regions or attractions [Wang, 2004; Cankurt and Subasi, 2015; Chang Jui Lin, 2011].

There are several studies on attraction recommendation (e.g., [Borras et al., 2014]), however their main focus is on the users/tourists and not on the attractions. In other words, those studies mostly focused on recommending to a given tourist the proper attractions based on the history of her preferences, her social network and other personal factors. In contrast, our main focus in this dissertation is on the prediction of the number of visits for the attractions themselves to help the responsible authorities, public and private attractions managers and owners to better plan the reception of their visitors.

In fact, there is a real need to develop robust prediction models that not only forecast well the future visits by considering seasonal aspects of tourism behaviour but also show flexibility to recent trends and events and idiosyncratic aspects of the attractions. To this aim, we go further, by *explicitly* incorporating aspects related to recency, seasonality and model specialization into our prediction models. We show improvements in our results in the task of fine-grained prediction, mainly regarding highly accurate predictions exploiting the three key aforementioned requirements of (1) *recency*, (2) *seasonality* and (3) *model specialization*. We defend that these requirements should be captured as *explicit features* or properties of models for tourism forecasting.

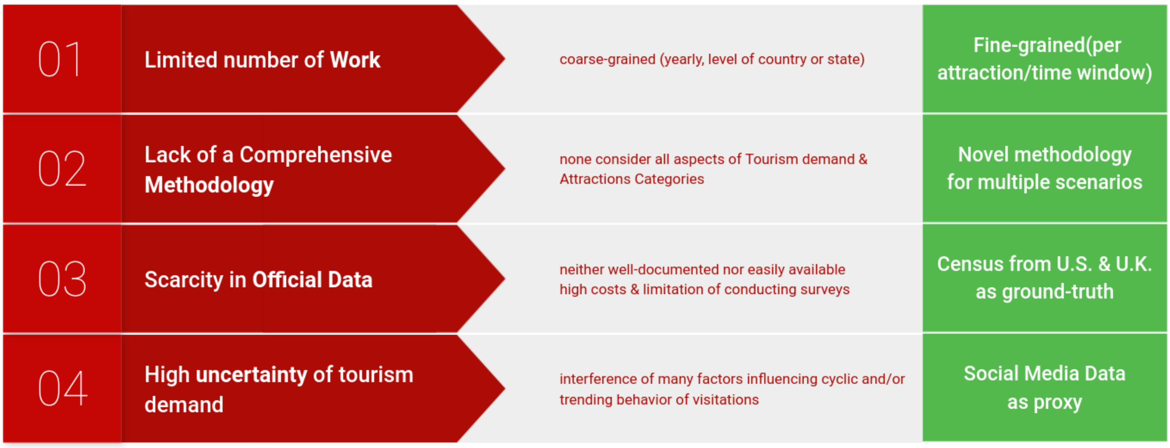
All in all, our solution to the challenges in fine-grained tourism prediction is a novel architecture using in jointly social media data and environmental features [10], adaptive to different scenarios of Tourism demand, while we also propose conjunctive inclusion of three main tourism requirements - recency, seasonality and model specialization in the prediction models not only to be able to capture the seasonal aspects of tourism demand but also fol-

[8] TripAdvisor log files, average monthly unique visitors, 2020

[9] https://review42.com/tripadvisor-statistics/

[10] In this work, we consider only climate-related features such as temperature and precipitation as relevant environmental features in the task of tourism demand prediction.

low the recent trends due to local/global changes. Finally Figure 1.1 summarizes what we discussed in this section naming the challenges in Fine-grained Tourism Demand Prediction and our novel solutions.



01	Limited number of Work	coarse-grained (yearly, level of country or state)	Fine-grained(per attraction/time window)
02	Lack of a Comprehensive Methodology	none consider all aspects of Tourism demand & Attractions Categories	Novel methodology for multiple scenarios
03	Scarcity in Official Data	neither well-documented nor easily available high costs & limitation of conducting surveys	Census from U.S. & U.K. as ground-truth
04	High uncertainty of tourism demand	interference of many factors influencing cyclic and/or trending behavior of visitations	Social Media Data as proxy

Figure 1.1: The challenges in Fine-grained Tourism Demand Prediction and our novel solutions

1.2 Problem Statement

The problem we face in this dissertation is forecasting the visitation for fine-grained touristic points. In addition to exploiting social media and environmental features, we also incorporate both recency and seasonality requirements. Moreover, we consider per-attraction model specialization. Given a touristic attraction, we define the prediction problem as follows.

First, we create equally spaced non-overlapping time windows with the same temporal granularity (e.g., a month, a week, a day, an hour, etc) for each time-series of the variables in social media, environmental data, recency and seasonality features in the format of $X = \{X_1, X_2, \ldots, X_m\}$ where m is the number of features. These time-series (e.g., number of reviews, average temperature, visits in the last month, visits in the last year) serve as the input of the prediction models. A time series X_i is a sequence $\{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)}\}$, where $x_i^{(t)}$ denotes the value of variable X_i measured (time-lagged) in time window t for the specific touristic attraction that is the target of prediction. **Measured** variables are social media and environmental features that have been measured at timestamp t while recency and seasonality features correspond to the history of visitation counts (i.e. response variable) in recent months or last year of visitation, which have been augmented and **time-lagged** to the time window t.

The objective function (f) is forecasting $y^{(t)}$, the tourism visitation at timestamp t in a target attraction with the lowest prediction error, giving the input vector X as the feature-set including social media, environmental, recency and seasonality features for each time window t in the interval of $[1, t-k]$ (for $k > 0$) as in Equation 1.1:

$$y^{(t)} = f\ (x_1^{(1)},\ x_2^{(1)},\ ...,\ x_m^{(1)},\ x_1^{(2)},\ x_2^{(2)},\ ...,\ x_m^{(2)},\ ...,\ x_1^{(t-k)},\ x_2^{(t-k)},\ ...,\ x_m^{(t-k)}) \tag{1.1}$$

where m is the number of available features[11]. In this work, we tackle the problem of predicting visitation counts at specific touristic places, exploiting external features such as social media data and environmental features besides idiosyncratic aspects of the attractions including recency, seasonality and model specialization into our prediction models.

1.3 Dissertation Hypothesis

The main research Hypothesis we investigate in this work is the following:

Main Hypothesis: Different factors - external and intrinsic - have a great influence on the accuracy of models for predicting visits in tourist attractions.

Our hypothesis is that external features such as timely number of users' reviews in **social media** and **environmental features** such as temperature, precipitation, air frost days, etc. correlate systematically with real visitations of touristic attractions. Therefore, by using social media data alongside environmental features, one can create more accurate tourism prediction models of tourism demands for fine grained touristic attractions. Second, there are intrinsic patterns in timely visitation of touristic attractions that could be explicitly imported into prediction models including **recency** and **seasonality** features. Recency considers the impact of recent events into the prediction models. Although most works in the literature focuses on the importance of seasonality as the main temporal aspect for tourism prediction, we argue that other temporal aspects should be considered as well [Moro and Rita, 2016] to assess whether and how recent events such as financial crises, new trends, epidemics/ pandemics[12], new infrastructures, may impact the predictions.

Seasonality focuses on the inherently cyclic behaviour of tourism demands. Indeed, seasonality is one of the main phenomena affecting tourism, corresponding to movements of a variable in a selected period of time, usually the year or from season to season [Hylleberg, 1992]. There are many works that explore seasonality implicitly in their prediction models including [Khatibi et al., 2019], [Petrevska, 2013] and [Kulendran and Wong, 2005].

[11] In some cases the objective function f combines input vector X and the response variable y.

[12] We have seen lately how the new COVID/19 pandemics has completely shut down the Tourism Industry worldwide. Forecasting its impact in the Tourism Industry could have helped to better manage the crisis.

Model specialization regards the situation in which we create a specialized individual model for each attraction. This can be advantageous since individual attractions may have very specific idiosyncratic patterns of visitations. On the other hand, there may be cases when we do not have enough data to train individual model for each site in which it is more viable to train single models for attractions of a given type to benefit from a vast amount of available social, climate and official data in the training process. Adopting all the external and intrinsic factors into **specialized models** for each attraction/group of attractions can be advantageous since individual attractions may have very specific idiosyncratic patterns of visitations.

We investigate our main hypothesis by answering some research questions. More specifically, our goal is to quantify the influence of each of the factors in the accuracy of our prediction task. As such, our main research questions are:

RQ 1: How online social media contents and environmental features influence the accuracy of predicting visits in touristic attractions? We compute the correlation of monthly real number of visits in each of touristic attractions with monthly total number of social media reviews in more than 70 U.S national parks and 27 U.K national museums and galleries in order to quantify their performance. In addition, we evaluate the feasibility of exploiting environmental data alongside social media in order to accurately forecast tourism demands in a fine-grained approach for specific attractions.

RQ 2: How recency, seasonality and model specialization (characteristic of attraction) influence the accuracy of predicting visits in tourist sites? Adopting our collected rich set of indoor and outdoor attractions, we explicitly exploit recency and seasonality features into global and specialized models in order to evaluate the impact of each of key requirements of tourism prediction. Furthermore, we analyse scenarios with data scarcity whether in recent or seasonal data for attractions where we show how the absence of recency or seasonality features drastically reduces the accuracy of prediction models.

1.4 Contributions and Outline of this Dissertation

In this dissertation, we aim to propose novel solutions to the challenges in fine-grained tourism prediction. We design novel architecture using in jointly social media data and environmental features, adaptive to different scenarios of Tourism demand, while we also propose conjunctive inclusion of three main tourism requirements - recency, seasonality and model specialization in the prediction models not only to be able to capture the seasonal aspects of tourism demand but also follow the recent trends due to local/global changes. We demonstrate that by explicitly exploiting the three proposed key requirements of tourism prediction as features in our models beside feeding the models with robust external features such as social media and environmental features, we can greatly improve prediction accuracy regarding the state-of-the-art results.

Figure 1.2 presents the flow of our analysis in the process of investigating each of research questions. We also highlight for each research question the issues we analysis. Our final goal is to develop robust prediction models that not only forecast well the future visits by considering seasonal aspects of tourism behaviour but also show flexibility to recent trends/events and idiosyncratic aspects of each attraction. In this way, the proposal of our work and methodology is **fine-grained** demand prediction where in the spatial aspect we go as low as attraction level and for time window as low as monthly granularity. However, the reason for restricting to attraction level and monthly prediction is the granularity of available official data as the ground-truth of our analysis.

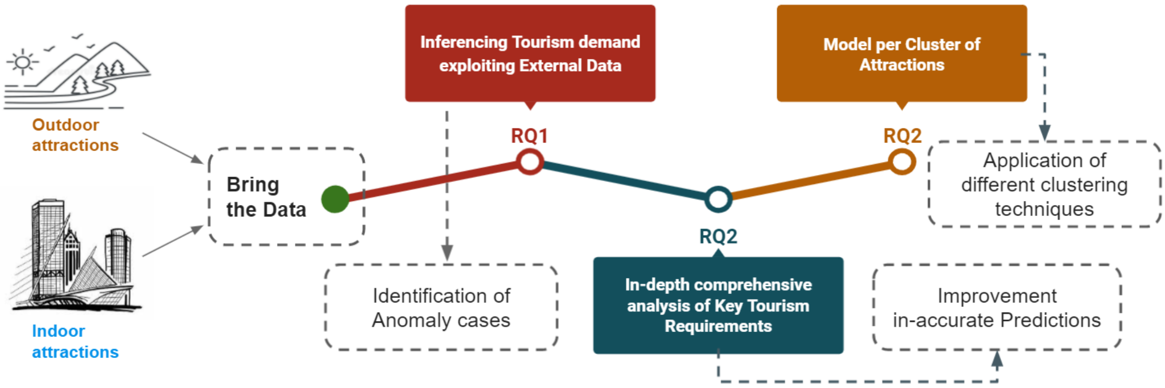

Outdoor attractions
Indoor attractions
Bring the Data
Inferencing Tourism demand exploiting External Data
RQ1
Identification of Anomaly cases
RQ2
In-depth comprehensive analysis of Key Tourism Requirements
Model per Cluster of Attractions
RQ2
Application of different clustering techniques
Improvement in-accurate Predictions

Figure 1.2: Comprehensive Fine-grained Tourism Demand Prediction

Our proposed models can even help to solve problematic or difficult prediction cases, poorly solvable by the current solutions. For instance, the National Portrait Gallery in U.K. saw a huge increase in social media reviews (over 50% by April 2015) but that was not accompanied by real world visits, causing the models to mistakenly follow the social patterns, ultimately implying in low accuracy. Another example is the Bryce Canyon national park in the U.S., in which the visits had untypical increases in the some months (more than 20% in Feb. to Sep. 2016 in comparison with the same period in 2015). That increase was not not reflected neither in the environmental features nor in the social media reviews, both inputs of the models. We claim that these situations can be dealt with by explicitly incorporating recency and seasonality features.

This dissertation is organized as follows. The main concepts of the work are discussed in a single chapter (Chapter 2), along with related work. In addition, we devoted one chapter (Chapter 3) to explain the methodology, in general, dataset specifications and evaluation metrics. Afterwards, for the sake of fluidity of the text we present the experimental results in separate chapters instantiating each type of attractions – outdoors and indoors. The detailed description of our main contributions is organized in the following Chapters:

- **Chapter 3** presents our collected dataset specification, experimental methodology and features exploited in the prediction models, and the evaluation metric. The investigation of both proposed **RQs** requires a rich dataset to permit in-depth analysis of different aspects of exploited features, category of attractions, multiple prediction models and effect of tourism requirements. To do this broad analysis, we collect, join and analyze five different datasets – TripAdvisor reviews and ratings, U.S National Park Service, U.S national climate data center, Department for Digital, Culture, Media and Sport of England and finally U.K national weather service covering two types of attractions (indoor and outdoor attractions). We made these datasets available online in our data in brief paper [Khatibi et al., 2020] for future experiments in the area of fine-grained tourism analysis. We are one of very few works in the literature to perform fine-grained (i.e., attraction-level, with at least monthly granularity) predictions of visitation counts exploiting both environmental and social media data to improve such predictions. Furthermore, we do this for more than one hundred attractions while most works focus on a single or just a few attractions.

- **Chapter 4** regards our study on RQs for **outdoor** attractions. First, we study the **RQ1** analyzing correlation and prediction results adopting official visitations, environmental features and social media data in more than 70 national parks in the U.S. Our experimental results reveal high accuracy levels (above 92%) for predicting tourism demand using features from both social media and environmental data. We compare the effectiveness of eight different prediction techniques, namely, Linear Regression, Support Vector Regression, General Regression Neural Network, Seasonal ARIMA, SARIMAX, LSTM and two naive models - naive recency and naive seasonality, for the tourism demand prediction task. We also perform a detailed failure analysis to inspect the cases in which the prediction results are not satisfactory.

  Then, regarding **RQ2**, we demonstrate that three tourism key requirements, i.e. recency, seasonality and model specialization are essential for fine-grained high-accuracy tourism demand prediction task. More than that, these requirements should be incorporated as explicit features into the learning models. Our experimental evaluation confirm our hypotheses, with observed gains over the other solutions. We also show that the explicit incorporation of such requirements into the models help to solve very hard-to-solve cases. One type is when social media reviews had a huge increase due to some aspects in the virtual world without implications in the real world causing the models mistakenly try to follow the patterns in social media features. The second type is when visits had untypical increases in the recent months without prominent changes or reflections in input of the models, i.e. environmental features and social media reviews. Consequently, these flaws can be corrected calibrating the models by adopting the explicit use of tourism requirement features.

We perform a factorial analysis in order to quantify the impact of each of the three requirements on the accuracy of the learned models in outdoor attractions. We show that the most impacting ones are indeed model specialization and seasonality but recency is very effective when there is not enough historical data about a specific attraction. We also study the possibility of creating group of similar attractions in order to build a single prediction model for each group, useful in scenarios where we have little information about some attractions. Finally, we perform a study on the performance of each of the tourism prediction requirements in cases with no recent or historical data for an attraction. We show that the absence of recency or seasonality features drastically reduces the accuracy of prediction models in different scenarios.

- **Chapter 5** focuses on studying **RQs** this time for **indoor** attractions. Again, first we focus on **RQ1** analyzing correlation and prediction results adopting official visitations, environmental features and social media data in 27 museums and galleries in the U.K. Using features from both social media and environmental data, we present experimental results with high accuracy levels (above 93%) in indoor attractions adopting eight different prediction techniques, namely, Linear Regression, Support Vector Regression, General Regression Neural Network, Seasonal ARIMA, SARIMAX, LSTM and two naive models - naive recency and naive seasonality, for the tourism demand prediction task. We also perform a detailed failure analysis to inspect the cases in which the prediction results are not satisfactory.

  Next, alike outdoors, we focus on **RQ2** in indoor attractions. We demonstrate that incorporating the three tourism key requirements – recency, seasonality and model specialization – as explicit features into the learning models is essential. Our experimental evaluation confirms our hypotheses, with observed gains over the other solutions. In addition, we elaborate that the explicit incorporation of such requirements into the models help to solve very hard-to-solve cases.

  Finally, similar to outdoor attractions, we do a factorial analysis of impact of each of the three requirements on the accuracy of the learned models in the scenario of indoors concluding that the most impacting ones are model specialization and seasonality. Alike outdoor attractions, we also evaluate the performance of models when we build a model for each group of attractions. We complete our study, analysing the performance of each of the tourism prediction requirements in cases with no recent or historical data for an attraction concluding that the absence of recency or seasonality features drastically reduces the accuracy of prediction models in different scenarios as like as what we observed in outdoors.

- **Chapter 6** compares our results with those in previous work in addition to summarizing our results to answer RQs in indoor versus outdoor attractions. We show that, for outdoor attractions, environmental and seasonal features have better predictive power while the opposite occurs for indoor attractions where social media and recency features play a more important role. In any case, best results, in all scenarios, are obtained when using all types of features, i.e. external data features jointly with key tourism requirements. Finally, we conclude this dissertation and present a discussion on directions for future work.

1.5 Publications and Submissions

The following is the list of the papers produced during the PhD period:

- Amir Khatibi, Ana Paula Couto da Silva, Jussara M. Almeida, Marcos André Gonçalves, **On the Role of Recency, Seasonality and Model Specialization in Fine-Grained Tourism Demand Prediction**, *submitted* to journal of Transactions on Intelligent Systems and Technology (ACM TIST) at 24-Dec-2020 - **Qualis A1, Impact Factor: 3.971**

- [Khatibi et al., 2020] Amir Khatibi, Ana Paula Couto da Silva, Jussara M. Almeida, Marcos André Gonçalves, **FISETIO: A FIne-grained, Structured and Enriched Tourism Dataset Indoor and Outdoor attractions**, *published* as Data-paper in Journal of Information Processing and Management (IP&M) 2020, Data in brief 28 (2020): 104906 - **Qualis A1, Impact Factor: 4.787**

- [Khatibi et al., 2019] Amir Khatibi, Fabiano Belem, Ana Paula Couto da Silva, Jussara M. Almeida, Marcos André Gonçalves, **Fine-Grained Tourism Prediction: Impact of Social and Environmental Features**, *published* in journal of Information Processing & Management (IP&M), 57(2), 102057 - **Qualis A1, Impact Factor: 4.787**

- [Khatibi et al., 2018] Amir Khatibi, Fabiano Belem, Ana Paula Couto da Silva, Dennis Shasha, Jussara M. Almeida, Marcos André Gonçalves, **Improving tourism prediction models using climate and social media data: A fine-grained approach**, *presented* at 12th International AAAI Conference on Web and Social Media, ICWSM 2018, 636-639 - **Qualis A1**

Chapter 2

Background and Related Work

This chapter has two main goals. First, we introduce core aspects of the machine learning techniques we exploit in our tourism prediction models (Section 2.1). Afterwards, we discuss studies related to the main dissertation contributions. We group these studies into three main topics: coarse-grained tourism prediction models, fine-grained prediction models and prediction models that include one or more tourism requirements,i.e., seasonality, recency and model specialization (Section 2.2).

2.1 Prediction Techniques Description

This section offers a description of the techniques we exploit for forecasting fine-grained tourist visit counts. There are many works from Keogh and his colleagues like [Shokoohi-Yekta et al., 2015; Yeh et al., 2017; Zhu et al., 2018] and [Zhu et al., 2019] where they study robust ways in the task of time series prediction. For example in [Shokoohi-Yekta et al., 2015], they mention that most of previous work has attempted to predict the future based on the current value of a time-series. However, for many problems the actual values are irrelevant. As a result, they propose novel algorithms to quickly discover high quality rules in time series datasets in order to accurately predict the occurrence of future events. In another work, [Yeh et al., 2017] study how to take advantage of weak labeled historical time series data to predict qualified outcomes.

There are also other studies in the literature where the authors adopt machine learning techniques into the context of Time Series prediction. For instanse in [Parmezan, 2016], authors propose a modification of the k-Nearest Neighbors (kNN) learning algorithm for time series prediction, namely the kNN-time series prediction in which time is an important factor. However, in this work we adopt most promising and used prediction techniques in the literature in the task of demand prediction.

Recall that our prediction task is to estimate $y^{(t)}$, the number of visits in a given touristic place in the timestamp t giving the input vector X as the feature-set including social media, environmental, recency and seasonality features for each time window in the interval of $[1, t-k]$.

2.1.1 Linear Regression

When there is a strong linear relationship between the response variable (i.e., visit count) and the predictor variables in the dataset, a linear regression model might be applied as a simple yet effective straightforward prediction technique. Indeed, linear regression has been shown to be quite cost-effective in various prediction tasks [Koutras et al., 2017; Vasconcelos et al., 2015]. The model generated by linear regression is given by:

$$f(X) = W \cdot X + b, \tag{2.1}$$

where W and b are the regression coefficients. Linear regression based estimators usually aim at minimizing the sum of squared residuals (differences between estimated and actual values of $y^{(t)}$).

2.1.2 Support Vector Regression

Support Vector Regression (SVR) is an extension of Support Vector Machines (SVM) widely used for regression tasks [Drucker et al., 1996]. SVR performs a "linear regression" in a high-dimensional feature space resulting from a (nonlinear) mapping provided by a kernel function. The linear model (in the feature space) is given by:

$$f(X) = \sum_{j=1}^{m} W_j g_j(X) + b, \tag{2.2}$$

where W is the weight vector to be ”learned”, $g_j(X)$ denotes a set of nonlinear transformations on the input feature set, and b is the ”bias” term. SVR pursues the best trade-off between the model's empirical error and the model complexity by constraining SVR regression function f(,) to the hyper-planes function class, and employing a margin around the hyper-plane. Moreover, f(,) only depends on a reduced set of the training data called the Support Vectors (SV), those which correspond to the active constraints in the optimization problem [Drucker et al., 1996] defined as:

$$L(y, f(X)) = \begin{cases} 0 & |y - f(X)| \leq \varepsilon \\ |y - f(X)| - \varepsilon & |y - f(X)| > \varepsilon \end{cases} \tag{2.3}$$

where y is the value to estimate.

The key parameters of SVR are the kernel function K, the margin of tolerance ε, and the trade-off C between the model complexity and the degree to which deviations larger than ε are tolerated.

SVR has been successfully employed to solve time series problems in many fields including the tourism industry. Indeed, it has been shown to outperform other techniques, such as Multi-Layer-Perceptron (MLP) regression, in the task of tourism demand prediction [Cankurt and Subasi, 2015].

2.1.3 General Regression Neural Network

General Regression Neural Networks (GRNNs) [Specht, 1991] are computing systems inspired by the biological structure of connected neurons that constitute the animal brain[1]. Similarly to other Artificial Neural Networks (ANNs), a GRNN is composed by a set of nodes (artificial neurons), which propagates signals (real numbers) across their connections, when activated. The output $f(X)$ is calculated by a non-linear function of the sum of its inputs:

$$f(X) = \frac{\sum_{i=1}^{N} w_i K(X, X_i)}{\sum_{i=1}^{N} K(X, X_i)}, \tag{2.4}$$

where w_i is the activation neuron weight corresponding to the training instance i, and K is the radial basis function (RBF) kernel. The RBF kernel is defined as $K(X,X_i) = e^{-d_i/2\sigma^2}$, where $d_i = (X - X_i)^T (X - X_i)$, which is the squared euclidean distance between the training instance X_i and the test input X. The tuning parameter σ controls the smoothness of a GRNN.

2.1.4 Seasonal ARIMA (SARIMA)

ARIMA (Auto Regressive Integrated Moving Average) is a classical time series forecasting method which was firstly proposed by Box and Jenkins [Box and Jenkins, 1976]. In this model, the future value of a time series is a linear function of previous values of the original series and random errors. In other words, ARIMA projects the future values of a series based entirely on its own inertia. Thus, the set of predictor variables X used by ARIMA consists of the past measurements of the response variable $y^{(t)}$, that is, $X = \{y^{(1)}, y^{(2)}, ..., y^{(t-k)}\}$, $k > 0$.

Since in this dissertation we consider seasonal effect as a feature in our prediction model, we apply the SARIMA, i.e. the seasonal version of the standard ARIMA model. SARIMA model is an equation in the following form:

[1] In our experiments, we use the grnn package of the language R[2]

$$f(t) = \frac{\Theta\ \theta\ \varepsilon^{(t)}}{\Phi\ \phi\ \Delta\ \delta}, \tag{2.5}$$

where Θ, Φ and Δ are polynomials that compute the seasonal auto-regressive, differences and moving average components, respectively, θ, ϕ and δ quantify the respective regular (non seasonal) polynomials and $\varepsilon^{(t)}$ is the estimation error.

2.1.5 Seasonal ARIMAX (SARIMAX)

Due to the importance of exogenous data (i.e., social media, environmental data, recency and seasonality features) in our prediction task, we also run SARIMAX models (SARIMA with exogenous variables) in our experiments. SARIMAX exploits not only the history of response variable, but also the input features, i.e. the external predictor variables, that is, the set of time series $X = \{x_i^{(1)}, x_i^{(2)}, ..., x_i^{(t)}\}$ where $x_i^{(t)}$ denotes the value of variable X_i measured (time-lagged) in time window t. The SARIMAX model is formulated as:

$$f(X,t) = \frac{\Theta\ \theta\ \varepsilon^{(t)}}{\Phi\ \phi\ \Delta\ \delta} + \beta X, \tag{2.6}$$

where the definition of the parameters Θ, Φ, Δ, θ, ϕ and δ is the same as equation 2.5.

2.1.6 Long Short-Term Memory Neural Network

Introduced in [Hochreiter and Schmidhuber, 1997], Long Short-Term Memory (LSTM) neural network models are well-suited to classification and regression as well as prediction tasks based on time series data. LSTMs have a notion of memory that may help capturing past trends in the data. The use of LSTMs in the context of tourism prediction is not new; in [Li and Cao, 2018] the authors apply LSTM to tourism flow prediction.

A LSTM network consists of a chain of cells – each LSTM cell is configured by four gates: input gate, input modulation gate, forget gate and output gate. Input gates take new inputs from outside and process newly incoming data. Memory gates take inputs from the output of the LSTM cell in the last iteration. Forget gates decide when to forget the output results, thus selecting the optimal time lag for the input sequence. Output gates take all results calculated and generate final output [Hochreiter and Schmidhuber, 1997].

Consider a time-series input represented as $X = \{x_i^{(1)}, x_i^{(2)}, ..., x_i^{(t)}\}$ where $x_i^{(t)}$ denotes the value of variable X_i measured (time-lagged) in time window t and hidden state cells $H = \{h^{(1)}, h^{(2)}, ..., h^{(t)}\}$. For $t = 1, ..., T$, LSTM computes:

$$f(X,t) = W_{hy}h^t + b_y \tag{2.7}$$

$$h_t = H(W_{hy}x^t + W_{hh}h^{t-1} + b_h), \tag{2.8}$$

where W and b are respectively weight matrices and bias vector parameters which need to be learned during model training.

2.1.7 Naive Models

In general, naive forecasting models are simple models that are based exclusively on historical observation. In the dissertation, we defined two naive models as our baselines based on seasonality and on recency of tourism activities - Naive-Seasonality and Naive-Recency. For a naive prediction with seasonality, a simple approach to determine $y^{(t)}$ in a time window t, is to pick the number of visits at $y^{(t-12)}$ in the available past data. Similarly, for recency, we predict $y^{(t)}$ based on the number of visits at $y^{(t-1)}$.

$$f(t) = \begin{cases} y^{(t-12)} & naive - seasonality \\ y^{(t-1)} & naive - recency \end{cases} \tag{2.9}$$

2.1.8 Summary of Prediction Techniques

In this section, we reviewed various prediction techniques in the task of tourism demand forecasting. Table 2.1 lists all prediction techniques considered in this dissertation, summarizing their main characteristics. We here exploit a very diverse set of techniques, aiming at investigating how each of them performs in our target prediction problem.

Table 2.1: Prediction Techniques Considered in this dissertation

Method	Exploit history of visits	Exploit social media and environmental features	Consider temporal dependency among data observations
SVM	×	✓	No
Linear Regression	×	✓	No
GRNN	×	✓	No
SARIMAX	✓	✓	Yes
SARIMA	✓	×	Yes
LSTM	✓	✓	Yes
Naive Models	✓	×	Yes

2.2 Related Work

In this section, we discuss the related studies to our work dividing into three main topics: coarse-grained tourism prediction models (Section 2.2.1); fine-grained prediction models (Section 2.2.2) and prediction models that include one or more tourism requirements, i.e. seasonality, recency and model specialization respectively in Sections 2.2.3, 2.2.4 and 2.2.5. Finally Section 2.2.6 summarizes the prior work distinguishing our contribution in a broad view.

2.2.1 coarse-grained Tourism Prediction

Previous work on predicting tourism visitation focused mainly on forecasting tourism demands at a country or state-level [Pai et al., 2014; Zhang and Zhang, 2011; Kamel et al., 2008; Cankurt and Subasi, 2015], which we refer to as *coarse-grained predictions*. These prior efforts used different sets of features to create robust prediction models. Some features that have already been exploited include climate related attributes (temperature, duration of sunshine and number of rainy days), US Dollar exchange rate, hotel bed capacity, visitation history, among others.

Authors of [Cankurt and Subasi, 2015] use Multi-Layer-Perceptron (MLP) regression and Support Vector Regression (SVR) models to make multivariate tourism forecasting for Turkey. They use a dataset composed of the monthly time-series of a large set of features including wholesale prices index, US Dollar selling, hotel bed capacity, number of tourism agencies in the country, number of the tourists coming from the top 10 visiting countries, exchange rate of the leading countries and 22 other variables, ranging from 1996 to 2013. Best prediction results were obtained by employing SVR. Yet, their work is at the level of an entire country, given that most features are publicly available only at that granularity level.

In [Khadivi and Ramakrishnan, 2016], the authors use Wikipedia usage trends (WUT) in order to forecast tourism demand of Hawaii Island. By collecting monthly time series of official tourism demand for Hawaii as well as Wikipedia pages and their usage trends, they try to explore how WUT influences the accuracy of tourism demand forecasts. They show that, on average, most of the Wikipedia reading activities occur about 4 to 8 months prior to the trip as the mean decision date for most of the activities are between 4 to 8 months before the actual arrival date. However, they report the accuracy of their prediction results only by RMSE (Root Mean Square Error) using an auto-regressive exogenous model where the external variable is a Wikipedia usage trend time series. RMSE is a measure of accuracy to compare forecasting errors of different models for a particular data and not between datasets, as it is scale-dependent [Hyndman and Koehler, 2006]. Although there are interesting statements and results in this work, there is no comparison of the proposed prediction models to other baselines nor any assessment of the results in a comparable manner.

[Huang et al., 2017] study the search engine trends in the top search engine in China, Baidu. The authors analyze the relationship between the internet search data in Baidu search engine (calling it as Baidu Index) and the actual tourist flow. The paper compares two ARIMA prediction models, with and without considering the Baidu Index. They report improving the results (RMSE value) by 12% when employing the Baidu index. However, their study is only for a single city, Beijing Forbidden City. They show that a long-run equilibrium relationship exists between the actual visitor numbers for the Forbidden City and Baidu keyword with a lag of 1-2 days.

There are also studies that exploit data from Location Based Social Networks (LBSN) (e.g., Foursquare and Yelp) to study mobility of tourists and citizens [Li and Chen, 2009; Cho et al., 2011; Hasan et al., 2013; Hossain et al., 2016], modeling friendship processes [Valverde-Rebaza et al., 2018] or traffic patterns, but mostly in a coarse-grained fashion. For instance, in [Georgiev et al., 2014], the authors study dining and shopping behaviors during the 2012 Summer Olympics in London using Foursquare check-in data. They model the impact of the Olympic Games on local retailers in different market segments by analyzing location-based social data collected from Foursquare. The authors suggest that spatial positioning of businesses as well as the mobility trends of visitors are primary indicators of whether retailers will rise their popularity during events such as the Olympic Games. They also argue that location-based information in social media data reflects the dynamic interaction of users within urban spaces, presenting an alternative means to predict the tourism demands for different business segments in a city or state.

Nonetheless, in our work, we have an extended study of about hundreds of attractions in different states of U.K and U.S. We show that in many attractions, the use of a single variable, history of visits, is not enough to have an accurate tourism prediction. We show that jointly adopting the history of visits, social media and environmental features can improve prediction results up to 94% in accuracy both in outdoor and indoor attractions.

2.2.2 Fine-grained Tourism Prediction

There are just a few prior studies on tourism prediction over finer granularities such as recreational sites, parks, galleries, hot-spots in the city, etc. These efforts, which we refer to as *fine-grained predictions*, aim at predicting visitation using either environmental features, such as monthly average temperature, amount of precipitation, humidity and cloudiness [Fisichelli et al., 2015; Hengyun Li and Li, 2016], **or** exploiting social media features such as user's check-ins, search engine trends [Höpken et al., 2018; Huang et al., 2017; Xiaoxuan et al., 2016] and geolocalized published photos [Ferrari et al., 2011; Spencer A. Wood and Lacayo, 2013; Khadivi and Ramakrishnan, 2016], but not both at the same time.

In this context, in [Spencer A. Wood and Lacayo, 2013] the authors use the locations of photographs in Flickr[3], a popular image hosting website, to estimate visitation counts in some recreational sites around the world. Their hypothesis is that pictures could indicate visitors, and furthermore, that photographs uploaded to an image-sharing website could record people's choices and provide useful data worldwide. They use information from the profiles of the photographers to derive travelers' origins in order to compare their estimations to empirical data. The authors show that visitation counts derived from field data and images are consistently scaled with a slope of 0.7, but the absolute visitation counts vary with local socioeconomic conditions and attributes of the site. The precision of predictions will hinge on the similarity of the studied sites in both geography and the types of attractions. In other words, they observe that absolute visitation rates are less variable across sites within nations or from a single destination type, such as a state park or an art gallery. As a final result, they suggest their method could serve more appropriately as a reliable proxy for empirical visitation and not exactly for accurate tourism prediction.

In [Fisichelli et al., 2015], the authors analyze the climate and visitation data for the U.S. national parks. They state that climate change will affect not only natural and cultural resources but also human visitation patterns. To demonstrate their hypothesis, they collect a dataset of historical monthly mean air temperature and visitation data from 1979 to 2013 at 340 parks and use it to project potential future visitation for years 2041 to 2060 based on two warming-climate scenarios and consequently two visitation-growth scenarios. After matching historical visitation data with historical monthly mean air temperature data, they argue that using a third-order polynomial temperature model explains 69% of the variation in historical visitation trends. They also argue that their model overestimates visitation at very high temperatures. Of the original 340 parks assessed, only 282 (83%) were temperature-sensitive, in that they showed strong relationships between visitation and temperature (adjusted $R^2 = 0.5$).

Another work that exploits environmental data [Hengyun Li and Li, 2016] links climate and seasonal tourism demand to study the effects of home climate, destination climate, and climate differences between destination and home on touristic demands. In their study, the authors consider different features of a destination including access to sea or lakes, availability of cultural and historical places, price, hospitality, accommodations, ease of access and cuisine. They show that home climate, destination climate, and climate difference count for most of the Hong Kongers' tourism demand considering 19 major tourism cities in mainland

[3] https://www.flickr.com/photos

China. In [Hengyun Li and Li, 2016], the authors stress that any research on the relationship between climate/weather and tourist behavior should not be based solely on how climatic dimensions affect tourists' comfort or well-being. Instead, the research should also consider that tourists might visit places with unfavorable or unpredictable weather conditions based on their interest in exploring places and landscapes they have not visited before [Steenjacobsen, 2001]. According to [Chang et al., 2006], novelty seeking is a key motivator for tourists when choosing a destination.

In [Höpken et al., 2018], the authors exploit travellers' web search behaviour as an additional input for predicting tourist arrivals. They analyze temporal relationships between search terms and tourist arrivals in a single attraction (a Swedish mountain) using the history of arrivals and Google web search. However, to avoid the inherent ambiguity of choosing query words, the authors define filters, considering only specific Google queries. Using ARIMA and ARIMAX as the prediction methods and Google Search Index (aggregated search query series by the most appropriate time lag) as the only external feature, they find that the inclusion of that feature leads to improvements in the prediction effectiveness.

In [Xiaoxuan et al., 2016], the authors propose a model with denoising (removal of noise) and forecasting by search engine using Hilbert-Huang Transform. They study (only) Jiuzhaigou National Park as their use-case. They report good MAPE results even with ARIMA (using only the history of visits) with further improvements due to denoising.

Another recent study [Volchek et al., 2018], investigates how the accuracy of tourism demand forecasting can be improved at the micro level using data from five London museums and different prediction methods such as SARIMA, SARIMAX and artificial neural network models [Iebeling Kaastra, 1996]. Their experiments are focused on evaluating different algorithms exploiting the Google Trends index as the main feature in order to predict the volume of arrivals to attractions. Google Trends index, however, is as a black-box whose internal mechanisms have not been revealed. Moreover, in that work, the authors experiment only with the top five most visited museums in London with free admission.

It is worth noting that in many of the above studies, their analysis is limited to a single or a few attractions; or very limited in terms of the type of attraction, location and exploited features. In contrast, in our work, by exploiting a rich set of features, including social media and environmental features data we analyze the result of multiple prediction methods in dozens of attractions in two categories of indoor and outdoor attractions. More importantly, we can achieve higher accuracy levels.

2.2.3 Seasonality Requirement

Several studies in the literature focus on the importance of seasonality as the main temporal aspect for tourism prediction. Seasonality has been defined as the inherently cyclic behaviour of tourism demands. Authors of [Hylleberg, 1992] states that seasonality is one of the main phenomena affecting tourism. According to them, seasonality is the systematic, although not necessarily regular, intra-year movement caused by changes in the weather, the calendar, and timing of decisions made by the agents of the economy, directly or indirectly through the production and consumption decisions. Authors of [Butler, 2001], instead, explain seasonality as a temporal imbalance in the phenomenon of tourism, which may be expressed in terms of dimensions of such elements as numbers of visitors, expenditure of visitors, traffic on highways and other forms of transportation, employment, and admissions to attractions.

Some researchers argue that tourism in most destinations is seasonal and tourism demand seasonality is caused by two basic components: climatic conditions and leisure time. 'Natural seasonality refers to the regular changes in climatic conditions at a particular destination during specific periods of the year [Koenig-Lewis and Bischoff, 2005]. They show that the climates of both the place of origin and destination influence the timing of travel, further determining the season length and quality of tourism in different regions. For example, certain types of tourism are highly affected by seasonality because climate is usually treated as an input in the creation of the tourism product, such as tourism for beaches, winter sports, and water sports[Martín, 2005], which we consider as outdoor attractions in our analysis.

Regarding periodicity, [Rosselló and Sansó, 2017] state that the main focus of interest had been annual seasonality, with studies that show the differences in tourism activity between different seasons. In contrast, in their work, they perform a decomposition analysis of yearly, monthly and weekly seasonalities of tourism demand. They do so by conducting an in-depth analysis of intra-monthly and intra-weekly tourism demands using entropy and relative redundancy measures. They show that seasonality is present in annual, monthly and weekly frequencies using the Balearic Islands airports as their case study. In addition, they show that monthly and weekly seasonality differs across geographical markets. Since variations during the year are often caused by the climate or other social factors, intra-monthly and intra-weekly changes in tourism demand should be more closely associated with institutional or social factors, due to non-working days during the week, work holidays and other events that take place at specific times, such as Christmas, school or university holidays and work vacations.

[Cuccia and Rizzo, 2011] focus their study on the impact of seasonality on cultural tourism – defined as tourism focused on cultural motivations, including visits to museums and archaeological sites. They analyze tourism seasonality in some selected destinations in Sicily, concluding that cultural destinations are less impacted by seasonality in tourism flows.

In a recent survey on tourism forecasting, Moro and Rita state that the most widely adopted statistical time series prediction method is the seasonal auto-regressive integrated moving average – SARIMA [Moro and Rita, 2016]. They claim that SARIMA is able to capture seasonality and recency implicitly in their forecasts.

2.2.4 Recency Requirement

A few prior studies analyze the effect of recency on tourism demands. In [Lim et al., 2018] and [Moro and Rita, 2016], the authors study temporal aspects considered as important to predict tourism visits. Particularly, in [Lim et al., 2018], an algorithm for recommending personalized tours is proposed using users' recent preferences as one of the variables of their model. In their tour recommendation algorithm, they enhance the models by a weighted update of user interests based on the recency of their visits giving more emphasis to more recent PoI [4] visits. They show improvements upon earlier tour recommendation work.

2.2.5 Model Specialization Requirement

Model specialization advocates creating specialized individual models for each touristic attraction. The main motivation is that particular attractions may have very specific intrinsic patterns of visitations. Studies such as [Richards, 2002] and [Leask et al., 2014] explore the tourists' motivations in the process of attraction selection. For instance, [Richards, 2002] identifies different motivations and behavioral patterns in visits to different types of museums. For example, tourists visiting the historic Rembrandt House are more likely to be accompanying other people, more likely to want to learn new things, as well as more likely to be in search of entertainment or local culture and history than those interviewed at the Stedelijk museum of modern art. They also find distance of the visitor from the origin, geographical origin of tourists, their socio-demographic characteristics, travel form and the period of staying in the destination are also important factors affecting the choice of attractions to visit. On the other hand, in [Leask et al., 2014], the authors study the generation Y[5] preferences, finding that this generation has his own profile and patterns of consumption. They discuss money spending preferences, the technology facilities in the attractions, the design of the place and the presence of information in social media as some of motivational

[4] Point of Interest (PoI): an entity of interest with well-defined location for example: museums, churches, waterfalls and coffee shops.

[5] Generation Y: the generation born in the 1980s and 1990s, comprising primarily the children of the baby boomers and typically perceived as increasingly familiar with digital and electronic technology.

differences. All in all, this serves as another factor that can affect differently the patterns of visitation in different touristic attractions motivating specialized models of visitation for each attraction.

2.2.6 Summary of Related Work

In this Section, we reviewed previous studies that provide motivation for the research goals proposed in this dissertation. In comparison to prior work, we perform fine-grained (i.e. attraction-level) prediction of visitation counts not for a single, but for a variety of attractions, exploiting both available social media and environmental data. To the best of our knowledge, we are one of few to perform such analysis by jointly exploiting both types of features to predict touristic demands. Our models are especially interesting to produce robust forecasting estimates for touristic places with low availability of visitation official census due to the high costs of surveys and the difficulty to access remote places. One of the novelties of our work is in comparing the effectiveness of using social media versus environmental features to predict visitation in different types of attractions, notably indoor and outdoor attractions.

Moreover, in our work we model and incorporate seasonality and recency in the prediction models while studying their impact on prediction performance, mainly for specialized prediction models for a set of attractions. Whereas in prior work, recency and seasonality features have been exploited indirectly, generally by using ARIMA-based models, in our work, we explicitly incorporate them as input features into our specialized models.

Table 2.2 summarizes the previous studies highlighting their contribution, whether they have been used environmental or any social media data and finally our contributions related to each work.

Table 2.2: Related work and our contributions

Related Work	Work Domain	Multiple Attr.	Ext. Data	Explicit Recency	Explicit Seasonality	Our work
[Cankurt and Subasi, 2015]	Predicting coarse-grained tourism demand for entire country Turkey using multiple socio-economic features	×	✓	×	×	use of environmental and social media features in a fine-grained prediction level
[Khadivi and Ramakrishnan, 2016]	Use of Wikipedia usage trends in order to forecast tourism demand of Hawaii reporting the accuracy of their prediction results only by RMSE	×	✓	×	×	use of environmental and social media features in more than 100 attractions
[Huang et al., 2017]	Analyze the relationship between the internet search data (Baidu in China) and the actual tourist flow only for a single city, Beijing Forbidden City	×	✓	×	×	Extended study of tens of attractions divided into two groups, studying the performance of different classes of features
[Li and Chen, 2009; Cho et al., 2011; Hasan et al., 2013; Hossain et al., 2016]	Use of Location Based Social Networks (LBSN) to study mobility of tourists and citizens in a coarse-grained fashion	×	✓	×	×	Fine-grained analysis level of information in social media networks
[Spencer A. Wood and Lacayo, 2013]	Use the locations of photographs in Flickr to estimate visitation counts in some recreational sites	✓	✓	×	×	Improving the accuracy of prediction models exploiting environmental features alongside explicit use of recency and seasonality factors
[Fisichelli et al., 2015]	Analyze the climate and visitation data for the U.S. national parks using a single model of third-order polynomial temperature model with an accuracy of 69%	✓	✓	×	×	Use of multiple prediction models exploiting social media features alongside explicit use of recency and seasonality factors

Related Work	Work Domain	Multiple Attr.	Ext. Data	Explicit Recency	Explicit Seasonality	Our work
[Höpken et al., 2018; Xiaoxuan et al., 2016]	Exploit travellers' Google web search and history of tourism arrivals to analyze temporal relationships between search terms and tourist arrivals in a single attraction (a Swedish mountain)	✓	✓	×	×	Extended study of tens of attractions exploiting environmental features alongside explicit use of recency and seasonality factors
[Volchek et al., 2018]	Tourism demand Prediction of top five most visited museums in London with free admission evaluating different algorithms exploiting the Google Trends index as the main feature	✓	✓	×	×	Compared to our work, the former is very limited in terms of the type of attraction, location and exploited features. while their main feature Google Trends index is a black-box with proved probability of overestimation problem
[Hylleberg, 1992]	Analyse seasonality as one of the main phenomena affecting tourism, i.e systematic, although not necessarily regular, intra-year movement caused by changes in the weather, the calendar, and timing of decisions made by the agents of the economy	×	×	×	✓	Extended study of tens of attractions exploiting external features of social media and environmental alongside explicit use of recency factor
[Butler, 2001]	Analyse seasonality as a temporal imbalance in the phenomenon of tourism, which may be expressed in terms of dimensions of such elements as numbers of visitors, expenditure of visitors, traffic on highways, employment, and admissions to attractions	×	×	×	✓	Extended study of tens of attractions exploiting external features of social media and environmental alongside explicit use of recency factor

Related Work	Work Domain	Multiple Attr.	Ext. Data	Explicit Recency	Explicit Seasonality	Our work
[Rosselló and Sansó, 2017]	Perform a decomposition analysis of yearly, monthly and weekly seasonalities of tourism demand showing that seasonality is present in annual, monthly and weekly frequencies using the Balearic Islands airports as their single case study	✓	×	×	✓	Extended study of tens of attractions exploiting external features of social media and environmental alongside explicit use of recency factor
[Cuccia and Rizzo, 2011]	Study the impact of seasonality on cultural tourism – defined as tourism focused on cultural motivations, including visits to museums and archaeological sites. They analyze tourism seasonality in some selected destinations in Sicily, concluding that cultural destinations are less impacted by seasonality in tourism flows	✓	×	×	✓	Extended study of tens of attractions exploiting external features of social media and environmental alongside explicit use of recency factor
[Moro and Rita, 2016]	Analyse and state that the most widely adopted statistical time series prediction method is the SARIMA which is able to capture seasonality and recency implicitly in forecasts	×	×	×	×	Analysis of multiple prediction models exploiting external features of social media and environmental alongside explicit use of recency factor
[Lim et al., 2018]	Propose an algorithm for recommending personalized tours based on users' recent preferences as one of the variables of their model enhancing their models by a weighted update of user interests based on the recency of their visits giving more emphasis to more recent visits	×	×	✓	×	Extended study of tens of attractions exploiting external features of social media and environmental alongside explicit use of seasonality factor

Chapter 3

Experimental Methodology

In this chapter, we present the methodology underlying the task of forecasting the visitation for fined-grained touristic points. We organize this chapter into five main sections, followed by a section with the chapter summary. First, in Section 3.1, we describe the datasets we collected as well as a set of basic statistics we can extract from them. Section 3.2 introduces the prediction architecture we propose to reach the main goal of this dissertation besides the parametrization of the learning models. Afterwards, Section 3.3 describes the metric to evaluate the accuracy of our predictions. Finally, in section 3.4 we discuss the factorial design configuration in order to compute the impact of each of tourism key requirements.

3.1 Data Collection

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 section, 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. We note that, for the sake of reproducibility, all our datasets are publicly available[1].

3.1.1 Dataset Description

Our data collection englobes information of outdoor and indoor touristic attractions. We used five different sources for our data collection, namely (1) the U.S. National Park Service,

[1] Available at Repository Mendeley - https://data.mendeley.com/datasets/t7bfhtzhxg/1

Table 3.1: Our FIne-grained, Structured and Enriched Tourism Dataset for Indoor and Outdoor attractions analysis (FISETIO)

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-processed data in 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 publicly available at Repository Mendeley DOI: 10.17632/t7bfhtzhxg.1 with Direct URL to data: https://data.mendeley.com/datasets/t7bfhtzhxg/1

(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. Table 3.1 shows the specification of these datasets. In the following we elaborate on the data sources for each of the attraction types.

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 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.

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, 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.

Division of attractions in outdoors and indoors. In [Westcott and Wendy Anderson, 2019], authors classify different types of tourist attractions throughout the world, whether natural or man-made and large or small. While [Stainton, 2021] enumerates main dominant categories of touristic attractions as natural, heritage(cultural), man-made and events. In another study [Li et al., 2017] authors analyse many factors impacting the tourist behavior in their visitations. However at the end, they highlight climate-related aspects as the most influential aspects on visitors behaviour. As a result, we distinguish our large attraction datasets based on their climate-related features which is whether they are indoors or outdoors. In the next section, we show that how the climate features such as temperature, humidity, precipitation, even sunny hours differs in this two principal category of attractions through their CCDF plots in Figures 3.1 and 3.2.

Nonetheless, we are aware of the possible cultural factor impact on the models since our attractions sets are in U.K. and U.S. and their visitation patterns could be impacted by English-people cultures. While, we may observe slightly different tourist behavior in other cultures such as Spanish, Portuguese, Latin, middle east and etc. This is an interesting point to study in a future work.

3.1.2 Dataset Overview

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 Figures 3.1 and 3.2. The y-axis shows the probability of the feature value exceeding the x-axis (P(X > x)). In Figure 3.1, 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 Figure 3.1(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 degrees.

Similarly, Figure 3.2 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 Figure 3.2(a), the CCDF-plot shows that for only 10% of the museums the total number of reviews was over 20 thousands.

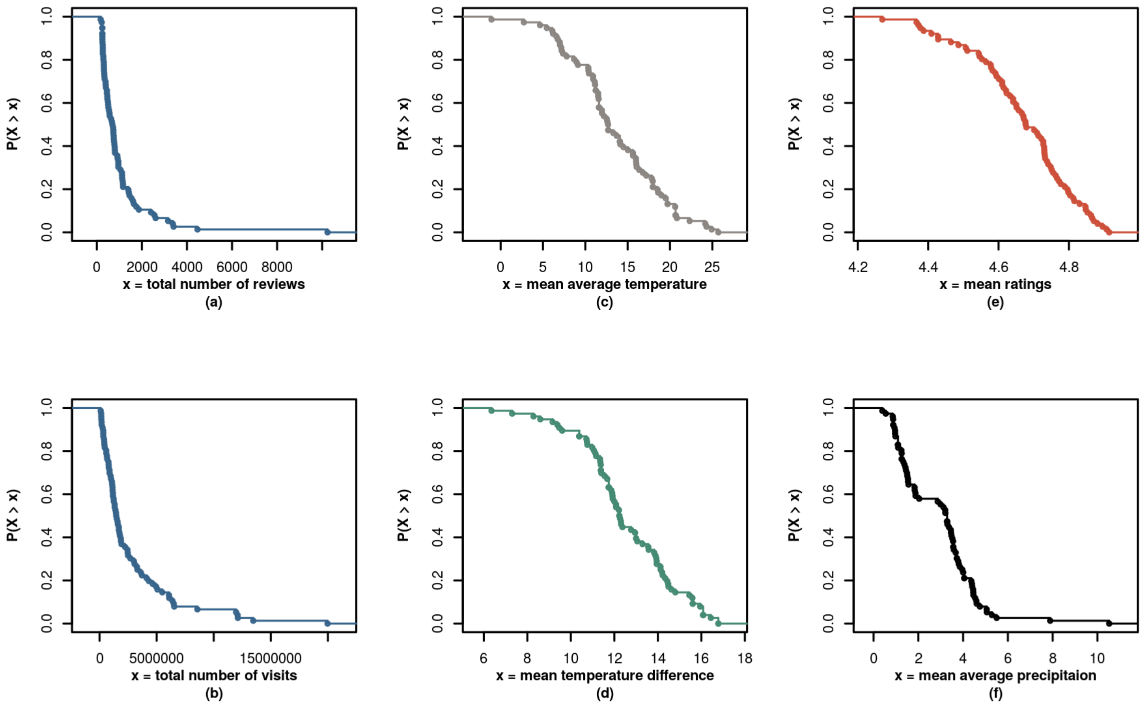


Figure 3.1: Complementary Cumulative Distribution Function (CCDF) plots for features in the outdoor dataset

3.1.3 Dataset Features Statistics

In this part, firstly, we present the basic statistics for each of the variables in the dataset. Next, we show the correlation analyses within variables in each dataset. Correlation analyses were performed to assess the strength of the relationships between pairs of variables extracted from our datasets. Different metrics can be employed to perform such analysis however

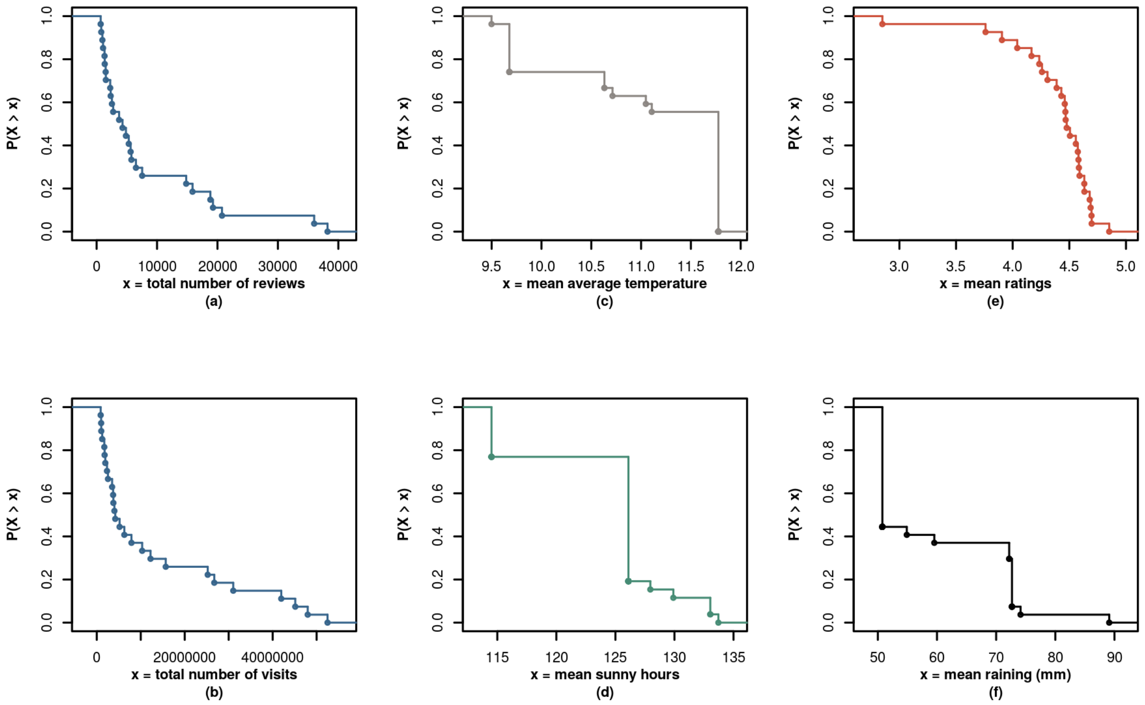


Figure 3.2: Complementary Cumulative Distribution Function (CCDF) plots for features in the indoor dataset

we used Pearson [Gautheir, 2001] correlation coefficient since it enables us to measure the linear dependency between the two input variables. In the following, Tables 3.2 and 3.3 present basic statistics for each category of attractions and corresponding variables/features.

Table 3.2: 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

Tables 3.4 (Parks) and 3.5 (Museums) provide the pearson correlation analysis of different features extracted from the environmental, social media and official datasets. For this analysis we calculate the correlation of features in batch, i.e. correlations are calculated along all the attractions and all historical points. Although some results are not that surprising – for instance, the positive correlations between temperature and number of visitations

Table 3.3: 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

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 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.

Table 3.4: 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 3.5: 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

3.2 Prediction Model Architecture and Parameter Tuning

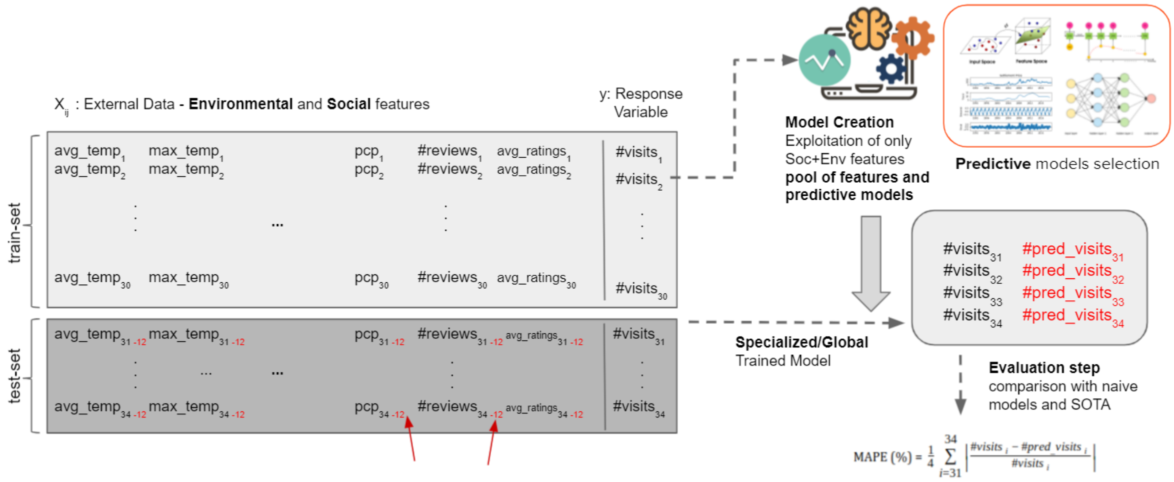

X_{ij} : External Data - **Environmental** and **Social** features
y: Response Variable
train-set
test-set
avg_temp$_1$ max_temp$_1$ pcp$_1$ #reviews$_1$ avg_ratings$_1$ #visits$_1$
avg_temp$_2$ max_temp$_2$ pcp$_2$ #reviews$_2$ avg_ratings$_2$ #visits$_2$
...
avg_temp$_{30}$ max_temp$_{30}$ pcp$_{30}$ #reviews$_{30}$ avg_ratings$_{30}$ #visits$_{30}$
avg_temp$_{31-12}$ max_temp$_{31-12}$ pcp$_{31-12}$ #reviews$_{31-12}$ avg_ratings$_{31-12}$ #visits$_{31}$
... ...
avg_temp$_{34-12}$ max_temp$_{34-12}$ pcp$_{34-12}$ #reviews$_{34-12}$ avg_ratings$_{34-12}$ #visits$_{34}$
Model Creation
Exploitation of only Soc+Env features
pool of features and predictive models
Predictive models selection
#visits$_{31}$ #pred_visits$_{31}$
#visits$_{32}$ #pred_visits$_{32}$
#visits$_{33}$ #pred_visits$_{33}$
#visits$_{34}$ #pred_visits$_{34}$
Specialized/Global
Trained Model
Evaluation step
comparison with naive models and SOTA
MAPE (%) = $\frac{1}{4}\sum_{i=31}^{34}\left|\frac{\#visits_i - \#pred\_visits_i}{\#visits_i}\right|$

Figure 3.3: Tourism demand prediction methodology adopting external data – social media and environmental features.

Figure 3.3 depicts how we split our dataset into training and test sets. It presents the external features – social media and environmental features – that we used in our prediction model and their relation with the number of visits occurred in each attraction (response variable). In our prediction architecture, since social media and environmental data may not be available at prediction time $X_{t=i}$, we exploit the values of the input feature in the last year ($X_{t=i-12}$) as the input of the models in the test case. This strategy has been used because of the annual seasonality behaviour of the tourism domain, as discussed in Section 2.2.

However, when we augment the key tourism requirements – recency and seasonality features – we update our architecture as in Figure 3.4, exploiting not only social and environmental features but also recent and seasonal features. Recency features consist of visit counts in the previous last 4 months (y-1, y-2, y-3, y-4) and their log values (log y-1, log y-2, log y-3, log y-4) while seasonality features are the number of visits in the last year, same period, i.e. (y-12, y-13, y-14, y-15) and their log values (log y-12, log y-13, log y-14, log y-15). Figure 3.4 presents the definition of each of features in the test-set. Again, for social media and environmental features, we exploit the values feature in the last year ($X_{t=i-12}$) while for recency features we have the value of visits in the last 4 months and for seasonality features the value of visits in the last year same period of the year.

We perform cross-validation to learn the prediction models as follows. For each attraction considered, we first divided each time series into two parts: the training set, consisting of the first x months of data (x = 30 for outdoor attractions and x = 76 in indoor attractions),

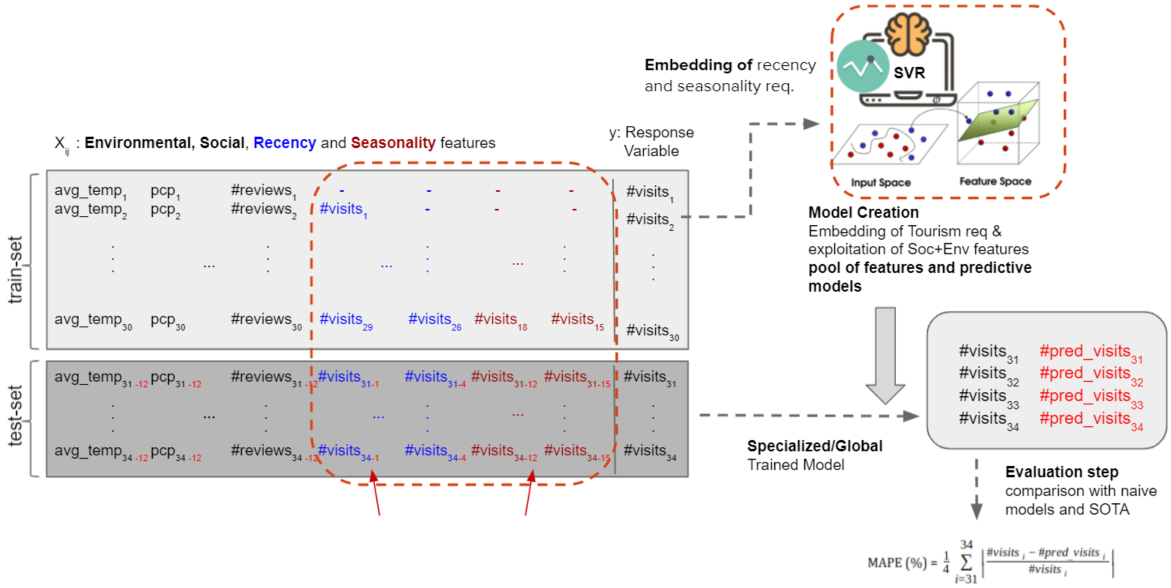

Embedding of recency and seasonality req.
SVR
Input Space
Feature Space
X_ij : Environmental, Social, Recency and Seasonality features
y: Response Variable
train-set
test-set
avg_temp_1 pcp_1 #reviews_1
avg_temp_2 pcp_2 #reviews_2 #visits_1
avg_temp_30 pcp_30 #reviews_30 #visits_29 #visits_26 #visits_18 #visits_15
#visits_1
#visits_2
#visits_30
avg_temp_31-12 pcp_31-12 #reviews_31-12 #visits_31-1 #visits_31-4 #visits_31-12 #visits_31-15 #visits_31
avg_temp_34-12 pcp_34-12 #reviews_34-12 #visits_34-1 #visits_34-4 #visits_34-12 #visits_34-15 #visits_34
Model Creation
Embedding of Tourism req & exploitation of Soc+Env features
pool of features and predictive models
#visits_31 #pred_visits_31
#visits_32 #pred_visits_32
#visits_33 #pred_visits_33
#visits_34 #pred_visits_34
Specialized/Global Trained Model
Evaluation step
comparison with naive models and SOTA
MAPE (%) = 1/4 Σ_{i=31}^{34} |(#visits_i − #pred_visits_i) / #visits_i|

Figure 3.4: Tourism demand prediction methodology adopting external data – social media and environmental features besides key tourism requirements – recency and seasonality features.

and the test set, consisting of the remaining months of data (4 months for both outdoor and indoor attractions). The training set is used to 'learn' the prediction model, while the test set is used for evaluating the learned model and reporting accuracy results. Furthermore, for those models that require parameter tuning, the training set is further split randomly into two parts: the first one, containing 30% of the training data, is used as validation set for parameter tuning. The second one is used for building the prediction function. Note that a specific model is learned (and later evaluated) for each attraction (for each park, museum or gallery), and thus, there is a different parameter choice for each of them. Parameter tuning for the models was performed as follows.

SVR. For SVR, we set the kernel function to "linear", because in preliminary experiments it produced the best results besides being more efficient (lower execution time). We varied the cost C parameter in the interval of $[2^{-5}, 2^{10}]$, and the best value varied for different attractions; however on average the best value was $C = 116$. The tolerance ε was tested in the range of $(0,1)$ with steps of 0.1 and we found 0.3 to be the best value of ε (again on average across all attractions).

SARIMA(X). Regarding SARIMA and SARIMAX models, we used the forecast package in R[2] in order to optimally find the best parameters (order of each polynomial) of the SARIMA model, as well as to find the seasonality pattern of the data.

LSTM. Related to LSTM, we have explored different network architectures, apply-

[2] Available at *https : //github.com/robjhyndman/forecast*

ing the ADAM optimizer [Nyamen Tato and Nkambou, 2018] for parameter optimization. ADAM is an adaptive learning rate optimization algorithm, designed specifically for training deep neural networks. Best results were obtained by: (i) normalizing all the variables in the range of (-1,1); (ii) using the mean-squared-error metric for the loss function; (iii) using a sequential model with one dense layer consisting 100 neurons using the Keras library in python[3]; and (iv) the following setting: number of epochs was set to 1000, dropout to 0.2 and batch size of 30.

The remaining techniques do not require manual tuning of parameters. Indeed, there are no tuning parameters for the linear regression method, while SARIMA and SARIMAX automatically determine the best order of each of their polynomials. Thus, the complete training set was used for deriving the models.

3.3 Evaluation Metrics

There are multiple metrics in the literature that have been used in the task of demand prediction, however we evaluate the accuracy of the prediction techniques by means of the Mean Absolute Percentage Error (MAPE) [Lewis, 1982]. MAPE has characteristics that facilitates comparison of results within different works since it is scale independent. In contrast, another metric Root Mean Square Error (RMSE) strongly depends on the data scale, which makes it inappropriate in comparison of results from different datasets or works. MAPE is a widely used measure of forecast error, being defined as:

$$MAPE(\%) = \frac{1}{M}\sum_{t=1}^{M} |\frac{y^{(t)} - \hat{y}^{(t)}}{y^{(t)}}| \tag{3.1}$$

where M is the number of forecasting periods, $y^{(t)}$ is the actual visitation count and $\hat{y}^{(t)}$ is the predicted visitation count, both for time window t. A lower MAPE(%) value indicates a smaller percentage of errors produced by the prediction model. One interpretation of MAPE(%) values was suggested by [Lewis, 1982] as follows: less than 10% is highly accurate forecasting, 10%-25% is good forecasting, 25%-50% is reasonable forecasting, and 50% or more is inaccurate forecasting.

[3] Keras is an open-source neural-network library written in Python.

3.4 Factorial Design definition of Tourism Key Requirements

Finally, we present the methodology applied to further investigate how recency, seasonality and model specialization requirements impact the prediction accuracy of different techniques, we perform a factorial design analysis over the correspondent features of each requirement to quantify the relative importance of each individual feature as well as their interactions on prediction accuracy. Factorial design techniques help to analyze the effect of each factor (requirement) as well as the effects of their interactions on the tourism demand (visits count) in each touristic attraction.

We employ the 2^k experimental design technique, since we are interested in determining the effect of k factors, each of which having two alternatives or levels. Such a design can be analyzed using a regression model to compute the main effect of a given factor x_i, subtract the average response of all experimental runs for which x_i was at its low (False) level from the average response of all experimental runs for which x_i was at its high (True) level[Jain, 1991]. The importance of a factor is measured by the proportion of the total variation in the response variable that is explained by this factor. In the following, we define the factors, their levels and the parameter space in our analysis. Moreover, the regression analysis is applied separately for each attraction class (indoors and outdoors).

Specifically, we employ a 2^k factorial design with $k = 3$ factors (i.e.,recency, seasonality and model specialization), each one with two levels (true or false). This design allows us to estimate the relative importance of each factor as well as all factor interactions on the response variable. This importance is estimated by the fraction of the total variation observed in the response that can be explained by each factor (or factor interactions). In the following, we define the considered factors and factor levels. Note that we perform a 2^k factorial analysis for each type of attraction (indoors and outdoors):

- Recency factor ($\mathscr{R}$): two levels of (1) True, if we use the visit counts in the previous last 4 months (y-1, y-2, y-3, y-4) and their log values (log y-1, log y-2, log y-3, log y-4) for training the model and; (2) False, otherwise.

- Seasonality factor ($\mathscr{S}$): two levels of (1) True, if we use the visit counts in last year (y-12, y-13, y-14, y-15) and their log values (log y-12, log y-13, log y-14, log y-15) for training the model and; (2) False, otherwise.

- Model Specialization factor ($\mathscr{M}$): two levels of (1) True, if we train an individual model for each indoor/outdoor venue and; (2) False, if we learn a unique model for all venues of each attraction class.

Defining the above factors and levels, we would have a 2^k factorial design for k = 3, results in 2^3 factorial design with a parameter space of $RxSxM = 2x2x2 = 8$ combinations.

3.5 Summary

In this chapter, we described the methodology we adopted to collect and process touristic attractions visitations in two categories of indoors and outdoors. We can summarize this process as follows:

- We collected a long-term dataset divided into two parts: outdoor and indoor attractions; we presented a complete analysis of our publicly available datasets - one of our main contributions in this dissertation. The resulting dataset is used in the Chapter 4 to analyse and predict the tourism demand in outdoor attractions (National Parks in U.S.) and in Chapter 5 for indoor attractions (Museums and Galleries in U.K.).

- We proposed an architecture to exploit not only external data - social media and environmental features in our prediction models but also tourism prediction requirements - recency, seasonality and model specialization.

- We also showed specific aspects of our methodology including the learning and parameterization of the prediction models as well as the evaluation metrics used in our study.

Chapter 4

Experimental Results - OUTDOOR Attractions

In this chapter, we present set of results to address our two research goals in the scenario of outdoor attractions, corresponding to more than 70 National Parks in the U.S. We start with correlation analysis of different features which are extracted from the environmental, social media and official datasets in Section 4.1. Then, we discuss prediction results using different combinations of external features, namely social media and environmental data in outdoor attractions that fulfill our first research goal (**RQ1**) in Section 4.2. For that analysis, we adopt the tourism demand prediction architecture which was presented in the previous Chapter (see Figure 3.3). We also identify the cases in which the combination of external features resulted in inaccurate predictions due to anomalies in the visitations in those attractions.

Next, in Section 4.3, we elaborate the prediction power of explicitly exploiting the three key requirements of tourism prediction as features in our models addressing our second research goal (**RQ2**). That study exploits the second tourism demand prediction architecture which was presented in the previous Chapter (see Figure 3.4). We also evaluate the prediction power of each tourism requirement in a factorial design analysis, for the cases with scarcity in the historical data and the cases where recency and seasonality features could improve the prediction of inaccurate cases. Finally, Section 4.4 studies the possibility of creating group of similar attractions in order to build a single prediction model for each group useful in scenarios where we have little information about some attractions. The summary of the main results and the conclusions are presented in Section 4.5.

We have also made all the datasets and codes publicly available that guarantees the reproducibility of the results[1].

[1] Available at Repository Mendeley - https://data.mendeley.com/datasets/t7bfhtzhxg/1

4.1 Correlation Results

This section aims to motivate the prediction of touristic demands exploiting both social media and environmental data by presenting correlation results of various features with the official number of visits for outdoor attraction. A high correlation means that two or more variables have a strong relationship with each other, while a weak correlation means that the variables are hardly related.

Table 4.1 presents correlations of each of the features with number of visits (#Visits), which were computed in batch, i.e. correlations are calculated across all attractions and all historical points in the outdoors dataset. The feature with the strongest (positive) correlation with #Visits is #Reviews (above 0.32). However, the correlations of most of the other features with #Visits are not negligible. Although such correlations are not the only aspect that influences their prediction power[2], they are an indicator that the considered features may bring some information to the prediction task. Additionally, considering the temperature features (Min, average and Max temperature), the correlation results with #visits are high. This shows the importance of temperature features in the case of outdoor attractions.

Table 4.1: Overall correlations of #Visits with other features for outdoor attractions. Strongest correlation is in bold.

Feature	Correlation with #Visits (Outdoor Attractions)	Feature Type
Min_Temp	0.26	Environmental
Avg_Temp	0.26	Environmental
Max_Temp	0.25	Environmental
Precipitation/Rainfall	0.005	Environmental
#Reviews	**0.32**	Social Media
Avg_Rating	-0.06	Social Media

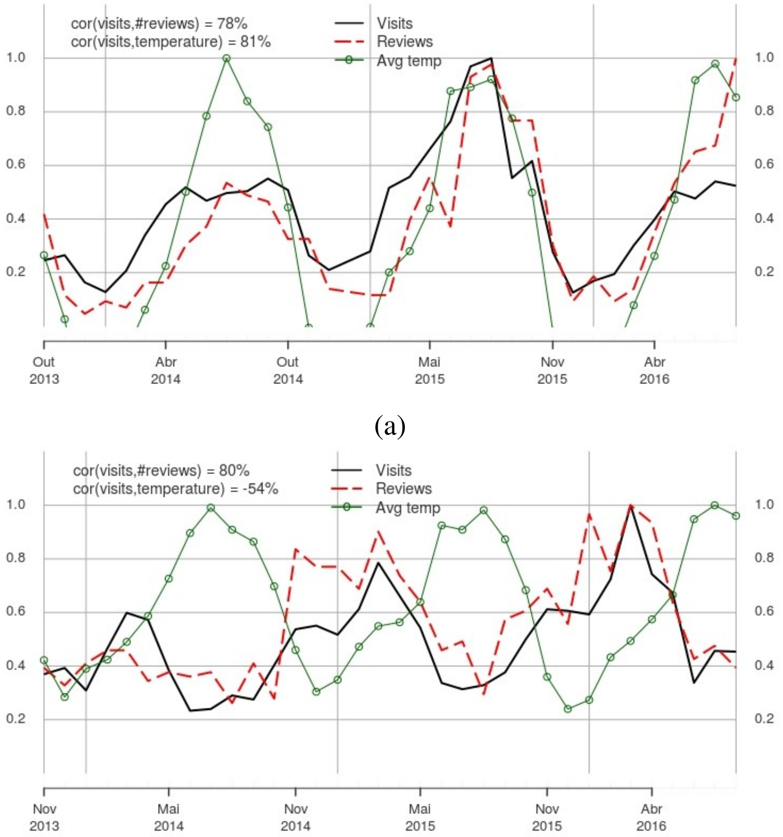
We further analyze the highest correlated feature with #Visits, namely #Reviews, for each type of attraction.We computed the correlation between monthly number of visits and monthly number of TripAdvisor reviews for each park separately, covering the period of September 2013 till September 2016. All correlation values were positive, but varied greatly across attractions. Thus, we grouped the parks into 3 categories: A (strong), B (moderate) and C (weak) based on the correlation between #Visits and #Reviews observed for the park. Table 4.2 (third column) shows percentage of parks in each category. Note that for 81% of the analyzed parks the correlation is at least moderate (strong for 60% of the parks).

In order to illustrate the observed correlation values, Figure 4.1 presents the time series of the number of TripAdvisor reviews, visitation counts, and average temperature

[2] Indeed we cannot claim any causality effect, but rather only some relationship between the variables.

Table 4.2: Distribution of attractions across three categories based on correlations between #Visits and #Reviews.

Category	Correlation (#Visits and #Reviews)	Outdoor Attractions
A	over 65% (strong correlation)	60% (45 parks)
B	50% to 65% (moderate correlation)	21% (16 parks)
C	less than 50% (weak correlation)	19% (15 parks)



(a)

(b)

Figure 4.1: Two examples of parks with strong positive correlations between number of reviews and official visits (category A in Table 4.2) with (a) strong positive correlation (81%) between average temperature and number of visits in Colorado National Park, and (b) moderate negative correlation (-51%) between average temperature and number of visits in Joshua Tree National Park (y-axes show normalized values of each time series in the scale of 0 to 1).

for two national parks in the dominant category (category A), with strong correlations between #Reviews and #Visits. Each time series is normalized by its maximum value in the time period analyzed. Focusing on Figure 4.1-a, note that all three time series are strongly positively correlated. In contrast, Figure 4.1-b illustrates the case of a park for which average temperature is negatively correlated with the other two time series. These results suggest great diversity in how social media and environmental features relate to visitation counts across different attractions.

Additionally, in order to investigate the performance of each of climate and social media features in the whole dataset, we used Random Forests [Breiman, 2001] to calculate the relative importance of each feature in predicting the variation in visitation. Figure 4.2 reveals the influence of each feature by percentage within the complete feature-set. As we can see, the social media feature - number of reviews - has the highest relative influence alongside the outdoor dataset. Then comes the environmental features - minimum and maximum temperature and precipitation. The average rating and average temperature appear in the end.

Finally, we highlight that social media features indeed have high predictive power in our outdoor dataset. However, it is also interesting to see that Min_temp comes in a close second place, which shows the complementarity of both types of features - social and environmental.

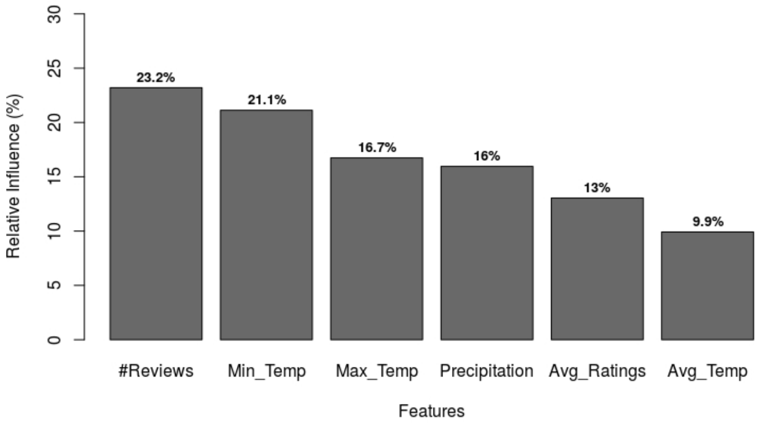


Figure 4.2: Features relative influence in percentage (%) - outdoor dataset

4.2 External Data Performance in Visits Prediction

We now present results of eight analyzed prediction methods, namely, Linear Regression (LR), SVR, GRNN, SARIMA, SARIMAX, LSTM and two naive models - naive recency and naive seasonality (all these methods were discussed in Chapter 2.1), when applied to our outdoor dataset. We compare the effectiveness of different combinations of external data – environmental (average, minimum and maximum temperatures, precipitation) and social media (#Reviews and average rating) features as predictor variables when exploited tourism demand prediction architecture as in Figure 3.3.

For the SARIMA and SARIMAX models, we also exploit the history of visit counts (#Visits) as input feature. We restrict our evaluation to a time period during which all parks have entries in the three datasets considered, namely environmental, social media and available official visits. This period corresponds to 34 months (November 2013 to August 2016): the initial 30 months were used to learn the models and then predict the next 4 months of visitations.

Table 4.3 reports MAPE results for each prediction model, when exploiting all available features. The table shows the percentages of parks for which the MAPE falls within each of the following ranges: high accuracy (MAPE $<$ 10%), good accuracy (10 to 25%) and low accuracy ($>$ 25%). In this scenario, we have the specialized SVR model with environmental and social features as the best model, predicting accurately for the highest percentage of attractions (almost 95% of Parks), considerably outperforming other models.

Best results for MAPE less than 10% (highly accurate results) are achieved by the naive-seasonality (42%) for the outdoor attractions. The success of the naive methods in highly accurate results (MAPE < 10%) is one of the reasons that motivates the adoption of seasonality in our models that we will investigate in the next section when explicitly incorporated as features into our best model (SVR). We can relate the success of naive seasonality in parks with the seasonal-cyclic behavior of climate in the outdoor attractions, as seasonality has been considered one of the main phenomena affecting tourism, principally due to changes in the weather conditions [Hylleberg, 1992].

Table 4.3: Percentages of outdoor attractions (parks) that fall into different ranges of MAPE value for each prediction technique. Bold values show prediction techniques with higher percentages of parks with good-to-high prediction results.

MAPE	SVR	SARIMAX	LR	GRNN	SARIMA	LSTM	n.recency	n.seasonality
lower 10%	22.4%	13.2%	15.8%	1.3%	7.9%	23.6%	6.58%	**42.11%**
10% to 25%	**72.4%**	55.26%	53,95%	25,00%	47.37%	60.6%	50.00%	40.78%
over 25%	5.26%	31.58%	30.26%	73.68%	44.74%	15.8%	43.42%	17.11%

For the sake of completeness and validation of the LR models (one linear regression model per each outdoor attraction), we also report the R-squared values (goodness of fitted models), R-Squared-Adjusted and F-Statistics. We had R-squared values in the interval of [0.23, 0.97] with the average value of 0.73; R-Squared-Adjusted [Clark et al., 1974] in the interval of [0.03, 0.96] having 0.65 as the mean value; F-Statistics within [1.16, 111.6] having average of 20.11. For 90% of the attractions, the p-value was $<$ 0.05 (95% of confidence)[3].

[3] The p-value for a model determines the significance of the model compared with a null model. For a

We compute the p-value to answer the following question: does this model explain the data significantly better than would just looking at the average value of the dependent variable? p-value less than < 0.05 means that the LR models in which are predicting the #Visits are significantly better than average value of #Visits at least for 90% of the attraction [Clark et al., 1974]. Table 4.4 shows statistical analysis of linear regression models.

Table 4.4: Statistical analysis of R-Squared, R-Squared Adjusted, F-statistics and P-value for outdoors linear Regression models

Metric	Min	Average	Max
R-Squared	0.23	0.73	0.97
R-Squared-Adjusted	0.03	0.65	0.96
F-Statistics	1.16	20.11	111.6
% of attractions with p-values < 0.05		90%	

4.2.1 Feature analysis

Once we provide the overall performance when both environmental and social features are used in our prediction task, we now turn to investigating how environmental and social media data attributes impact the prediction accuracy of different techniques. Here and in the following, we refer to the complete set of features as F and we compare the performance of different prediction techniques using F as input as well as using a single feature (or subset of features) or all but one feature f (referred to as $F - f$). In such discussion we focus on the attractions for which the predictions had good to high accuracy, that is, MAPE $< 25\%$, and take as a metric of comparison the percentage of parks for which the prediction fell within this range of MAPE value.

Table 4.5 shows the prediction performance of each technique using different sets of features, namely all features (F) and single features. Recall that the SARIMA technique likewise the naive models - naive recency and naive seasonality exploit only the history of visit counts and the SARIMAX model, which is a boosted version of SARIMA, also adds exogenous features to improve its results. We should stress that the SARIMAX results shown in Table 4.5 are always boosted by incorporating history of visits as a feature. This may cause an unfair comparison of the performance of the individual features within prediction methods.

As shown in Table 4.5, the best results with the complete set of features are obtained when we employ the SVR prediction model, while the second best prediction model is linear

linear model, the null model is defined as the dependent variable being equal to its mean.

Table 4.5: Prediction results (% of parks with MAPE $<$ 25%) for outdoor attractions: all features versus single features. For each prediction technique, individual features with best prediction results are marked in bold.

Features	SVR	SARIMAX	LR	GRNN	SARIMA	n.recency	n.seasonality
All Features (F)	94.74%	68.42%	69.73%	26.31%	-	-	-
Min_temp	**85.49%**	71.05%	67.10%	27.63%	-	-	-
Avg_temp	81.57%	**72.36%**	65.78%	**30.26%**	-	-	-
Max_temp	80.26%	69.73%	**68.42%**	**30.26%**	-	-	-
#Reviews	75%	61.84%	65.78%	26.31%	-	-	-
Precipitation	36.84%	52.63%	34.21%	25%	-	-	-
Avg_Ratings	31.57%	52.63%	31.57%	**30.26%**	-	-	-
History #Visits	-	-	-	-	**55.26%**	**56.58%**	**82.89%**

regression. SARIMAX wins the third place. Finally, SARIMA, using only history of visits, and GRNN have the weakest prediction results. Once again, we find that GRNN loses in performance, compared to all other techniques, by a very large gap. Note that the best result of SARIMAX is obtained when using only the average temperature as an exogenous feature (as opposed to all other features). Moreover, SVR with any of the temperature features or even with #Reviews as single input provides better results than any other technique. In any case, the best overall result is produced when all features are used, corroborating the importance of jointly employing social media and environmental data for enhancing the tourism prediction accuracy.

We delve further into the performance of the best approach, SVR, by showing, in Figure 4.3, how its performance improves as we introduce new features, starting with the worst individual feature (average ratings) and adding the others in increasing other of their individual effectiveness. We note that features Avg_Rating and Precipitation, in isolation, do not present good prediction performances (lower than 40% accuracy), which is consistent with the low correlation results of these features with the number of visits. The inclusion of the social media feature #Reviews substantially increases the prediction accuracy. Further improvements are obtained when adding temperature features.

4.2.2 Feature ablation analysis

Another interesting analysis of the predictive power of the features, for different prediction techniques, is presented in Table 4.6. For each feature (or feature subset) f, the table shows performance results of each technique considering the set containing all prediction features but f, that is, $F - f$. Once again, we use as a performance metric the percentage of parks for

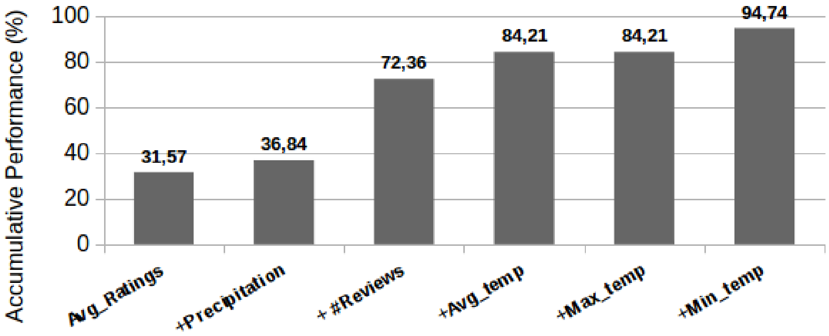


Figure 4.3: Performance of SVR on outdoor attractions (% of parks with MAPE $<$ 25%) as we introduce new features.

which prediction achieved MAPE $<$ 25%. For instance, row (F - *Environmental features*) shows the performance of the models when all the environmental features (i.e., minimum, average and maximum temperatures as well as precipitation) are disregarded. Similarly, row (F - *Social media features*) presents the results when the two considered social media features, namely number of reviews and average rating, are removed from the input feature set. Note that Table 4.6 does not show results for SARIMA, as this technique considers only the history of visit counts as input feature set.

Table 4.6 shows that Precipitation and average ratings are the weakest features in the feature set. Their removal from the feature set either did not cause any impact (e.g., for SVR) or actually helped prediction performance (other models). Interestingly, fixing a prediction model (e.g., SVR) and disregarding the social media features, the use of all temperature features alone (i.e., feature set F - *Social media features*) results in lower performance than having only one of the temperature features (see for instance, Table 4.5: 85.49% for only Min_temp). This may be due to some noise and inconsistencies that arise when all features are employed together. On the other hand, eliminating a single temperature feature from the complete set (i.e., keeping the social media features) does not change the prediction quality. This effect, which holds for SVR, SARIMAX and Linear Regression, may be explained by the fact that the temperature features are highly correlated among themselves (see discussion below). Thus, the removal of one of them is compensated by the others. However it is worth mentioning the importance of at least one environmental feature, especially if used as input to SVR, as the removal of the complete feature subset (F - *Environmental features*) causes great drop in performance (even greater than if the social media features are disregarded).

Furthermore, it can be inferred from this analysis that #Reviews is the most contributing individual feature to F, as removing it causes the largest drop in performance (for all models but GRNN, for which minimum and average temperatures cause somewhat greater performance reduction). This observation is consistent with the correlation results previously reported.

Table 4.6: Prediction Results (% of parks with MAPE $<$ 25%) for outdoor attractions: all features versus all but one feature (or feature subset). Values in bold show the most contributing individual feature for each prediction technique.

Features	SVR	SARIMAX	Linear Regression	GRNN
All Features (F)	94.74%	68.42%	69.73%	26.31%
F - Environmental Features	76.31%	65.78%	65.78%	27.63%
F - Social Media Features	81.57%	67.10%	61.84%	25%
F - Min_temp	94.74%	71.05%	72.36%	23.68%
F - Avg_temp	94.74%	72.36%	72.36%	**23.68%**
F - Max_temp	94.74%	71.05%	72.36%	26.31%
F - Precipitation	94.74%	72.36%	68.42%	28.94%
F - #Reviews	**88.16%**	**67.10%**	**60.52%**	26.31%
F - Avg_Ratings	94.74%	72.36%	72.36%	27.63%

4.2.3 Factorial design of the best features

We perform a factorial design analysis over the best features to quantify relative importance of each individual feature as well as their interactions on prediction accuracy. Considering the results in Table 4.5, we concluded that Avg_temperature, Min_temperature, Max_temperature and #Reviews stand out as the most promising individual features. However, the three temperature features are strongly correlated among themselves for all parks, as one might expect. Indeed, the Pearson correlation coefficients between Avg_temperature and Max_temperature, Avg_temperature and Min_temperature as well as Min_temperature and Max_temperature, computed across all attractions are 0.99, 0.99 and 0.98. Thus, choosing only one should be enough to capture all the effect of temperature on visitation count. We chose Min_Temperature, as the resulting percentage of parks with MAPE $<$ 25% is marginally better than if Avg_temperature or Max_temperature is used (look at Table 4.5).

Thus our present analysis focuses on the effect of Min_Temperature and #Reviews on visitation count (#Visits) in each given park. To that end, we employ an ANOVA test [Fisher, 1921], which is a statistical procedure for analyzing the amount of variance that is contributed to a sample by different factors. The ANOVA test was applied separately for each park. It takes as input the time series of the two factors (Min_Temperature and #Reviews) as well as the time series of the response variable (#Visits), and outputs the percentage of the total variation observed in the response variable that can be explained by each individual factor as well as by their interaction. We chose to employ an ANOVA type II test, which is recommended in case of unbalanced designs [Herr, 1986] (an unbalanced design has an unequal number of observations for level combinations), as is ours.

The results of the ANOVA test varied greatly across parks, reflecting the great variations in the correlations between the factors and the response variable. For presentation purposes, we consider once again the categories of parks by strength of the correlation between #Reviews and #Visits (categories A, B and C in Table 4.2), grouping ANOVA results for parks in each category. Table 4.7 shows these results, presenting, for each park category, the mean and maximum contribution of each individual factor as well as of their interaction to #Visits. It also shows, for individual factors and interaction, the percentage of parks (in each category) for which its contribution was statistically significant (p-value < 0.01).

According to Table 4.7, for parks in category A, which account for most (60%) parks in the outdoor dataset, factor #Reviews explains on average 23% of the variation in #Visits. Yet, its contribution can be as high as 76%. Moreover, such contribution is statistically significant for 64% of the parks in this category. Min_Temperature explains, on average, 30% of the variation (reaching up to 67%), and such contribution is statistically significant for 82% of the parks, having thus a stronger effect on #Visits. Similar results are observed for parks in categories B and C, but with a weaker importance of #Reviews, as expected given the definition of the categories. In all cases, the contribution of the interaction of the two factors is considerably smaller than that of the individual factors.

Overall, we find that, for some parks (especially those in category C), the relative importance of #Reviews and Min_Temperature is somewhat small, suggesting that, for such parks, other features (not captured in our dataset) may have more predictive power. However, for most parks in our dataset (60%), by using only the two selected factors – number of reviews and minimum temperature – along with their interaction, one is able to capture and explain most variation (60%) in the number of visits. This result demonstrates that social media and environmental features are complementary and that #Reviews and Min_Temp (and their interactions) are the two most important features in the considered set, corroborating our previous analyses.

4.2.4 Anomaly cases - inaccurate predictions

In this section, we inspect attractions (national parks) for which our best prediction technique – SVR using all features – did not result in good predictions (i.e., MAPE $> 25\%$). We identify and discuss the possible reasons and the effects that caused our proposed prediction methodology to obtain such unsatisfactory results.

Table 4.7: Contribution of #Reviews and Min_Temperature (and their interaction) to explain variation in #Visits in outdoor attractions (Results of ANOVA type II analysis. Categories refer to Table 4.2. Columns p-value include percentage of attractions, in each category, for which the effect of each factor/interaction is statistically significant.).

Factor	Category A (60% of parks)			Category B (21% of parks)			Category C (19% of parks)		
	p-value (< 0.01)	Mean contrib.	Max contrib.	p-value (< 0.01)	Mean contrib.	Max contrib.	p-value (< 0.01)	Mean contrib.	Max contrib.
#Reviews	64%	23%	76%	38%	13%	35%	7%	5%	16%
Min_temp	82%	30%	67%	56%	27%	85%	60%	25%	68%
(#Reviews*Min_-temp)	22%	7%	27%	6%	5%	13%	0%	2%	8%

We identify that for 4 out of 76 U.S national parks, the SVR results are not good (MAPE $>$ 25%). We thoroughly explored the design space in terms of parameter values and yet the results were always unsatisfactory. We then searched for external reasons that could explain such poor predictions. We checked the official website of the U.S National Park while looking in different blogs, history records and websites like Wikipedia about each of the poorly predicted parks. Our investigation uncovered some possible explanations for the poor prediction performance on each of these parks. Overall we find anomalous patterns in either the number of monthly visits or the number of monthly reviews, which are not followed by the other variable.

Table 4.8 lists the four parks identified, and provides, for each of them, the best MAPE result obtained with SVR, the park category (in terms of correlation between #Reviews and #Visits), the identified anomaly in the number of monthly visits/reviews, the corresponding effect (if any) on the number of monthly reviews/visits as well as a possible reason behind these out-of-pattern behaviors. For instance, the prediction accuracy for the *Sunset Crater Volcano National Monument* was very poor (best MAPE equal to 184.81%). This attraction falls into category C, that is, the correlation between #Reviews and #Visits is weak (less than 50%). Based on our investigation, we learned that there was a decrease of around 10% in the number of visits at the end of 2015 and early 2016 and a larger decrease from April to September 2016. These patterns are not matched by the visitations observed in prior years. Yet, no noticeable difference occurred in the number of monthly reviews on TripAdvisor. We conjecture that a possible reason for the observed decrease in number of visits is a report of steam rising from the crater, which generated an eruption scare, but the steam was later determined to be related to a forest fire.

Table 4.8: Description of cases of inaccurate predictions on outdoor attractions.

Park Name	MAPE (%)	Cat	Anomaly	Effect on #Reviews	Possible Reason
Sunset Crater Volcano National Monument	184.81%	C	Decrease in number of visits (about 10%) from November 2015 to January 2016 in comparison with the same period in 2014 and 2015 and a larger decrease (over 60%) in number of visits in the period of April to September 2016, compared to the previous years	Same behavior as previous years	Report of **steam rising from the crater and Eruption scare** on June 5, 2015; however, the cause of the steam was determined to be a forest fire later on.
Bryce Canyon National Park	35.19%	C	Considerable increase in number of visits (more than 20%) starting in February 2016 till September 2016 in comparison with the same period in 2015	Same behavior as previous years plus a slightly decrease in May 2016 compared to May 2015	**Waiving entrance fees** for a total of fifteen days in 2016 as a way to encourage people to get outdoors and enjoy.
Hovenweep National Monument	26.53%	C	Increase in number of visits (about 15%) between May and October 2015 in comparison with the same period of 2014 and a bigger increase (over 30%) in the same period of 2016	Roughly same behavior as previous years	In July 2014, the International Dark-Sky Association **designated Hovenweep an International Dark Sky Park**.
Denali National Park and Preserve	36.48%	B	in this case anomaly was in #Reviews	Huge increase (over 50%) in number of monthly reviews in the period of May to September 2016 with the peak in August 2016 compared to the same period in year 2015	Aug. 30, 2015, when President **Barack Obama** directed Secretary of the Interior Sally Jewell to **rename the mountain to Denali**.

4.3 Augmentation of Tourism Prediction Requirements

We showed in Section 4.2 that applying external data – social and environmental features – greatly improve the accuracy of our prediction task. In this section, we go further demonstrating that, by adding the three proposed key requirements of tourism prediction as explicit features in our models, we can greatly improve prediction accuracy of the results presented in the previous section. Consequently, in Section 4.3.1 we discuss the model specialization, one of the main tourism prediction requirement while in Section 4.3.2 we present the analysis of other two requirements, i.e. seasonality and recency. For experiments in this section, we adopt the second tourism demand prediction architecture which was presented in the previous Chapter (see Figure 3.4).

4.3.1 Model Specialization

We provide evidence of the importance of considering model specialization as an explicit requirement for tourism prediction in two types of Specialized and Global prediction models. Training specialized models for each individual attraction allows the models to learn specific patterns of visitation at each touristic point. However, the application of model specialization may be considerably jeopardized when we do not have enough data to train individual models for each site. In this case, it is more viable to train and apply a single global model taking advantage of the complete social and environmental (training) data for multiple attractions.

Specialized vs. Global Prediction Models. We compare the prediction accuracy of the specialized models, reported in the previous Section, with that of a global model trained with all attractions of outdoor. Figure 4.4 illustrates the construction of specialized and global models. In specialized models, we train a model particularly with features of each attraction while for global models, the model receives feature observations of all attractions building a model. In our comparison, we employ the SVR method as it produced the best results among the tested methods. The experimental setup is similar, with the difference we exploit the second tourism demand prediction architecture which was presented in previous Chapter (see Figure 3.4).

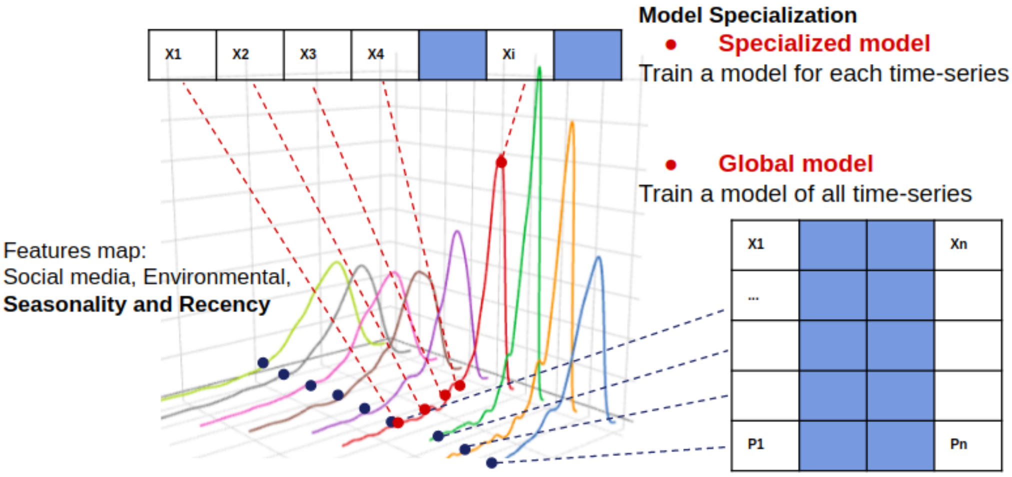


Figure 4.4: Model specialization: Specialized and Global models

Results are shown in Table 4.9. We observed that in the case of outdoor attractions, the global model has a good MAPE (MAPE $< 25\%$) only for 18% of parks while specialized models have a much better performance with over 94% of parks having good prediction accuracy.

To further illustrate our argument that individual models are more adequate to compare idiosyncratic aspects of each attraction, Table 4.10 presents the learned coefficients for the specialized and global models for two attractions - National Park (NP) in Aztec Ruins and in Big Cypress, for which good results were obtained with the specialized models. As we can see in this Table, the relative importance of features are different for these two attractions, despite the good results of their respective individual models – both with good MAPE

Table 4.9: Percentages of outdoor attractions that fall into different ranges of MAPE value for specialized and global models using the SVR model.

MAPE	SVR Global Model	SVR Specialized model
lower 10%	5.26%	22.4%
10% to 25%	13.16%	72.4%
over 25%	89.47%	5.26%

values (MAPE $<$ 25). For instance, in the specialized model for NP-Aztec Ruins, the environmental feature (Max_Temp) has a higher value than the social media feature (number of Reviews) while the opposite is true in case of NP-Big Cypress. On the other hand, the global model, which tries to capture an "average" behavior for all national parks, fails to return accurate predictions for NP-Aztec Ruins. Interestingly, even with the global model, the relative importance of the features, as captured by the respective coefficients for each feature, is consistent with our previous discussions.

Table 4.10: Comparison of coefficients within global and specialized models with 95% of confidence; NP: National Park, **Spec: Specialized model, Glo: Global Model**. The coefficient with highest value in each model is in bold.

Model	Min_temp	Avg_temp	Max_temp	Precipitation	#Reviews	Avg_Ratings	**MAPE_Glo**	**MAPE_Spec**
Spec - NP Aztec Ruins	0.18	0.17	**1.27**	0.007	1.08	0.01	818.2	11.8
Spec - NP Big Cypress	1.96	0.11	0.06	0.04	**2.09**	0.05	20.64	13.4
Glo - all Parks	-0.7	**1.74**	-0.99	0.01	0.25	-0.007	-	-

4.3.2 Recency and Seasonality

We now shift our attention to demonstrate how the addition of the other two tourism requirements, i.e. recency and seasonality, into the most accurate prediction models in the previous section (SVR models with Social and Environmental features) can improve the accuracy of forecasting the visitations for fine-grained touristic points. In the next, we present this analysis first for global models and then for the specialized models separately. It is worth noting that the reason behind this separation is isolating the effect of model specialization in the study of recency and seasonality features since there may be influence between these requirements.

Global model (Model specialization = OFF). Here, we show the predictivity power of trained Global model with Environmental and Social data (hereafter called GloES) augmented with seasonality and recency tourism features for outdoors tourism attraction. Table 4.11 show the results. In the case of outdoor attractions, the GloES has a good MAPE (MAPE < 25%) only for about 18% of parks while introducing recency and seasonality as features into the GloES significantly improves the accuracy of the prediction task to 87% of parks with accurate results (4 times more attractions in the accurate interval of MAPE). *GloES+recency+seasonality* produces good predictions (MAPE < 25) for about 87% of the parks. Those results however, are worse than when we apply specialization (if data availability allows), mainly for highly accurate predictions (MAPE < 10). We will show the results in the next. In any case, the good accuracy provided by the GloES with recency and seasonality encourage its application for the cases in which there is a lack of enough training data for specific attractions.

Table 4.11: GloES Prediction results augmented with the other two tourism requirements - seasonality and/or recency trained with 76 national parks in the U.S. (outdoors). Results are in **bold** for the best prediction models.

	Parks			
MAPE	GloES	GloES +recency	GloES +seasonality	GloES +recency +seasonality
MAPE<10	5.26%	3.95%	**30.26%**	**30.26%**
MAPE<25	18.42%	46.05%	81.58%	**86.84%**

Specialized models (Model specialization = ON). Table 4.12 shows the prediction performance of the models when all the three tourism prediction requirements are present i.e. model specialization, seasonality and/or recency. As previously discussed, for outdoor attractions, the specialized models without the new features (SpecES) have a good performance (MAPE < 25%) – over 95% for parks (column *SpecES* in Table4.12).

Table 4.12: SpecES Prediction results augmented with the other two tourism requirements - seasonality and/or recency training for each of the 76 national parks in the U.S. (outdoors). Results in bold for the best prediction models.

	Parks			
MAPE	SpecES	SpecES +recency	SpecES +seasonality	SpecES +recency +seasonality
MAPE<10	22.37%	40.79%	48.68%	**50.00%**
MAPE<25	94.74%	94.74%	**96.05%**	**96.05%**

Considering the results in Table 4.12, we note that for parks the combination of all tourism requirements (fifth column in Table 4.12) performs the best (96% for MAPE < 25 and 50% for MAPE < 10). We will analyze this aspect further in Section 4.3.4, when we perform a factorial analysis over the tourism requirements.

Regarding the high accuracy cases (MAPE < 10), we record that the combination of SpecES with the other two key tourism requirements- recency and seasonality- produced the best overall results. In more details, for the outdoor attractions, comparing the results in Tables 4.3 and 4.12, for (MAPE < 10), *SpecES+recency+seasonality* has highly accurate predictions for 50% of the parks compared to around 42% using the naive-seasonality.

Figure 4.5 illustrates the MAPE error of naive models (y-axis) versus SpecES model (x-axis) for all outdoor attractions. In the left side of this Figure, we observe that error for naive recency is higher than error of SpecES for most of the cases pushing the orange points (attractions) to the y-axis side. However in the right side of Figure 4.5, we have the error for naive seasonality which is comparable to the SpecES model but still has higher MAPE for some of the attractions. All in all, SpecES represents 20% improvement for MAPE < 10 compared to naive seasonality and about 600% improvement over naive recency.

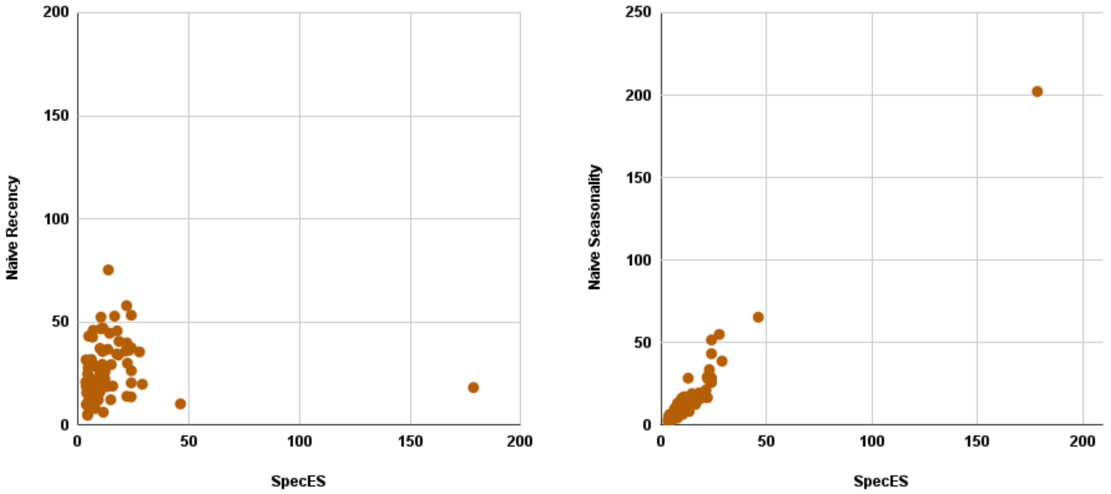


Figure 4.5: Scatter plot of the MAPE error of naive models (y-axis) versus SpecES model (x-axis) for all outdoor attractions (naive recency left and naive seasonality).

4.3.3 Features ablation - scenarios with scarcity in tourism prediction requirements

As discussed in the previous sections, learning specialized models trained with the complete information regarding social, environmental, recency, and seasonality information considerably improves the accuracy of the prediction models. However, having full information regarding recency and seasonality is not always guaranteed. In the following, we further investigate the individual impact of recency and seasonality in the prediction task in scenarios without full availability of historical information on (number of) visits, social media and environmental data for touristic attractions. For these analyses, we revisit the prediction architecture and redefine the training and test sets when necessary.

Only Recency - scarcity in seasonal data: In scenarios that we do not have enough historical information for an attraction, i.e., we only have very recent data on visits, social media and environmental data of a touristic place, we can exploit recency features in order to improve the prediction of the future visitation. This situation may occur, for instance, for new attractions or attractions that have only started to collect (visitation) data very recently. To simulate this scenario in our datasets, we only use the last four (4) months of the historical data of each attraction to train each prediction model while filtering out the rest of the data. Figure 4.6 presents our revised prediction architecture to deal with this new prediction scenario.

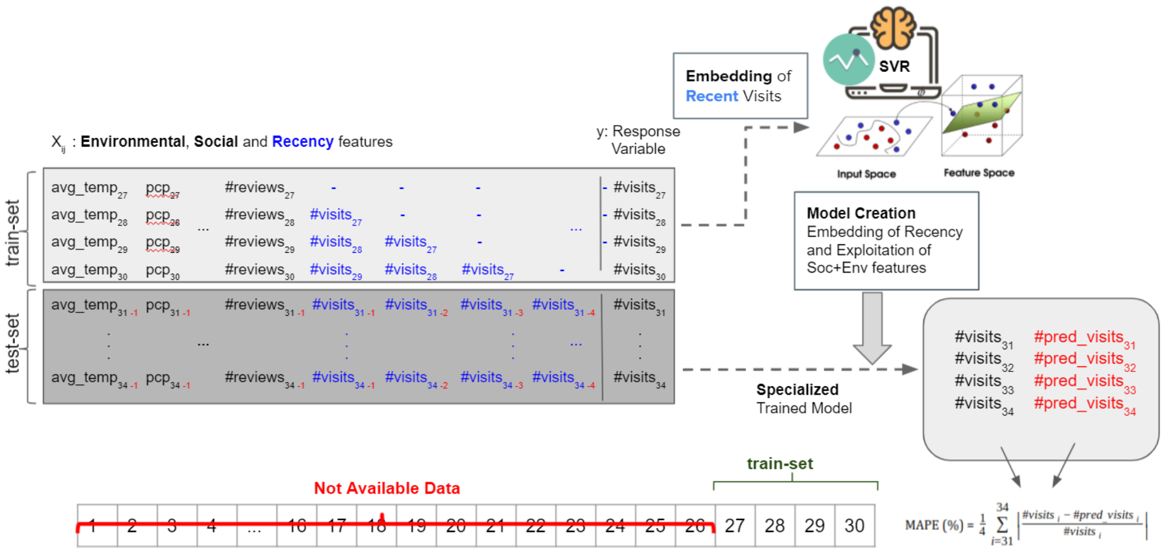
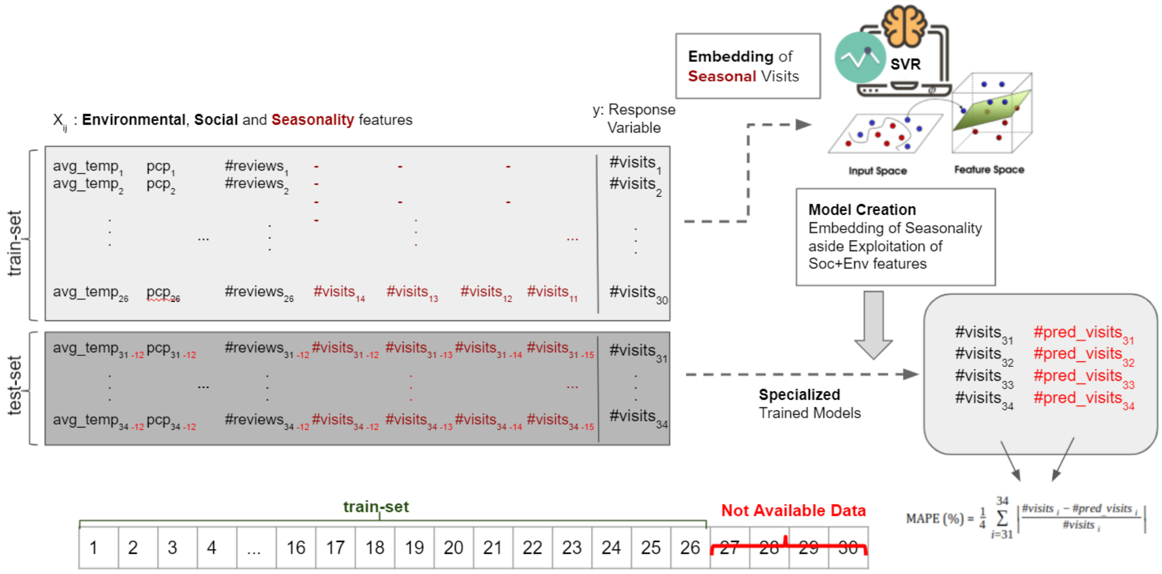

Embedding of Recent Visits
SVR
Input Space
Feature Space
X_ij : Environmental, Social and Recency features
y: Response Variable
train-set
test-set
avg_temp_27 pcp_27 #reviews_27 - - - - #visits_27
avg_temp_28 pcp_28 ... #reviews_28 #visits_27 - - ... - #visits_28
avg_temp_29 pcp_29 #reviews_29 #visits_28 #visits_27 - - #visits_29
avg_temp_30 pcp_30 #reviews_30 #visits_29 #visits_28 #visits_27 - #visits_30
avg_temp_31 -1 pcp_31 -1 #reviews_31 -1 #visits_31 -1 #visits_31 -2 #visits_31 -3 #visits_31 -4 #visits_31
avg_temp_34 -1 pcp_34 -1 #reviews_34 -1 #visits_34 -1 #visits_34 -2 #visits_34 -3 #visits_34 -4 #visits_34
Model Creation
Embedding of Recency and Exploitation of Soc+Env features
Specialized Trained Model
#visits_31 #pred_visits_31
#visits_32 #pred_visits_32
#visits_33 #pred_visits_33
#visits_34 #pred_visits_34
Not Available Data
train-set
1 2 3 4 ... 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
MAPE (%) = 1/4 Σ_{i=31}^{34} |(#visits_i − #pred_visits_i) / #visits_i|

Figure 4.6: Tourism demand prediction methodology in scarcity of seasonal data adopting social media, environmental and recency features

Since we do not have the features of the last 12 months to evaluate the prediction model, we adopted two different scenarios for defining the input value of each feature in the test-set: (i) *last month* case, in which we use the previous month information as the input of the model and; (ii) *mean of 4-months* case, in which we use the mean of each feature of the train-set as the input feature values of the models.

Table 4.13 (two leftmost columns) shows the results. The percentage of parks with an accurate prediction (MAPE<10) is quite low in both cases (about 2%). Regarding good predictions (i.e., MAPE<25), using the last month as the input has a slightly better performance(25%) than using the mean of 4-months features (21%) in outdoors.

Only Seasonality - scarcity in recent data: Likewise the recency features, we analyze the performance of seasonality features when we do not have the most recent data available. This may happen in cases when data collection is periodical (or seasonal) and lasts longer periods and the most recent data is not yet available for prediction. In this scenario, we can adopt seasonality features, i.e. number of visits, social media and environmental data in the previous years in order to predict the future visitation, if this information is available. Figure 4.7 shows our revised prediction architecture to deal with this prediction scenario.



Embedding of Seasonal Visits
SVR
Input Space
Feature Space
X_{ij} : Environmental, Social and Seasonality features
y: Response Variable
train-set
test-set
avg_temp$_1$ pcp$_1$ #reviews$_1$ #visits$_1$
avg_temp$_2$ pcp$_2$ #reviews$_2$ #visits$_2$
avg_temp$_{26}$ pcp$_{26}$ #reviews$_{26}$ #visits$_{14}$ #visits$_{13}$ #visits$_{12}$ #visits$_{11}$ #visits$_{30}$
avg_temp$_{31-12}$ pcp$_{31-12}$ #reviews$_{31-12}$ #visits$_{31-12}$ #visits$_{31-13}$ #visits$_{31-14}$ #visits$_{31-15}$ #visits$_{31}$
avg_temp$_{34-12}$ pcp$_{34-12}$ #reviews$_{34-12}$ #visits$_{34-12}$ #visits$_{34-13}$ #visits$_{34-14}$ #visits$_{34-15}$ #visits$_{34}$
Model Creation
Embedding of Seasonality aside Exploitation of Soc+Env features
Specialized Trained Models
#visits$_{31}$ #pred_visits$_{31}$
#visits$_{32}$ #pred_visits$_{32}$
#visits$_{33}$ #pred_visits$_{33}$
#visits$_{34}$ #pred_visits$_{34}$
train-set
Not Available Data
1 2 3 4 ... 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
MAPE (%) = $\frac{1}{4}\sum_{i=31}^{34}\left|\frac{\#visits_i - \#pred\_visits_i}{\#visits_i}\right|$

Figure 4.7: Tourism demand prediction methodology in scarcity of recent data adopting social media, environmental and seasonality features

For this, we do not use the most recent historical data of each attraction and use only the remaining historical data for training the prediction model. For constructing the training-set, we define two cases regarding the unavailability of historical data: (i) unavailable history of the last 4 months of each feature; (ii) unavailable last 12 months (last year) of each feature. The first case corresponds to the situation where we do not have the previous last 4 months (y-1, y-2, y-3, y-4) while the second case is when we do not have one complete cycle of historical data (annual seasonality) [Rosselló and Sansó, 2017].

Table 4.13 (two rightmost columns) presents the results for outdoor attractions. The percentage of parks with an accurate prediction (MAPE<10) is much higher in the first case when only the last 4 months of the historical data is unavailable in comparison to the case when the complete historical data of the last year is missing (43% versus 0%). This behaviour is similar for good predictions (MAPE<25) in outdoors attractions (85% versus 21%). These results again suggest the importance of the historical data. In other words, having the last trends of visitations besides the periodical/historical behaviors is essential for an accurate prediction.

Table 4.13: Scarcity in seasonal historical and recent data - Evaluating performance of recency and seasonality features in 76 national parks in the U.S.

	Only Recency		**Only Seasonality**	
MAPE	last month	mean of 4-months	unavailable last 4 months	unavailable last year
MAPE<10	1.32%	2.63%	43.00%	0.00%
MAPE<25	25.00%	21.00%	85.00%	21.00%

4.3.4 Factorial analysis of tourism prediction requirements

In this section we investigate the impact of each of tourism prediction requirements, i.e. recency, seasonality and model specialization by means of a factorial design analysis. We employ a regression analysis for evaluating the amount of variation in the prediction results that can be explained by each factor (and interaction). We adopted a $2^k r$ experimental design technique, in order to estimate the effect of $k = 3$ factors (recency, seasonality and model specialization), each of which having two levels (requirement is incorporated into the model or not, for the prediction task) and with r replications per configuration.

As reported in Section 3.2, applying cross-validation along with the SVR model produces small variations in prediction results due to the stochastic nature of the task. In order to reduce this variation and increase the accuracy of results, we executed each experiment several times to calculate the average and standard deviation of the variation of results. We estimated the adequate number of runs based on 95% confidence level and accepted error percentage of 2%, as being 5 runs. In our factorial analysis, the response variable is the % of outdoors attractions that fall in each MAPE range and we want to estimate the importance of each factor (interaction) on the variation observed in those % of touristic attractions. When all three requirements are turned off, we use the global SVR model (non-specialized model trained for all outdoors attractions using only the Environmental and Social media features, i.e. absence of all three factors.

In Table 4.14, we show the variation explained by each tourism requirement on the prediction results outdoors attractions. We observe that model specialization and then seasonality have the largest contributions. In the case of MAPE < 25%, we have also a sig-

Table 4.14: Contribution of each of tourism prediction requirements: recency, seasonality, model specialization and their interactions into the response variable in outdoors attractions - parks; results for MAPE $<$ 10% and MAPE $<$ 25% in 5 runs. The contributions higher than 5% are in **bold** face.

	contribution (%)	
Requirements	MAPE $<$ 25	MAPE $<$ 10
Recency	1.1	1.1
Seasonality	**24.6**	**32.6**
Model spec.	**46.0**	**58.5**
Recency, Seasonality	0.7	2.0
Recency, Model spec.	1.2	2.0
Seasonality, Model spec.	**24.4**	1.8
Recency, Seasonality, Model spec.	1.8	0.9
Residuals	0.2	1

nificant contribution of the interaction between these two factors - seasonality and Model specialization (24.4%).

In addition, we observe that model specialization, in relative terms, is more important to the variation observed for MAPE < 10 than for the results for MAPE < 25 (explains 59% versus 46%). One may say that if we need highly accurate prediction results (MAPE < 10%), the use of specialized models becomes even more important.

We can also see in Table 4.14 that the impact of recency and its interactions with other factors on the prediction results are almost negligible. Despite that, recency can improve results (look for instance at the second and third columns in Table 4.12), indicating that we should use it, mainly if the seasonality features are not available.

Seasonality alone has more than 24% of contribution for MAPE < 25. This indicates that when we have only the historical data for an attraction, we can significantly improve accuracy by injecting seasonality features into the model as input variables. We also observe that seasonality has even a higher impact (32.6% versus 19.8%) in outdoor attractions for very accurate prediction results (MAPE < 10). This is in alignment with what we have discovered in our previous work [Khatibi et al., 2019] when we showed that in outdoor attractions the impact of climate features is high, considering that the climate features have a high correlation with seasonality. [Butler, 2001].

4.3.5 Drill down analysis of encapsulated features in recency and seasonality factors

In the previous section, we have quantified the impact of each of the tourism prediction requirements. In the following, we delve further into the role that each of the recency and seasonality features (introduced in the Section 3.4) play regarding the prediction task accuracy. We will do so by analyzing the learned coefficients of the **global models** in the outdoor scenarios. In other words, we will use the global models as an **analytical tool (only)**. We chose to do so because such analysis would be very complex (if not impossible) if we consider all the 76 models produced with specialization (one for each outdoor attraction).

As can be seen in Table 4.11 (outdoors), the impact of the incorporation of the recency and seasonality features into the global models is similar to that of the specialized models, with significant improvements over the case in which we do not use such features, for MAPE < 10% and MAPE < 25%, although results are not as good as with the latter.

Table 4.15 shows the learned coefficients of global models in outdoor scenarios, respectively. In more details, for this analysis, we built global models for all attractions of each type, adopting each time a different feature-set: (I) soc + env: global model trained having only social media and environmental features in the feature-set; (II) recency (soc + env + rec): global model having recency features in addition to the social media and environmental features; (III) seasonality (soc + env + seas): global model having seasonality features in addition to the social media and environmental features and; (IV) seasonality+recency (soc + env + rec) + seas): global model having all features including social media, environmental, recency and seasonality features.

Considering the learned coefficients in the table, indicates the high importance of number of reviews and then maximum temperature in the simplest model. In the recency model (soc + env + rec), instead, higher weights are given to the number of visits in the last month (y-1 feature) and average temperature; y-12 and y-14 in the seasonality model; and finally y-12 and y-1 for complete feature-set model, which is consistent with our previous discussions in the factorial design analysis for outdoor attractions.

4.3.6 Improving accuracy of anomaly cases

In Section 4.2.4, we identified a small set of outdoor tourism attractions for which our best prediction models – exploiting social media and environmental features – performed poorly. In this section, we evaluate whether the incorporation of seasonality and recency features into the specialized models for these attractions can help to mitigate the found problems. We analyse Bryce Canyon National Park in the U.S. as an example.

In the Bryce Canyon national park, the difficulty was that the considerable increase in number of visits (more than 20% starting in February 2016 until September of the same year in comparison with the same period in 2015) was not accompanied by social media reviews (same behavior as previous years plus a slightly decrease in May 2016 compared to May 2015) (Figure 4.8). A possible reason was the waiving of the entrance fees in 2016. Again,

Table 4.15: The coefficients of features of global (single) model for all 76 U.S. National Parks adopting each time a different set of features: (I) only social media and environmental features (soc+env), (II) social media, environmental and recency features (soc+env+rec), (III) social media, environmental and seasonality features (soc+env+seas), (IV) complete feature set: social media, environmental, seasonality and recency feature (soc+env+rec+seas). The **bold** face shows the top 2 features in each column.

Features	soc+env	soc+env+rec	soc+env+seas	soc+env+rec+seas
tmin	0.006	-0.290	-0.001	-0.011
tavg	**0.021**	**0.580**	0.002	0.022
tmax	0.002	-0.278	0.000	-0.010
temp_dif	-0.015	0.003	0.003	0.003
pcp(rain)	0.000	-0.002	-0.003	-0.001
revs	**0.278**	0.019	0.001	0.000
rating	-0.007	0.005	0.009	0.005
month	-0.007	-0.038	-0.001	-0.005
y-1	-	**1.218**	-	**0.340**
y-2	-	-0.297	-	0.014
y-3	-	-0.070	-	0.018
y-4	-	0.066	-	0.014
log y-1	-	-0.007	-	0.034
log y-2	-	0.023	-	0.004
log y-3	-	-0.023	-	-0.013
log y-4	-	0.001	-	-0.001
y-12	-	-	**0.947**	**0.928**
y-13	-	-	0.026	-0.262
y-14	-	-	**0.031**	-0.018
y-15	-	-	-0.010	-0.038
log y-12	-	-	0.009	-0.013
log y-13	-	-	-0.003	-0.024
log y-14	-	-	-0.012	-0.005
log y-15	-	-	0.003	0.013

by explicitly exploiting seasonality and recency we were able to capture such anomalies, reducing the mean percentage error of the models from 35% to 24%, i.e. 30% reduction in the prediction error. (Table 4.16).

Table 4.16: accuracy of difficult cases incorporating explicit tourism prediction requirements in outdoor attractions

Attraction	MAPE - previous results (SpecES)	MAPE - SpecES+recency+seasonality
U.S. Bryce Canyon National Park (outdoor)	35.19%	**24.43%**

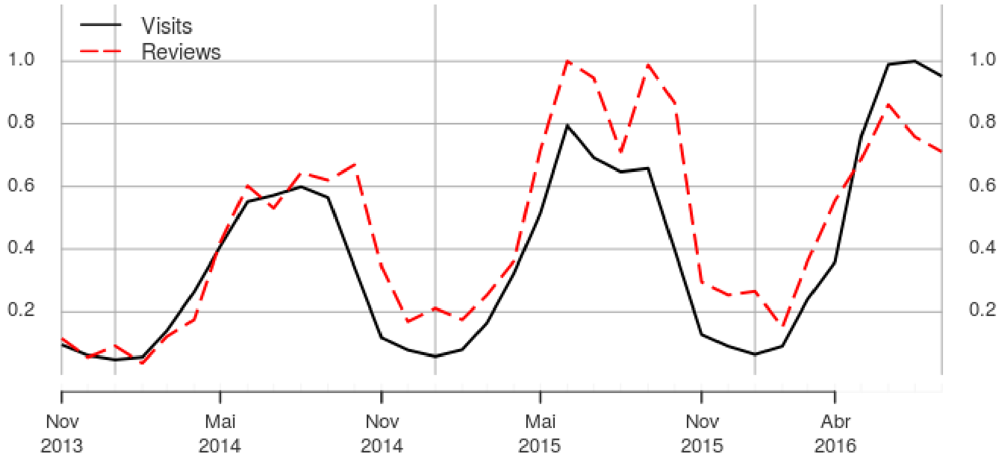


Figure 4.8: Temporal evolution of number of visits and social media comments in Bryce Canyon National Park in the U.S. (the values in y-axis are normalized by the maximum value of each variable in the whole period).

4.4 Clustering of Attractions

In this section, we analyse the similarity of various outdoor attractions using multiple clustering algorithms in order to build a single prediction model for each group/cluster of attractions. We then evaluate the single model's error accuracy with the specialized model of each attraction in that particular cluster. Applying clustering techniques is particularly interesting in cases for which we have little information about some attractions. In this scenario, we can exploit prediction models of similar attractions to complete information by ”transferring' information learned from attractions with more information. Obviously these attractions should have high similarity in their visits patterns, particularly in their recency and seasonality features requirements for this idea to work. Here, we define clusters of attractions running k-means [Jin and Han, 2010] and dB-scan [Sander, 2010], two well-known clustering algorithms in the literature. For better clustering visualization we use PCA analysis [Jolliffe and Cadima, 2016].

We first turn our attention to extract clusters using k-means. The total number of clusters k we chose is the one that maximizes the silhouette quality measure [Wang et al., 2017]. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to +1 – a high value indicates that the object is well matched to its own cluster and loosely matched to neighboring clusters. Figure 4.9 depicts the silhouette metric varying k from 2 to 15. Best silhouette value is achieved when k is equal to 7. Figure 4.10 plots the distribution of parks in each cluster for k = 7 showing that clusters 1 and 2 share some features that neither first nor second principal components are able to distinguish. However, more than 80% (PC1 +

PC2 = 36.1 + 45.1) of variations in clusters can be explained by the two main components.

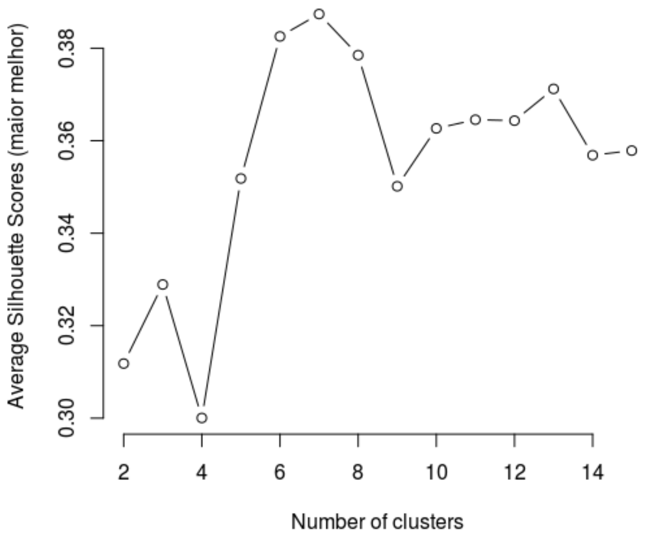


Figure 4.9: Average silhouette score for different number of clusters using k-means clustering algorithms for 76 outdoor attractions

We further analyze the results of clustering outdoor attractions using DB-scan clustering technique due to the geometry of the cluster distribution. Figure 4.11 presents the distribution of parks after clustering these attractions with k = 5 clusters (cluster 1 represents noise shown as red points). One problem in this scenario is the high percentage of points clustered as noise by DB-scan (about 30% of attractions). On the other hand, we observe a better division of outdoor attractions again exploiting PCA analysis to visualize the distribution of clusters (Figure 4.11) having PC1 + PC2 = 70%. All in all, it seems that k-means was not able to cluster well even with a larger principal component variation explained (about 80%) with respect to DB-scan (about 70%).

Once we have applied different clustering models, which split similar attractions into groups, we can build a prediction model for each cluster. For every new attraction to be analyzed, we attribute this attraction to one of the clusters, then we run its prediction model to calculate the corresponding number of visits to this venue. Table 4.17 shows the error accuracy for different number of clusters in outdoor attractions. Extreme k values, such as k = 1, reduce the prediction model to the previous global model for all attractions (alike GloES+recency+seasonality model presented in Section 4.3.2) and k = 76, ends to specialized model for each attraction (same as SpecES+recency+seasonality model discussed in Section 4.3.2). We observe that clustering the outdoor attractions into 5 and 7 clusters and

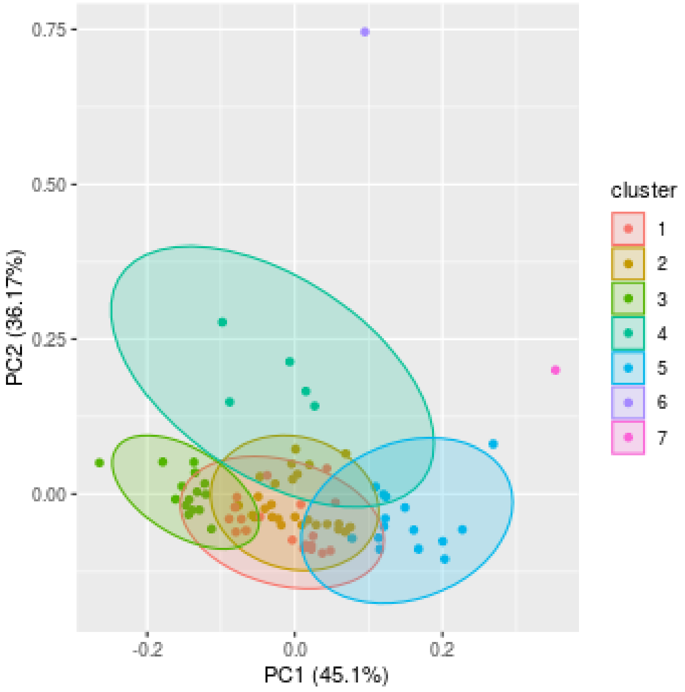


Figure 4.10: Clustering of parks using k-means clustering algorithms adopting social media, environmental, recency and seasonality features

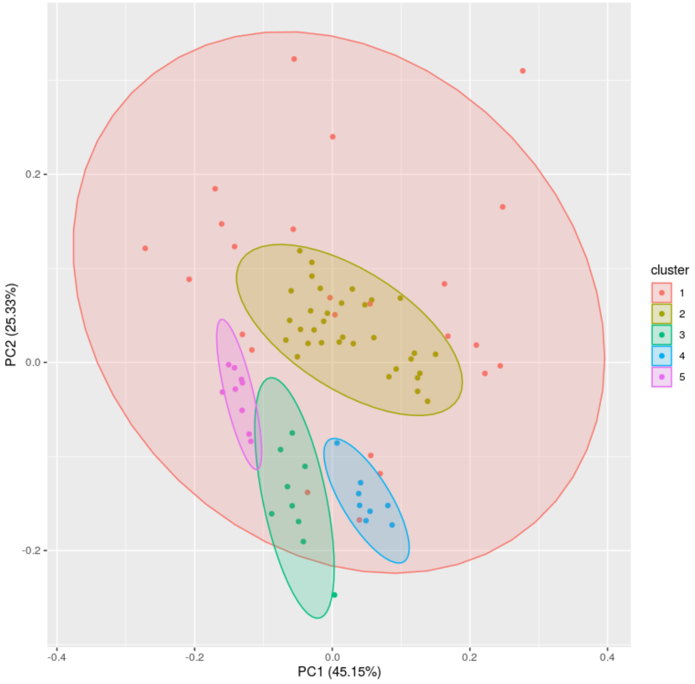


Figure 4.11: Clustering of parks using k-means clustering algorithms adopting social media, environmental, recency and seasonality features

running a model for each cluster increased the error of many attractions in comparison to GloES+recency+seasonality model (global model where practically k = 1). For MAPE < 25%, we have 30% of attractions with good accuracy results using the global model while this percentage reduces to 28% and 25% respectively for k = 5 and k = 7. With respect to MAPE < 10%, again we could not obtain better results than a global model using k = 5 and k = 7, i.e. 82% and 87% in order versus 87% for Global model.

All in all, we could not find evidence that grouping the clusters into more than one cluster (k>1) could improve the accuracy of prediction models with respect to the global model. In other words, we can observe that creating a single cluster (k=1) for outdoor attractions offers the best results in terms of MAPE values for both, MAPE < 10% and MAPE < 25%.

Table 4.17: Percentage of National Parks in each interval of MAPE using different number of clusters in order to cluster outdoor attractions

No of clusters	MAPE<10%	MAPE<25%
k = 1 (Global Model)	30	87
k = 5 (DB-scan)	28	82
k = 7 (k-means)	25	87
k = 76 (Specialized Models)	50	96

4.5 Summary and Remarks in outdoor attractions

In this Chapter, we presented the results obtained so far with respect to research goals of this dissertation. We demonstrated that by explicitly exploiting three key requirements of tourism prediction as explicit features in our models, besides feeding the models with robust external features such as social media and environmental features, we can greatly improve prediction accuracy. We can summarize our main contributions by answering the research questions as follows:

- We exploited external data – social media features along with with environmental data – to forecast touristic demand at outdoor attractions considering in 76 National Parks in the U.S. Accurate prediction results ($MAPE < 25\%$) for more than 93% of the outdoor attractions were achieved using only external features when SVR prediction model is applied. These results support our hypothesis in **RQ1** of high predictability when jointly exploiting social media and environmental features as a tool for tourism demand forecasting for places and attractions, mainly for scenarios with unavailable official visitation census. We evaluated the performance of each category of features, i.e social media and environmental features in tourism prediction in outdoor attractions. The environmental features had a better performance as expected. However, best results were obtained when both types of features were used together.

- We investigated the impact of exploiting recency and seasonality features alongside social media and environmental data to improve the performance of specialized prediction models for outdoor touristic attractions. Our experiments addressed **RQ2** by demonstrating that fulfilling tourism demand prediction requirements, i.e. recency, seasonality and model specialization can increase the accuracy of forecasting touristic demand at fine-grained levels in comparison to results with only external features (absence of recency and seasonality features). In fact, the exploitation of the complete set of key tourism requirements outperforms the best baselines results for high accuracy cases ($MAPE < 10\%$) for outdoor attractions (50% versus baseline 42%). Moreover, accurate prediction results ($MAPE < 25\%$) were achieved for more than 96% of the outdoor attractions when the SVR prediction model was applied exploiting the seasonality and model specialization.

- We analyzed the impact of each of the tourism prediction requirements individually and their interactions applying a 2^k factorial design analysis. The results indicate the higher importance of model specialization factor (46% in $MAPE < 25\%$), then seasonality (above 24% of contribution) and its interaction with model specialization (more than 24%) and, finally, recency features with a low importance (less than 2%) in the scenario of outdoors. However, as explained we should not discard recency features since in attractions without available seasonality features, recency features improve the accuracy of prediction models (models with only social media and environmental features incremented by recency features). We quantified the performance of each of the three tourism prediction factors (requirements) in the learned models, observing the higher impact of model specialization and seasonality features in model accuracy. But even the less impacting recency features can increase the accuracy of the models, mainly when we do not have the seasonal data of an attraction available.

- We studied different clustering techniques to group similar attractions in order to build a single prediction model for each cluster. This could be potentially useful in scenarios for which we have little information about some specific attractions. We observed that in the outdoors scenario, clustering could not outperform the global model in our experiments. We will further investigate the reasons for this behavior in the future.

- We provided a thorough experimental analysis regarding the cases for which our proposed prediction methodology was not able to produce satisfactory results. We elaborated on possible causes – most cases were related to changes in data distributions caused by unpredictable events including steam rising from a crater and eruption scare, waiving attraction's entrance fees and renaming a national park.

Chapter 5

Experimental Results - INDOOR Attractions

This chapter will replicate the analysis in Chapter 4, only now, for indoor attractions. We present our experimental results to address our two research goals in the scenario of indoor attractions, i.e. 27 Museums and Galleries in the United Kingdom. First, we present our correlation analysis of different features extracted from the environmental, social media and official datasets in Section 5.1. Then Section 5.2 discusses prediction results using different combinations of external features, namely social media and environmental data in indoor attractions that fulfill our first research goal (**RQ1**). Again, for that analysis, we adopt the tourism demand prediction architecture which was presented in Chapter 3 (see Figure 3.3).

Next, Section 5.3 elaborates the prediction power of explicitly exploiting the three key requirements of tourism prediction as features in our models addressing our second research goal (**RQ2**). That study exploits the second tourism demand prediction architecture which was presented in Chapter 3 (see Figure 3.4). We even evaluate the prediction power of each tourism requirement in a factorial design analysis, in cases with scarcity in historical data and cases where recency and seasonality features could improve the prediction of inaccurate cases. Furthermore, we identify the cases, where the combination of external features resulted in inaccurate predictions while addition of key tourism requirements could improve the accuracy of those cases.

Finally, Section 5.4 studies the possibility of creating clusters of similar indoor attractions in order to build a single prediction model for each group useful in scenarios that we have little information about some attractions. We finish this chapter by a summary of all obtained results in Section 5.5. As like as outdoor attractions, we note that, for the sake of reproducibility, all our datasets as well as tools and other pieces of code are publicly available[1].

5.1 Correlation Results

This section aims to motivate the prediction of touristic demands exploiting both social media and environmental data by presenting correlation results of various features with the official number of visits for indoor attractions.

Table 5.1 presents correlations of each of the features with number of visits (#Visits) for the indoor attractions. These correlations were computed across all attractions in the indoors dataset. As we can see, the feature with the strongest (positive) correlation with #Visits is #Reviews (0.45). However, we can also see that the correlations of most of the other features with #Visits are not negligible. Although such correlations are not the only aspect that influences their prediction power[2], they are an indicator that the considered features may bring some information to the prediction task. Additionally, considering the temperature features (Min, average and Max temperature), the correlation results with #visits are also considerable again showing the importance of temperature features even in the case of indoor attractions.

Table 5.1: Overall correlations of #Visits with other features for indoor attractions. Strongest correlation is in bold.

Feature	Correlation with #Visits	Feature Type
Min_Temp	0.12	Environmental
Avg_Temp	0.14	Environmental
Max_Temp	0.14	Environmental
Precipitation/Rainfall	-0.13	Environmental
Air_Frost_days	-0.09	Environmental
Sunny_Hours	0.08	Environmental
#Reviews	**0.45**	Social Media
Avg_Rating	0.04	Social Media

In the following, we further analyze the highest correlated feature with #Visits, namely #Reviews, for indoor attractions. The last column of Table 5.2 shows the percentage of museums and galleries that fall in each of the three aforementioned categories (A, B and C), which were defined based on the correlations between monthly number of visits and monthly number of reviews in the period of November 2011 till July 2018 for the indoor attractions. Note that, the percentage of attractions with moderate-to-strong correlations (categories A and B) is 48%.

[1] Available at Repository Mendeley - https://data.mendeley.com/datasets/t7bfhtzhxg/1

[2] Indeed we cannot claim any causality effect, but rather only some relationship between the variables.

Table 5.2: Distribution of attractions across three categories based on correlations between #Visits and #Reviews.

Category	Correlation (#Visits and #Reviews)	Indoor Attractions
A	over 65% (strong correlation)	30%(8 museums)
B	50% to 65% (moderate correlation)	18% (5 museums)
C	less than 50% (weak correlation)	52% (14 museums)

Later on, we will see that the feature #Reviews is indeed one of the best predictors in the prediction task for indoor attractions.

In Figure 5.1 we present a Random Forest analysis for indoor dataset, i.e., the influence, by percentage, of each feature in the predictability within the indoor attractions. The social media feature - number of reviews - has again the highest predictive. Then comes the other social media feature - average rating. Only after those we have environmental features - sunny hours during the month, minimum temperature, rainfall, maximum temperature and air frost days respectively. However, in this case, #Reviews is more than five times more influential than the best environmental feature. Indeed the social features together account for 65% of the predictive capability in this dataset. In any case, the environmental features still have some contributions in the prediction task, mainly when considered altogether.

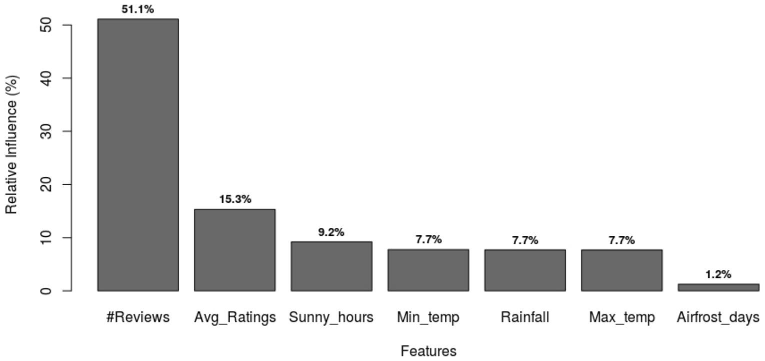


Figure 5.1: Features relative influence in percentage (%) - indoor dataset

5.2 External Data Performance in Visits Prediction

We now shift our attention to the prediction performance of external features in indoor attractions. Recall that in this case the environmental features available include average, minimum and maximum temperatures, as well as rainfall, sunny hours and days of air frost. Since the total monthly number of reviews for all attractions before November 2011 was very low, as shown in Figure 5.2, for the prediction period, we considered the 80 most recent months (Nov. 2011 till July 2018), training all our models with the first 76 months in order to predict the next 4 months of visitations. Also, since GRNN had a much poorer performance than the other techniques for outdoor attractions, we here focus our analyses on the seven more competitive approaches, namely, SVR, Linear Regression (LR), SARIMA, SARIMAX, LSTM and two naive models - naive recency and naive seasonality (all these methods were discussed in Chapter 2.1).

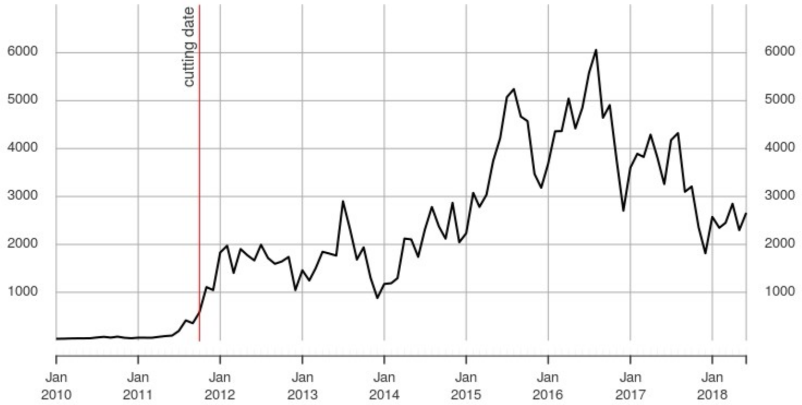


Figure 5.2: Number of TripAdvisor reviews (y-axis) for all indoor attractions.

Table 5.3 shows the percentage of museums and galleries for which the prediction, using each considered technique, fell within the ranges of high accuracy (MAPE $<$ 10%), good accuracy (10% $<$ MAPE $<$ 25%) and low accuracy (MAPE $>$ 25%). As it can be seen in this Table, we have the specialized SVR model with Environmental and Social features as the best model, predicting accurately for the highest percentage of attractions (almost 93% of Museums), considerably outperforming other models. However, best results for MAPE less than 10% (highly accurate results) are achieved by the naive-recency (26%) for the indoor attractions. The success of the naive methods in highly accurate results (MAPE < 10%) is one of the reasons that motivates the adoption of recency and seasonality in our models that we will investigate in the next section when explicitly incorporated as features into our best model (SVR).

Again, for the sake of completeness and validation of the LR models (one linear regression model per each indoor attraction), we report the R-squared values, R-Squared-Adjusted and F-Statistics. We had R-squared values in the interval of [0.15, 0.88] with the average value of 0.47; R-Squared-Adjusted in the interval of [0.07, 0.86] having 0.42 as the mean value; F-Statistics within [1.76, 68.7] having average of 13.0. For 93% of the attractions, the p-value was $<$ 0.05 (95% of confidence). In other words, with 95% of confidence, the LR models are significantly better than the average value of #Visits at least for 93% of the muse-

Table 5.3: Percentages of indoor attractions (museums and galleries) that fall into different ranges of MAPE value for each prediction technique. Bold values show the prediction technique with higher percentage of attractions with good-to-high prediction results.

MAPE	SVR	LR	SARIMAX	SARIMA	LSTM	n.recency	n.seasonality
lower 10%	14.81%	7.41%	7.41%	7.41%	11.1%	**25.93%**	18.52%
10% to 25%	**77.78%**	55.56%	62.96%	48.15%	62.97%	44.44%	62.96%
over 25%	7.41%	37.04%	29.63%	44.44%	26%	29.63%	18.52%

ums [Clark et al., 1974]. Likewise outdoor models analysis, we report R-Squared, R-Squared Adjusted, F-statistics and P-value statistics for indoor linear models in Table 5.4.

Table 5.4: Statistical analysis of R-Squared, R-Squared Adjusted, F-statistics and P-value for indoors linear Regression model

Metric	Min	Average	Max
R-Squared	0.15	0.47	0.88
R-Squared-Adjusted	0.07	0.42	0.86
F-Statistics	1.76	13.0	68.7
% of attractions with p-values < 0.05		93%	

5.2.1 Feature analysis

Table 5.5 shows the prediction performance of all models, with all features and with the individual features in isolation. We can see that airfrost_days and #Reviews in general produce the best or close to the best results when used in isolation. There are cases in Linear Regression and SARIMAX models in which prediction with only one feature could produce better results than having all the features in prediction. Again, we hypothesize that this may be due to the noise when we have all the environmental or social media features together in the prediction models. Remind that a similar behavior was detected in the outdoor scenario (Table 4.5) for both SARIMA and GRNN – the prediction with average temperature was better than the one with all features. On the other hand, SVR demonstrated to be very robust and insensitive to any existing noise, being able to extract useful information with the incorporation of more features. Finally, among the environmental features, Air_frost_days individually produced the best performance within most of the prediction models.

As before, we should mention that SARIMAX results are always boosted by the incorporation of the history of visits as a mandatory feature, making the comparison somewhat unfair when comparing individual features performance within prediction methods. As it can be observed in Table 5.5, SARIMAX has accurate prediction for about 85% of indoor attractions when we use only Rainfall (and history of visits) as features.

In any case, all features present close (and usually good) results, mainly when applied together with SVR. And, differently from the outdoor scenario, the gains of the model using all features are much higher, as the results with all features is much better than any feature in isolation – 19% of improvements over the best individual feature when using SVR - the model that produced the overall best results with all features (92.59%). This may indicate that all features can somewhat contribute to the whole prediction process.

Table 5.5 shows the prediction performance of each technique using all features (F) and only individual features in isolation. Consistently with results in Table 5.3, we find that the best results with the complete set of features are obtained with SVR, as it was also observed for outdoor attractions. The second best prediction model with the complete set of features is SARIMAX (as opposed to Linear Regression, which was the case for outdoor attractions).

Moreover, out of the individual features, airfrost_days, rainfall and #Reviews in general produce the best or close to the best results when used in isolation. Indeed, among the environmental features, airfrost_days and rainfall individually produced the best performance for all prediction models, outperforming even the complete set of features for Linear Regression and SARIMAX[3]. We hypothesize that this may be due to some noise that may be introduced when all the environmental features are used together. However, unlike the other techniques, SVR demonstrated to be very robust and insensitive to any existing noise, being able to extract useful information with the incorporation of more features.

Focusing on our best technique, namely SVR, we observe that the improvements from jointly exploiting all features over using any single feature, is more impressive for indoor attractions (19% improvement over the best individual feature) than for outdoor attractions (10% improvements). In any case, we find once again that, the joint use of all features contribute significantly to the whole prediction process.

5.2.2 Feature ablation analysis

We also compare the performance of each technique for the complete feature set F and for all features but one (or a subset) $F - f$. The results, in terms of percentage of museums and

[3] Indeed, for SARIMAX, the feature rainfall, produced the single best result. However, recall that SARIMAX is always boosted by the incorporation of the history of visits as a mandatory feature, making the comparison with other techniques when using only individual features somewhat unfair. As it can be observed in Table 5.5, SARIMAX has an accurate prediction for about 85% of indoor attractions when we use only rainfall (and history of visits) as features.

Table 5.5: Prediction results (% museums and galleries with MAPE $<$ 25%) for indoor attractions: all features versus single features. For each prediction technique, individual features with best prediction results are marked in bold.

Features	SVR	LR	SARIMAX	SARIMA	n.recency	n.seasonality
All Features	92.59%	62.96%	70.37%	-	-	-
Min_temp	70.37%	**70.37%**	74.07%	-	-	-
Max_temp	70.37%	59.26%	66.67%	-	-	-
Airfrost_days	**74.07%**	**70.37%**	77.78%	-	-	-
Rainfall	**74.07%**	**70.37%**	**85.19%**	-	-	-
Sunny_hours	**74.07%**	55.56%	62.96%	-	-	-
#Reviews	**74.07%**	**70.37%**	77.78%	-	-	-
Avg_Ratings	70.37%	**70.37%**	74.07%	-	-	-
History #Visits	-	-	-	**55.56%**	**70.37%**	**81.48%**

galleries with good-to-high prediction accuracy (MAPE $<$ 25%), are shown in Table 5.6.

Note that the social media feature #Reviews is the most important feature, as its removal causes the largest drop in performance for all prediction techniques. On the other extreme, rainfall, maximum temperature and sunny hours are the weakest features in our set, as their removal causes the smallest decrease in performance (for SVR) and may actually improve the results (for Linear Regression and SARIMAX). Once again, this may be explained by a strong correlation of these features with others in the remaining set (e.g., maximum and minimum temperature, which have correlation of 96% in the whole dataset), as well as the small role that these environmental features may play in one's experience indoors. However, out of the environmental features, **air_frost_days** and minimum temperature have a significant impact on prediction performance, notably on our best technique (SVR).

5.2.3 Factorial design of the best features

We also performed an ANOVA type II test to quantify the relative importance of the most important factors (and their interactions) in our models. We chose to focus our analysis on two factors – the best representative of the social feature as well as the best representative of the environmental features – to make a parallel with the analysis of the previous section.

Considering the results in Tables 5.5 and 5.6, we selected #Reviews as the representative of social media features. As mentioned, air_frost_days is one of the best environmental features. We also note that air_frost_days has a strong (negative) correlation with both minimum and maximum temperatures (Pearson coefficient of -0.74 and -0.72, respectively), and a moderate (negative) correlation with Sunny_hours (-0.54). Given such observations, we considered Air_frost_days to be the representative of environmental features.

Table 5.6: Prediction Results (% of museums and galleries with MAPE < 25%) for indoor attractions: all features versus all but one feature (or feature subset). For each prediction technique, values in bold show most contributing individual feature to the feature set for each model.

Features	SVR	Linear Reg.	SARIMAX
All Features (F)	92.59%	62.96%	70.37%
F - Environmental Features	77.77%	70.37%	70.37%
F - Social Media Features	74.07%	51.85%	59.26%
F - Min_temp	77.77%	77.77%	77.77%
F - Max_temp	81.48%	77.77%	77.77%
F - Air_frost_days	77.77%	70.37%	70.37%
F - Rainfall	85.18%	66.66%	74.07%
F - Sunny_hours	81.48%	70.37%	74.07%
F - #Reviews	**74.07%**	**55.55%**	**62.96%**
F - Avg_Ratings	77.77%	62.96%	74.07%

The feature air_frost_days by itself has also some unique information to contribute to the prediction process, as show by the pairwise Pearson Correlation with the other environmental features, for all museums, as per below:

- -0.74 correlation for (Air_frost_days, Min_temp);
- -0.72 for (Air_frost_days, Max_temp);
- 0.02 for (Air_frost_days, Rainfall); and
- -0.54 for (Air_frost_days, Sunny_hours).

Table 5.7 shows the results of the ANOVA test, once again segmented for three categories of attractions. We see that #Reviews explains the largest fraction of the data variation that can be explained by the considered factors, whereas air_frost_days and their interaction play a much less important role. However, unlike observed for outdoor attractions, the two factors explain most variation for only a smaller fraction of attractions. For example they explain up to 53% of the total variation, on average, for only 30% of the attractions. In other words, a great part of the data variation remains unexplained by the two considered factors. These results suggest that, consistently with our previous analyses, other features (included or not in our dataset) have also important role in the final prediction results.

Table 5.7: Contribution of #Reviews and Air_frost_days (and their interaction) to explain variation in #Visits in indoor attractions (Results of ANOVA type II analysis. Categories refer to Table 5.2. Columns p-value include percentage of attractions, in each category, for which the effect of each factor/interaction is statistically significant.).

Factor	Category A (30% of Museums)			Category B (18% of Museums)			Category C (52% of Museums)		
	P-value (< 0.01)	Mean contrib.	Max contrib.	p-value (< 0.01)	Mean contrib.	Max contrib.	p-value (< 0.01)	Mean contrib.	Max contrib.
#Reviews	100%	50%	71%	100%	29%	36%	36%	9%	20%
Air_frost_days	13%	2%	8%	40%	4%	11%	14%	5%	18%
(#Reviews* Air_frost_days)	13%	1%	7%	0%	1%	2%	0%	1%	3%

5.3 Augmentation of Tourism Prediction Requirements

In this section, we aim to demonstrate that similarly to outdoors attractions, by explicitly exploiting the three key requirements of tourism prediction – recency, seasonality and model specialization – as features in our models alongside the external data, i.e social media and environmental features, we can greatly improve prediction accuracy of the results presented in the previous section (Section 5.2). For experiments in this section, we adopt the second tourism demand prediction architecture which was presented in Figure 3.4.

5.3.1 Model Specialization

Now, we provide further evidence of the importance of considering model specialization as an explicit requirement for tourism prediction.

Specialized vs. Global Prediction Models. Here, we compare the prediction accuracy of the specialized models, reported in the previous Sections, with that of a global model trained with all attractions of indoor. In this comparison, we employ the SVR method as it produced the best results among the tested methods. The experimental setup is similar, with the difference being only on the training set, that now contains data from all indoor attractions. Results are shown in Table 5.8. We observed that in the case of indoor attractions, the global model has a good MAPE (MAPE < 25%) only for 11% of museums while specialized models have a much better performance with over 93% of museums having good prediction accuracy.

Table 5.8: Percentages of indoor attractions that fall into different ranges of MAPE value for specialized and global models using the SVR model.

MAPE	SVR Global Model	SVR Specialized model
lower 10%	3.7%	14.81%
10% to 25%	7.4%	77.78%
over 25%	88.89%	7.41%

To further illustrate our argument that individual models are more adequate to compare idiosyncratic aspects of each attraction,, Table 5.9 presents the learned coefficients for the specialized and global models for two attractions - National History Museums (NHM) in Kensington and in Trint, for which good results were obtained with the specialized models. As we can see in this Table, the relative importance of features are different for these two attractions, despite the good results of their respective individual models – both with good MAPE values (MAPE $<$ 25). For instance, in the specialized model for NHM-Kensington, the environmental feature (Min_Temp) has a higher value than the social media feature (number of Reviews) while the opposite is true in case of NHM-Tring. On the other hand, the global model, which tries to capture an "average" behavior for all attractions, fails to return accurate predictions for both NHMs, specially for NHM-Tring. Interestingly, even with the global model, the relative importance of the features, as captured by the respective coefficients for each feature, is consistent with our previous discussions.

Table 5.9: Comparison of coefficients within global and specialized models with 95% of confidence; NHM: National History Museum, **Spec: Specialized model, Glo: Global Model**. The coefficient with highest value in each model is in bold.

Model	Max_temp	Min_temp	Air_frost days	Rainfall	Sunny hours	#Reviews	Avg Ratings	**MAPE Glo**	**MAPE Spec**
Spec - NHM Kensington	-1.69	**1.24**	0.17	0.07	1.09	0.01	-0.5	35.3	11.02
Spec - NHM Tring	-0.08	-0.06	0.11	0.04	0.046	**0.35**	0.03	762.0	20.2
Glo - all Museum	0.12	-0.08	0.003	-0.023	-0.038	**0.52**	-0.056	-	-

5.3.2 Recency and Seasonality

We now shift our attention to demonstrate how the addition of the other two tourism requirements, i.e. recency and seasonality, into the most accurate prediction models in the previous section (SVR models with Social and Environmental features) can improve the accuracy of forecasting the visitations for fine-grained touristic points this time for indoor attractions. In the next, we present this analysis first for global models and then for the specialized models separately. It is worth noting that the reason behind this separation is isolating the effect of model specialization in the study of recency and seasonality features since there may be influence between these requirements.

Global model (Model specialization = OFF). The application of model specialization may be considerably jeopardized when we do not have enough data to train individual models for each site. In this case, it is more viable to train and apply a single global model taking advantage of the complete social and environmental (training) data for multiple attractions (hereafter called GloES - Global model with Environmental and Social data). In here, we show the predictive power of trained GloES augmented with seasonality and recency tourism features for indoors tourism attractions. Table 5.10 show the results. In the case of indoor attractions, the GloES has a good MAPE (MAPE < 25%) only for about 11% of museums.

We can also observe in the Table 5.10 that introducing recency and seasonality as features into the GloES significantly improves the accuracy of the prediction task. *GloES+recency+seasonality* produces good predictions (MAPE < 25) for about 74% of the museums, putting almost 6 times more attractions in the accurate interval of MAPE. Those results however, are worse than when we apply specialization (if data availability allows), mainly for highly accurate predictions (MAPE < 10). We will show the results in the next. In any case, the good accuracy provided by the GloES with recency and seasonality encourage its application for the cases in which there is a lack of enough training data for specific attractions. In any case, the good accuracy provided by the GloES with recency and seasonality encourage its application for the cases in which there is a lack of enough training data for specific attractions.

Table 5.10: GloES Prediction results augmented with the other two tourism requirements - seasonality and/or recency training with 27 museums in the U.K. (indoors). Results are in **bold** for the best prediction models.

	Museums			
MAPE	GloES	GloES +recency	GloES +seasonality	GloES +recency +seasonality
MAPE<10	3.7%	7.41%	**25.93%**	7.41%
MAPE<25	11.11%	48.15%	**77.78%**	74.07%

Specialized models (Model specialization = ON). Training specialized models for each individual attraction allows the models to learn specific patterns of visitation at each touristic point. Table 5.11 shows the prediction performance of the models when all the three tourism prediction requirements are present i.e. model specialization, seasonality and/or recency. As previously discussed, for indoor attractions, the specialized models without the new features (SpecES) have a good performance (MAPE < 25%) – over 92% for museums (column *SpecES* in Table 5.11.)

Table 5.11: SpecES Prediction results augmented with the other two tourism requirements - seasonality and/or recency, trained for each of the 27 museums in the U.K. (indoors). Results in bold for the best prediction models.

	Museums			
MAPE	SpecES	SpecES +recency	SpecES +seasonality	SpecES +recency +seasonality
MAPE<10	14.81%	29.63%	**48.15%**	**48.15%**
MAPE<25	92.59%	**96.30%**	**96.30%**	92.59%

Considering the results in Table 5.11, we note that the combination of only seasonality and model specialization for museums (fourth column in Table 5.11) results in a slightly higher accuracy (96% for MAPE < 25 and 48% for MAPE < 10) than when all features are used. We will analyze this aspect further in Section 5.3.4 when we perform a factorial analysis over the tourism requirements.

Regarding the high accuracy cases (MAPE < 10), we record that the combination of SpecES with the other two key tourism requirements- recency and seasonality- produced the best overall results. In more details, for the indoor attractions (Table 5.11), *SpecES+recency+seasonality* produced high prediction accuracy for about 48% of the museums compared to 26% obtained by the naive-recency (Table 5.5), the best baseline in this category.

Figure 5.3 illustrates the MAPE error of naive models (y-axis) versus SpecES model (x-axis) for all indoor attractions. In the left side of this Figure, we observe that error for naive recency is higher than error of SpecES for many of the cases specially those with MAPE > 25, pushing the blue points (indoor attractions) to the y-axis side. However in the right side of Figure 5.3, we have the error for naive seasonality which is comparable to the SpecES model but still has higher MAPE for some of the attractions. All in all, SpecES represents about 150% improvement for MAPE < 10 compared to naive seasonality and 85% improvement over naive recency.

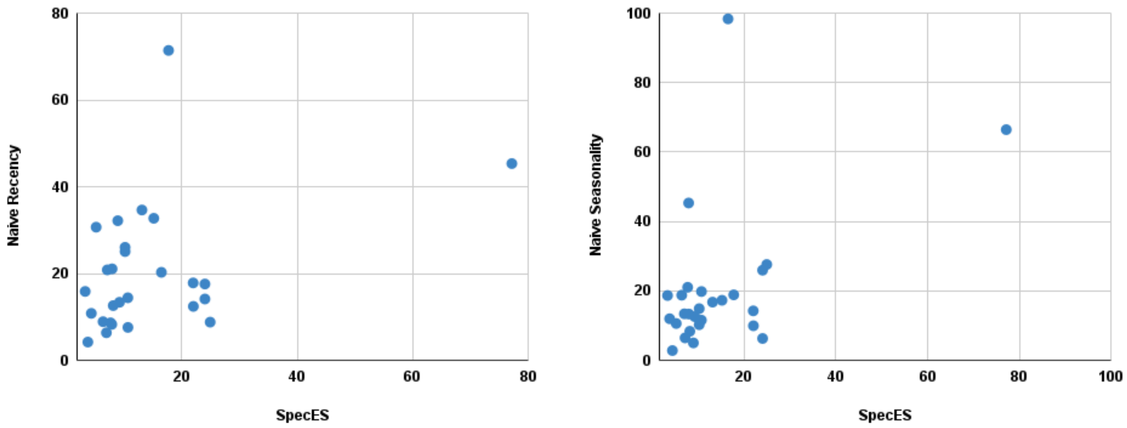


Figure 5.3: Scatter plot of the MAPE error of naive models (y-axis) versus SpecES model (x-axis) for all indoor attractions (naive recency left and naive seasonality).

5.3.3 Features ablation - scenarios with scarcity in tourism prediction requirements

Now we investigate the individual impact of recency and seasonality in the prediction task in scenarios without full availability of historical information on (number of) visits, social media and environmental data for indoors touristic attractions. For these analyses, the prediction architecture, the division of training and test sets would be exactly the same as what we discussed in the previous chapter (Figures 4.6 and 4.7).

Only Recency - scarcity in seasonal data: As we presented in the outdoor attractions in the previous chapter, when we do not have enough historical information for an attraction, i.e., we only have very recent data on visits, social media and environmental data of a touristic place, we can exploit recency features in order to improve the prediction of the future visitation. Likewise outdoors, we simulate this scenario using only the last four (4) months of the historical data of each attraction to train each prediction model while filtering out the rest of the data. Again, since we do not have the features of the last 12 months to evaluate the prediction model, we adopted two different scenarios for defining the input value of each feature in the test-set: (i) *last month* case, in which we use the previous month information as the input of the model and; (ii) *mean of 4-months* case, in which we use the mean of each feature of the train-set as the input feature values of the models.

Table 5.12 (two leftmost columns) shows the results. The percentage of museums with an accurate prediction (MAPE<10) is low in both cases ($\approx$ 15%) in museums. Regarding good predictions (i.e., MAPE<25), using the last month as the input has the same results as using the mean of 4-months features (37% in museums in both cases).

Table 5.12: Scarcity in seasonal historical and recent data - Evaluation of performance of recency and seasonality features in 27 Museums in the U.K.

	Only Recency		Only Seasonality	
MAPE	last month case	mean of 4-months case	unavailable last 4 months	unavailable last year
MAPE<10	14.81%	14.81%	44.44%	29.63%
MAPE<25	37.00%	37.00%	81.48%	74.00%

Only Seasonality - scarcity in recent data: Likewise the recency features, we analyze the performance of seasonality features when we do not have the most recent data available. This may happen in cases when data collection is periodical (or seasonal) and lasts longer periods and the most recent data is not yet available for prediction. In this scenario, we can adopt seasonality features, i.e. number of visits, social media and environmental data in the previous years in order to predict the future visitation, if this information is available. Likewise outdoors, for this, we do not use the most recent historical data of each attraction and use only the remaining historical data for training the prediction model. For constructing the training-set, we define two cases regarding the unavailability of historical data: (i) unavailable history of the last 4 months of each feature; (ii) unavailable last 12 months (last year) of each feature. The first case corresponds to the situation where we do not have the previous last 4 months (y-1, y-2, y-3, y-4) while the second case is when we do not have one complete cycle of historical data (annual seasonality) [Rosselló and Sansó, 2017].

Table 5.12 (two rightmost columns) presents the results for indoor attractions. The percentage of museums with an accurate prediction (MAPE<10) is much higher in the first case when only the last 4 months of the historical data is unavailable in comparison to the case when the complete historical data of the last year is missing (44% versus 30%). These results again suggest the importance of the historical data. In other words, similarly to outdoors, having the latest trends of visitations besides the periodical/historical behaviors is essential for an accurate prediction in indoor attractions.

5.3.4 Factorial analysis of tourism prediction requirements

We investigate the impact of each of tourism prediction requirements by means of a factorial design analysis. However, this time the response variable is the % of indoors attractions that fall in each MAPE range and we want to estimate the importance of each factor (interaction) on the variation observed in those % of touristic attractions. Again, when all three requirements are turned off, we use the global SVR model (non-specialized model trained for all indoors attractions using only the Environmental and Social media features, i.e. absence of all three factors).

Table 5.13: Contribution of each of tourism prediction requirements: recency, seasonality, model specialization and their interactions into the response variable in indoors attractions - museums; results for MAPE $<$ 10 and MAPE $<$ 25 in 5 runs. The contributions higher than 5% are in **bold** face.

	contribution (%)	
Requirements	MAPE $<$ 25	MAPE $<$ 10
Recency	1.0	0
Seasonality	**21.1**	**19.8**
Model spec.	**55.9**	**70.0**
Recency, Seasonality	**5.5**	1.8
Recency, Model spec.	0.1	1.3
Seasonality, Model spec.	**13.3**	2.9
Recency, Seasonality, Model spec.	0.7	1.6
Residuals	2	3

In Table 5.13, we show the variation explained by each tourism requirement on the prediction results in indoors attractions. We observe that model specialization and then seasonality have the largest contributions. In the case of MAPE < 25%, we have also a significant contribution of the interaction between these two factors - seasonality and Model specialization (13.3%).

In addition, we observe that model specialization, in relative terms, is more important to the variation observed for MAPE < 10% than for the results for MAPE < 25% (explains 70% versus 56% of result variation for Museums) again indicating that if we need highly accurate prediction results (MAPE < 10%), the use of specialized models becomes even more important.

We can also see in Table 5.13 that the impact of recency and its interactions with other factors on the prediction results are almost negligible. Despite that, recency can improve results (look for instance at the second and third columns in Table 5.11), indicating that we should use it, mainly if the seasonality features are not available.

Likewise outdoors attraction, seasonality alone has more than 20% of contribution for MAPE < 25%. This again indicates that when we have only the historical data for an attraction, we can significantly improve accuracy by injecting seasonality features into the model as input variables.

5.3.5 Drill down analysis of encapsulated features in recency and seasonality factors

Here we quantify the impact of each of the tourism prediction requirements, recency and seasonality features in the prediction accuracy. We will do so by analyzing the learned coefficients of the **global models** in the indoor scenarios. As can be seen in Tables 5.10 (indoors), the impact of the incorporation of the recency and seasonality features into the global models is similar to that of the specialized models, with significant improvements over the case in which we do not use such features, for MAPE < 10% and MAPE < 25%, although results are not as good as with the latter.

Table 5.14 shows the learned coefficients of global models in indoor scenarios, respectively. Likewise the outdoors attractions, for this analysis, we built global models for all attractions of each type, adopting each time a different feature-set: (I) soc + env: global model trained having only social media and environmental features in the feature-set; (II) recency (soc + env + rec): global model having recency features in addition to the social media and environmental features; (III) seasonality (soc + env + seas): global model having seasonality features in addition to the social media and environmental features and; (IV) seasonality+recency (soc + env + rec) + seas): global model having all features including social media, environmental, recency and seasonality features.

Looking at the learned coefficients in the table, indicate the high importance of number of reviews and then maximum temperature in the simplest model. In the recency model (soc + env + rec), instead, higher weights are given to the number of visits in the last two months (y-1 and y-2 features). The number of visits in the last year (y-12) and in 15 months before (y-15) are more relevant when seasonality is incorporated into the model (soc + env + seas). Finally, visits in the last year and in the last month (y-12 and y-1) contribute more to the accuracy of the complete model (soc + env + rec + seas). Interestingly, the impact of visits in the last year, same period (y-12) has a larger weight than visits in the last month (y-1) which is aligned with what we observed in the factorial analysis of the impact of tourism prediction requirements – seasonal features have more contribution to the model than recency ones.

5.3.6 Identification of anomaly cases and improving their accuracy with tourism prediction requirements

In this section, we first inspect attractions (museums and galleries) for which our best prediction technique – SVR using external data features did not result in good predictions (i.e., MAPE $>$ 25%). We identify and discuss the possible reasons and the effects that caused our proposed prediction methodology – first tourism prediction architecture in Figure 3.3 –

Table 5.14: The coefficients of features of global (single) model for all 27 U.K Museums adopting each time a different set of features: (I) only social media and environmental features (soc+env), (II) social media, environmental and recency features (soc+env+rec), (III) social media, environmental and seasonality features (soc+env+seas), (IV) complete feature set: social media, environmental, seasonality and recency feature (soc+env+rec+seas). The **bold** face shows the top 2 features in each column.

Features	soc+env	soc+env+rec	soc+env+seas	soc+env+rec+seas
tmin	-0.093	0.025	-0.004	-0.031
tavg	0.024	0.005	-0.001	0.001
tmax	**0.116**	-0.011	0.002	0.026
air_frost_days	0.004	0.005	0.000	0.008
rain	-0.022	0.006	0.003	0.018
sunny_hr	-0.037	0.020	0.004	0.022
revs	**0.517**	0.010	0.004	0.002
rating	-0.051	-0.001	0.000	-0.004
month	-0.007	-0.060	-0.002	-0.026
y-1	-	**0.511**	-	**0.407**
y-2	-	**0.258**	-	0.158
y-3	-	0.156	-	0.033
y-4	-	0.054	-	0.026
log y-1	-	0.025	-	0.089
log y-2	-	-0.003	-	-0.033
log y-3	-	-0.032	-	-0.015
log y-4	-	0.003	-	-0.017
y-12	-	-	**0.764**	**0.658**
y-13	-	-	0.038	-0.291
y-14	-	-	0.084	-0.051
y-15	-	-	**0.089**	0.034
log y-12	-	-	0.011	-0.028
log y-13	-	-	-0.002	-0.027
log y-14	-	-	-0.003	0.036
log y-15	-	-	-0.007	0.003

to obtain such unsatisfactory results. Then, in continuous, we reveal the prediction models accuracy for those inaccurate cases adding key tourism prediction requirements as explicit features into input feature-set.

For 2 out of the 27 U.K. museums and galleries considered, the prediction accuracy of SVR using only social media and environmental features was not satisfactory. These attractions are listed in Table 5.15 along the anomalous pattern in the number of visits, the corresponding effect on the number of TripAdvisor reviews and a possible reason that explain such anomaly. In one such case, the Royal Armouries Fort Nelson museum, both the number of visits and the number of TripAdvisor reviews experienced a huge increase in April 2018, possibly due to a campaign motivating visitors to visit a new sculpture that was opened to the public that month.

Table 5.15: Museums and Galleries Failure Analysis

Museums/ Gallery	MAPE (%)	type	Anomaly	Effect on #Reviews	Possible Reason
Royal Armouries Fort Nelson	55.44	Museum	Huge increase (over 350%) in the number of visits by April 2018 w.r.t the previous years in the same period. However, using only social media number of reviews and average ratings (and not the content of reviews) and environmental features we could not predict this anomaly.	similarly to the visits, a huge increase in the number of social media reviews w.r.t to the same period of the previous years; interestingly, the social media could respond quickly to the anomaly in the same manner	According to the official News [4] and social media comments, the **visitors were motivated to visit the ionic poppy sculpture 'wave'** has been opened to public in April 2018; there are tens of comments in this period using words like: 'A must visit to see the poppies', 'amazing wave', 'poppy visit', spectacular wave', etc.
U.K. National Portrait Gallery	137.90	Gallery	in this case anomaly was in #Reviews	Major increase in the number of social media reviews by April 2015	Based on the annual report published by National Portrait Gallery[5], the audience grew on a national and international level during 2015/16 with an **increased number of people having access to exhibitions**, displays and the collection online, and **through the gallery's website**.

As this event (in April 2018) was in our test period, we could not capture the anomalous increase in number of visitors. Perhaps a more successful approach would benefit from exploiting the contents of the TripAdvisor reviews. The other attraction, U.K. National Portrait Gallery, experienced a complete mismatch in the trends of number of visits and number of reviews by April 2015; while the former maintained the historical pattern, the latter exhibited a major increase, possibly due to a reportedly increase in the number of people having access to exhibitions through the gallery's website.

To further illustrate these anomalous patterns, Figure 5.4 shows the time series of visitation and reviews for the National Portrait Gallery. Note the major increase in the number of social media reviews starting in April 2015 and lasting until roughly September 2016. Given the great importance of social media features to prediction accuracy for indoor attractions, such deviant behavior caused the SVR model to make predictions that were very off. The same was also observed for Linear Regression. However, on the other hand, SARIMAX and SARIMA, managed to be robust to such anomalous patterns, resulting in very good prediction results (MAPE of 12.24 and 10.48 respectively). This is because regression methods (SVR and Linear Regression) do not deal properly with temporal dependencies among data observations in a time-series variable nor give more weights to recent observations (last lags). ARIMA-based models, on the other hand, manage to maintain a good performance by exploiting an extra feature (history of visits) and give more weights to the recent observations for the visits.

[4] http://www.farnhamherald.com/article.cfm?id=126532&headline=WaveofpoppiesopensatFortNelson&sectionIs=news

[5] https://www.npg.org.uk/assets/files/pdf/accounts/npgaccounts2015-16.pdf

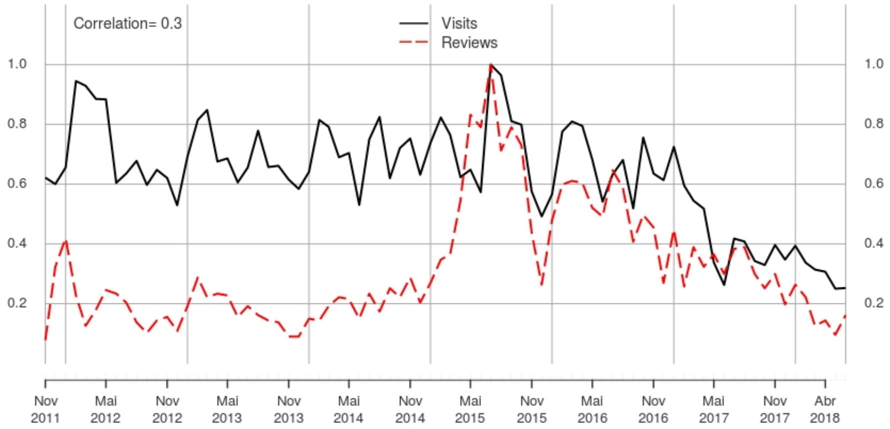


Figure 5.4: Temporal evolution of number of visits and social media comments in National Portrait Gallery in the U.K. (the values in y-axis are normalized by the maximum value of each variable in the whole period).

Now, we evaluate whether the incorporation of seasonality and recency features into the specialized models adopting second tourism architecture presented in Chapter 3, Figure 3.3 can help to mitigate the discussed anomalies in these attractions. We analyse the National Portrait Gallery in the U.K. as an example.

In the case of National Portrait Gallery, as we discussed, the social media reviews had a non-typical major increase by April 2015 but there was a gradual decrease in the number of visits (Figure 5.4). This atypical behaviour could be explained considering the annual report published by National Portrait Gallery[6], informing that the virtual audience grew on a national and international level during 2015/16 with an increased number of people having access to exhibitions, displays and the collection **online** through the gallery's website. As a result, we observed more social media activity but less in-site visitations. The incorporation of the recency and seasonality features helped to detect this behavior change and consequently improved the model accuracy, i.e. 90% reduction in the prediction error (a reduction of 137% mean percentage error to 13%). (Table 5.16).

Table 5.16: accuracy of difficult cases incorporating explicit tourism prediction requirements in indoor attractions

Attraction	MAPE - previous results (SpecES)	MAPE - SpecES+recency+seasonality
U.K. National Portrait Gallery (indoor)	137.90%	**13.34%**

All in all, regarding the tourism prediction architectures which were presented in Chap-

[6] https://www.npg.org.uk/assets/files/pdf/accounts/npgaccounts2015-16.pdf

ter 3, we observe that proposal of second architecture (Figure 3.4) is really relevant when we consider the improvement of results over first methodology (Figure 3.3) not only in scenario of indoors (as we showed in above) but also in outdoors (as we discussed in section 4.3.6).

5.4 Clustering of Attractions

As for the outdoor attractions, we analyse the similarity of indoor attractions using clustering algorithms in order to build a single prediction model for each group/cluster of attractions. As before, we define clusters of attractions running k-means and DB-Scan along with PCA analysis for better clustering visualization. When Using k-means for museums and galleries, the total number of clusters *k* that maximizes the silhouette quality measure was k = 3. Figure 5.5 depicts the silhouette metric varying k from 2 to 15. It can be seen that after k = 3, there is another peak in the Figure when k = 6 with a high silhouette score, making k= 6 another possible candidate.

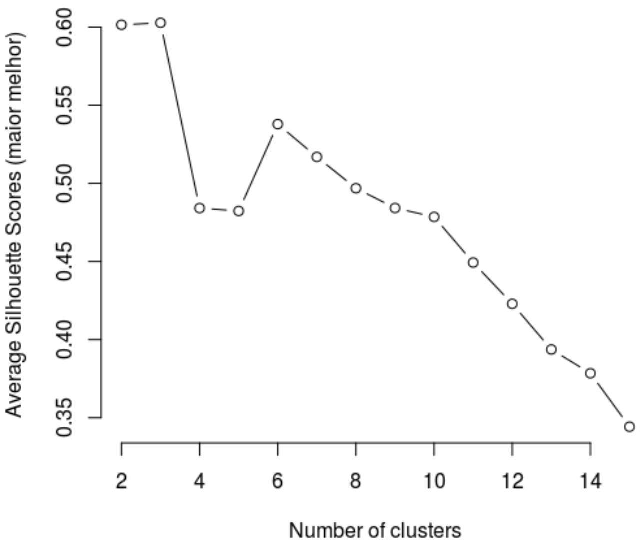


Figure 5.5: Average silhouette score for different numbers of clusters using k-means clustering algorithms for 27 indoor attractions

Figure 5.6 plots the distribution of museums in each cluster (k = 3 in the left side and k = 6 in the right side). In both scenarios, two principal components (PC1 and PC2) explain more than 92% (PC1 + PC2 = 84.72 + 7.53) of variations in clusters. However, clustering the museums in 6 clusters just breaks cluster 1 (when k = 3) into two new clusters: 4 and

5. At the end of this section, we will discuss the number of clusters k that obtained the best prediction results for indoor attractions.

We further analyse the results of clustering indoor attractions using the DB-scan technique. Figure 5.7 presents the distribution of museums after clustering these attractions with k = 3 clusters (cluster 1 is noise, marked as red points in the plot). It seems that DB-scan could not separate the indoor attractions well, having PC1 + PC2 = 75%. All in all, it seems that k-means was more successful in the clustering task having a larger principal component variation explained (about 92%) with respect to DB-scan (about 75%).

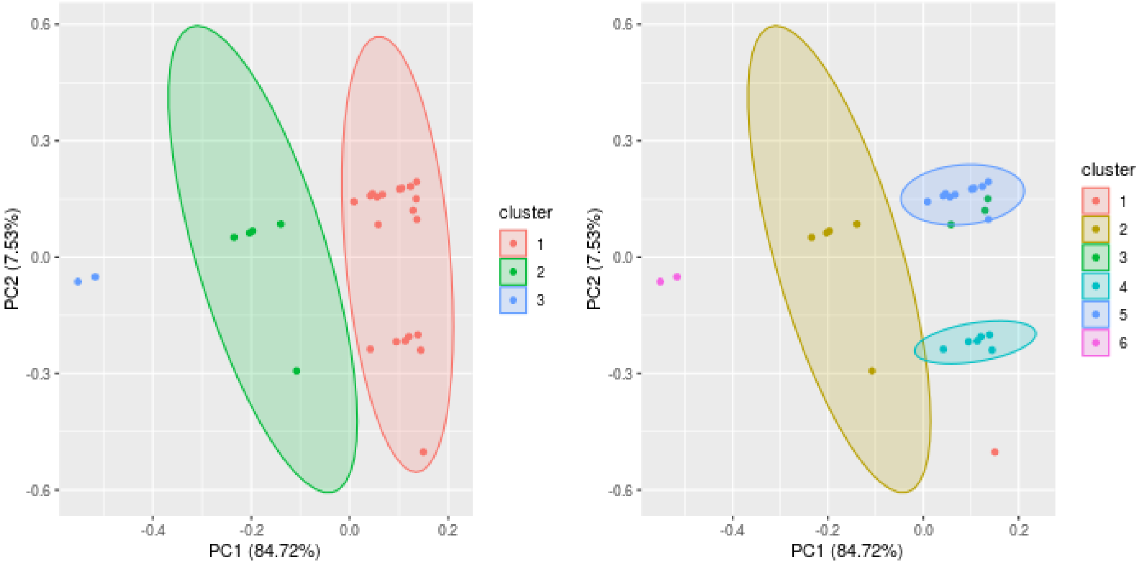


Figure 5.6: Clustering of museums using k-means (k = 3 left and k=6 in right) clustering algorithms adopting social media, environmental, recency and seasonality features

Considering different results for clustering indoor attractions using k-means and DB-scan, we present the results after building a prediction model for each cluster of attractions and use that model to predict the visits of attractions in that cluster. Table 5.17 shows the error accuracy for different number of clusters in indoor attractions. Notice that k = 1 reduces the prediction model to the previous global model for all attractions (i.e., GloES+recency+seasonality model presented in Section 5.3.2) and k = 27 is equivalent to specialized model for each attraction (same as SpecES+recency+seasonality model shown in Section 5.3.2). We observe that clustering the indoor attractions in 3 clusters and running a model for each cluster did not change the percentage of attractions with highly accurate results (MAPE < 10%) in comparison to the Global model. However it reduced the percentage of attractions with good prediction results (MAPE < 25%) from 74% (one global model for all indoor attractions) to 52% (k = 3 clusters). On the other hand, clustering the attractions into 6 clusters, increased the percentage of highly accurate predictions (MAPE < 10%) to

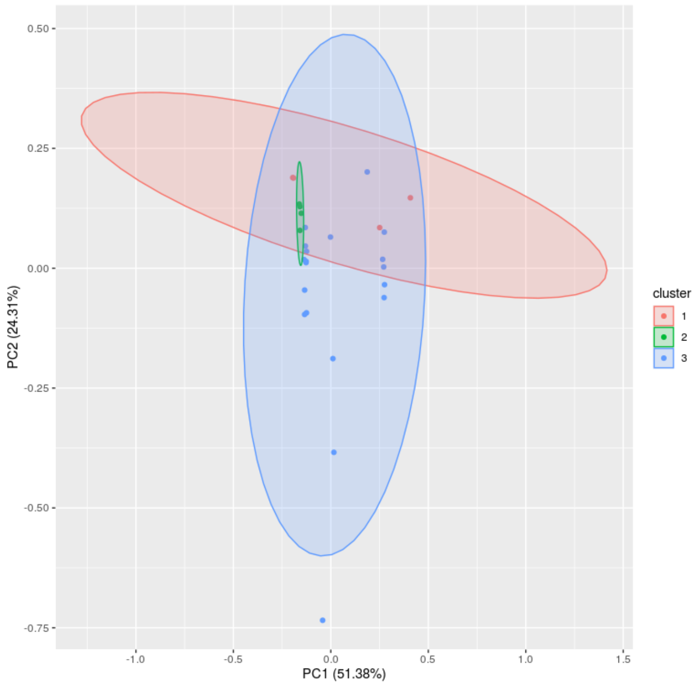


Figure 5.7: Clustering of museums using k-means clustering algorithms adopting social media, environmental, recency and seasonality features

11% , i.e. around 4% better than using a global model. Results for MAPE < 25% remain basically the same in this scenario.

All in all, we observe that by increasing the number of clusters k, the models become more specialized for a group of attractions. Consequently, the percentage of attractions with the MAPE < 10% (very accurate results) increases but this is not always the case for results with MAPE < 25%. In other words, clustering the indoors in 6 clusters using k-means offers better results w.r.t the global model (MAPE < 10%) and equal results for MAPE < 25%. But this is not enough to surpass the use of a single model per attraction.

Table 5.17: Percentage of Museums and Galleries in each interval of MAPE using different number of clusters in order to cluster indoor attractions

No of clusters	MAPE<10%	MAPE<25%
k = 1 (Global model)	7.4	74
k = 3	7.4	52
k = 6	11.11	74
k = 27 (Specialized models)	48.15	93

5.5 Summary and Remarks in indoor attractions

In this Chapter, we presented a similar analysis to that performed in the previous Chapter; this time for indoor attractions. Interestingly, in some cases we observed different results and behaviours in the performance of the features and tourism requirements, possibly due to idiosyncratic characteristics of each type of attraction. We presented the results we obtained with respect to the research goals of this dissertation, demonstrating that by explicitly exploiting three key requirements of tourism prediction as explicit features in our models, besides feeding the models with robust external features – social media and environmental features – we can greatly improve prediction accuracy. We can summarize our main contributions towards answering our research questions as follows:

- We exploited external data – social media features along with environmental data – to forecast touristic demand at 27 Museums and Galleries in the U.K. We computed correlation of number of visits with social media and environmental features collected from different sources. Accurate prediction results ($MAPE < 25\%$) were achieved for more than 93% of the indoor attractions using only external features and SVR prediction model. These results again support our hypothesis in **RQ1** of high predictability when jointly exploiting social media and environmental features as a tool for tourism demand forecasting for places and attractions, mainly for the scenarios with unavailable official visitation census.

- We investigated the impact of exploiting recency and seasonality features alongside social media and environmental data to improve the performance of specialized prediction models for indoor touristic attractions. Our experiments addressed **RQ2** by providing evidence that adopting tourism demand prediction requirements – recency, seasonality and model specialization – can increase the accuracy of forecasting touristic demand at fine-grained levels in comparison to results obtained with only external features (absence of recency and seasonality features). Indeed, the exploitation of a complete set of key tourism requirements outperformed the best baselines results for high accuracy cases ($MAPE < 10\%$) for indoor attractions (48% versus baseline 22%). Furthermore, accurate prediction results ($MAPE < 25\%$) were achieved for more than 96% of the indoor attractions when the SVR prediction model was applied exploiting seasonality and model specialization.

- We analyzed the impact of each of the tourism prediction requirements individually and their interactions applying a 2^k factorial design analysis. Results indicate a higher importance for the model specialization factor (above 55% in $MAPE < 25$), then seasonality (more than 21% of contribution) and its interaction with model specialization (about 15%) and finally recency features with a low importance (less than 2%). However, we emphasized that we should not discard recency features since in attractions without available seasonality features, recency can improve the accuracy of prediction models (models with only social media and environmental features incremented by recency features).

- We experimented with different clustering techniques in order to group similar indoor attractions to build a single prediction model for each cluster. We observed that in the indoor scenario, clustering indoor attractions into 6 clusters using k-means offered better results w.r.t the global model (MAPE < 10%) and equal results for MAPE < 25%.

- We provided an in-depth analysis of indoor attractions for cases for which our proposed prediction methodology was not able to produce satisfactory results. We elaborated on possible causes – they were mostly related to changes in data distributions caused by unpredictable events including: waiving attraction's entrance fees; the availability of an online system to access exhibitions; and boosting visits by exhibiting an ionic poppy sculpture in one of the Museums.

Chapter 6

Conclusions and Future Research Directions

In this chapter we summarize the main achievements of this dissertation. We also present a discussion on future research directions. In mode details, in Section 6.1 we compare our results with previous studies. Next, we describe this dissertation's achievements in Section 6.2 breaking down in one subsection for each research question containing a brief summary of the goal and the obtained results. Finally in Section 6.3, we present a broad discussion and some research directions for future work.

6.1 Comparison with Previous Work

There are three main prior studies that can be somewhat compared to ours in one way or another. In [Fisichelli et al., 2015], the authors analyze the climate and visitation data for U.S. national parks using a third-order polynomial temperature model and argue that it explains 69% of the variation in historical visitation trends. In Chapters 4 and 5, we showed that by jointly exploiting social media and environmental data, even a simple linear regression model can produce good prediction results for about 70% of the outdoor attractions and 63% of indoor attractions while a more robust algorithm such as SVR produces good-to-high prediction results for more than 93% of the attractions (both indoor and outdoors).

In [Khadivi and Ramakrishnan, 2016], the authors exploited Wikipedia usage trends in order to forecast tourism demand in Hawaii. They show that on average, most of the Wikipedia reading activities occur about 4 to 8 months prior to the trip as the mean decision date for most of the activities are between 4 to 8 months before the actual arrival date. However, they report the accuracy of their prediction results only by RMSE using an auto-regressive exogenous model where the external variable is a Wikipedia usage trend time series. RMSE is a measure of accuracy that should be used to compare forecasting errors of different models for a particular data and not between datasets, as it is scale-dependent [Hyndman and Koehler, 2006]. Although there are interesting statements and results in that work, there is no comparison of prediction models to other baselines nor assessment of the results in a comparable manner.

Finally, in [Spencer A. Wood and Lacayo, 2013], the authors use the geo-tagged photos in Flickr to estimate visitation counts in some recreational sites around the world. They report the relationship between the empirical estimates of mean annual visitor user-days and those derived from photographs (this is best described by a polynomial function with $R^2 = 0.386$). They also claim that categorizing the recreational parks into more specific profiles could improve the correlations. However, they do not report results on such categorization.

All in all, our prediction results are strongly superior to the prior efforts having used a mixture of environmental and social media features employed by a linear kernel SVR model.

6.2 Current Achievements

Recall from Chapter 1, that this dissertation aimed at tackling two main research questions:

- RQ 1: How online social media contents and environmental features influence the accuracy of predicting visits in touristic attractions?
- RQ 2: How recency, seasonality and model specialization influence the accuracy of predicting visits in tourist sites?

In the following we discuss the obtained results for each research goal.

6.2.1 RQ 1: How online social media contents and environmental features influence the accuracy of predicting visits in touristic attractions?

In RQ1, we focused on exploiting external data, i.e. social media and environmental features in order to improve the accuracy of prediction models for touristic attractions. To that end, we investigated this research question in-depth in two scenarios of indoors and outdoors in Chapters 4.2 and 5.2 respectively. We exploited Social Media features along with Environmental data to forecast touristic demand at fine-grained levels such as recreational sites, parks, museums, galleries, etc. We computed correlations on the number of visits with social media and environmental features in the several collected datasets from five different sources, encompassing 76 National Parks in U.S. and 27 Museums and Galleries in England.

As one of our main contributions, all collected datasets are publicly available so that others can replicate our results and produce new insights.

Accurate prediction results (*MAPE* $< 25\%$) for more than 93% of the indoor and outdoor attractions were achieved when SVR prediction models were applied. These results support our hypothesis of high predictability when jointly exploiting social media and environmental features as a tool for tourism demand forecasting for places and attractions, mainly for the scenarios with unavailable official visitation census.

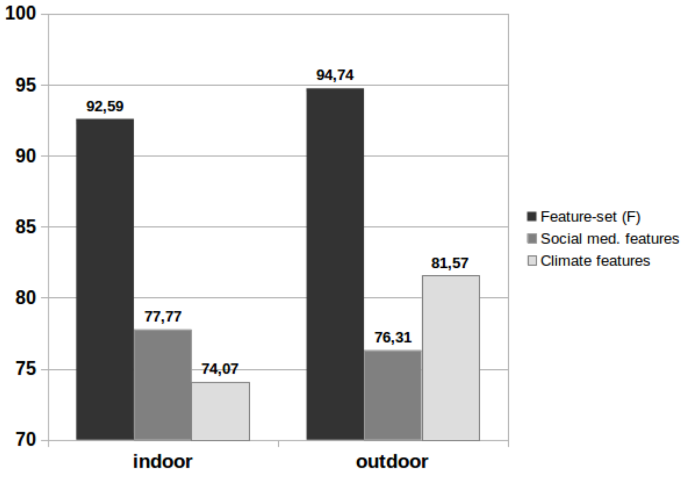


Figure 6.1: Comparison of feature sets for indoor and outdoor attractions (% of attractions with good to high prediction accuracy of SVR).

Figure 6.1 and Table 6.1 summarize our results for indoor and outdoor attractions. The figure shows, side by side, the performance of our best prediction technique – SVR – using only social media features, using only environmental features and using the complete set of features (F). These results are also present in Tables 4.6 and 5.6. Note that the percentages of attractions for which the prediction using only social media features had good-to-high accuracy (MAPE results below 25%) are roughly the same for both types of attractions (77.7% and 76.3% for indoor and outdoor attractions, respectively). However, the environmental features, when used in isolation, have a much better performance in outdoor attractions (81.5% of attractions with MAPE $<$ 25%, as opposed to 74% for indoor attractions). This might be expected given the greater role that weather often has on someone's experience in outdoor locations. We also stress that we here consider only two social media features. Other features, extracted for example from the reviews' textual content, or from other social networks, might contribute to further improve prediction accuracy. In any case, we note that the joint use of both types of features offered a great boost in performance over any individual feature subset, for both indoor and outdoor attractions.

Having a social media feature - number of reviews – as one of the most important features in predicting tourism arrival is not totally unexpected since social media data reflects real world events such as epidemics, economic fluctuations, natural and social disasters or discussions in a very short manner of time. In [Pan et al., 2011], the authors state that the influence of the Internet has been significantly transforming the tourist industry in a number of ways. It has become one of the most efficient means of reaching new tourist markets and is now the leading information source for tourists due to the many online tourism communities it supports [Liu and Park, 2015; Pantano and Di Pietro, 2013]. Moreover, the number of reviews in social media websites not only captures a collective vision of the community for each attraction but it is also less biased than opinions published by public organs or other third parties. According to [Song and Liu, 2017], previous studies on tourism have mostly been based on surveys or experts' views, which have limitations in terms of representativity and scale. Such population samples, based on induced behaviors of general users may not completely reflect the reality or the view of real users who are really interested in touristic attractions. In contrast, studies based on tourism social media data try to capture a more realistic situation, based on real, non-artificially induced behaviors (e.g., likes or comments) using large samples of users.

According to [Song and Liu, 2017] social media data are of utmost importance because of three main characteristics, namely, (a) reliability, (b) new information flows and (c) real time and nowcasting. First, social media is reliable because the produced data are based on users' actual actions, not on surveys. In other words, actual actions such as likes or explicit comments have been analyzed rather than stated intentions or answers to questions. Second, social media is a new source of information flow produced by tourists themselves. It enriches the knowledge of tourism businesses' target market and is very useful for analyzing the consumers' demand for different tourism products and services [Perdana et al., 2014]. Third, social media data is usually real time and allows nowcasting, that is, the use of real-time data to describe contemporary activities before official data sources are made available [Bollier et al., 2010].

Table 6.1 provides a complimentary look on the relative performance of the prediction techniques. We here rank the models based on the number of attractions for which it was the winner approach among all considered techniques, producing the best prediction result (smallest MAPE). As shown in the table, SVR has the best overall performance for both indoor and outdoor attractions, being the winner for 35 (out of 76) outdoor attractions and 14 (out of 26) indoor attractions. SARIMA is the second best option, being the winner method for 15 outdoor and 7 indoor attractions. Unlike SVR, SARIMA performs well only when the response variable (number of visits) is normally well behaved along the period of time with regards to trend and seasonality. SARIMAX and linear regression come in third and fourth in the ranking, whereas GRNN occupies the last position of the ranking.

Table 6.1: Comparison of prediction techniques: number of attractions each method produced the best (lowest) MAPE.

dataset	SVR	Linear Reg.	GRNN	SARIMAX	SARIMA
outdoor	35	13	2	11	15
indoor	14	2	-	4	7

We also provided a deep analysis for the experiments in which our proposed prediction methodology was not able to produce satisfactory results. We elaborated on possible causes and for each of those cases, they were mostly related to changes in data distributions caused by unpredictable events including steam rising from a crater and eruption scare, waiving attraction's entrance fees, renaming a national park, provision of an online system to access exhibitions and boosting visits with an ionic poppy sculpture.

6.2.2 RQ 2: How recency, seasonality and model specialization influence the accuracy of predicting visits in tourist sites?

In RQ2, we focused on analyzing the effects of key tourism prediction requirements, i.e. recency, seasonality and model specialization – on prediction accuracy of fine-grained tourists' visits. Our experimental evaluation confirmed our hypotheses, with observed gains over the previous results. We also showed that the explicit incorporation of such requirements into the models helped to solve very hard-to-solve anomaly cases.

In mode details, we investigated the impact of exploiting recency and seasonality features alongside social media and environmental data to improve the performance of specialized prediction models for touristic attractions (indoor and outdoor). Our experiments showed that adopting tourism demand prediction requirements, i.e. recency, seasonality and model specialization can increase the accuracy of forecasting touristic demand at fine-grained levels such as recreational sites, parks, museums, galleries, etc in comparison to results obtained with only external features (absence of recency and seasonality features). In fact, the use of a complete set of key tourism requirements outperformed the best baselines results for high accuracy cases ($MAPE < 10\%$) for both indoor (48% versus baseline 22%) and outdoor attractions (50% versus baseline 42%). Furthermore, accurate prediction results ($MAPE < 25\%$) were achieved for more than 96% of the indoor and outdoor attractions when SVR was applied exploiting the seasonality and model specialization.

We discussed that recency considers the impact of recent events into the prediction models arguing that temporal aspects should be considered to assess whether and how recent events such as financial crises, new trends, epidemics/ pandemics and new infrastructures may impact the predictions [Moro and Rita, 2016]. On the other hand, seasonality focuses on the inherently cyclic behaviour of tourism demands. Seasonality is one of the main phenomena affecting tourism, corresponding to movements of a variable in a year or from season to season [Hylleberg, 1992]. In addition, we highlighted how model specialization can be advantageous since individual attractions may have very specific idiosyncratic patterns of visitations.

We also analyzed the impact of each of the tourism prediction requirements individually and their interactions applying a 2^k factorial design analysis. The results indicated the higher importance of model specialization factor (about 50% in $MAPE < 25\%$), then seasonality (more than 21% of contribution) and its interaction with model specialization (about 15%) and finally recency features with a low importance (less than 6%). We emphasized that we should not discard recency features since in attractions without seasonality features available, recency features can improve the accuracy of prediction models (models with only social media and environmental features incremented by recency features). We also pointed out that the higher impact of seasonality in outdoor attractions in comparison with indoor attractions (MAPE < 10%) was somewhat expected [Khatibi et al., 2019].

To have a deeper understanding of the impact of the recency and seasonality aspects, we explored how scarcity in historical data – recent and seasonal features – impacted the prediction accuracy of models. We defined two scenarios of (i) only recency – when the seasonal data is not available, and (ii) only seasonality – when no recent information is available for the attractions. The observation was that recent trends of visitation are essential in the accuracy of the models since the absence of the last 12 months of recent data deteriorated a lot the accuracy of the models in comparison to the scenario with absence of the last 4 months. Finally, we showed how explicit incorporation of seasonality and recency features into the specialized models of indoor and outdoor attractions could improve the accuracy of the tourism demand in attractions in which the state-of-the-art models could not provide an accurate prediction.

All in all, we encourage that our **Comprehensive Fine-grained Demand Prediction analysis** can be generalized and applied in many other areas of demand forecasting in which we may have seasonality behavior or effect of recent events in the demand oscillations. In this work, we applied our proposed extensive methodology in the area of Tourism demand only as a use-case in order to quantify effect of different factors in multiple scenarios while we also proposed novel solutions in treating challenges like data scarcity or lack of official data.

6.3 General Discussion and Future Work

Although exploiting data from social media is fascinating, a critical question that will determine their utility for forecasting future visitation is: how well do they reflect on-the-ground visitor surveys and records? In this work, we showed that there is a strong relationship between the number of reviews and visitation field-based records for a large fraction of the attractions, particularly those that are outdoors. This may provide a powerful new tool for forecasting tourism demands, helping tourism accommodations to get prepared even when there is no prior survey for their regions (or one is not even possible), only by using freely available social media data empowered by environmental records. However, correlating environmental and social data with official visits demonstrated to be key to motivate the simplicity of our prediction model.

In addition, the evaluation of our techniques on two different types of attractions, indoors and outdoors, revealed the relative effectiveness (in terms of prediction performance) of each category of features, i.e, social media and environmental features, versus the mixture of both. Our results showed that that social media features (notably number of TripAdvisor reviews) are the most important ones in the case of indoor attractions such as museums and galleries while for the outdoor attractions like national parks, environmental features (notably average temperature) play a more important role. In any case, our experiments clearly show the great benefits of combining both classes of features for both types of attractions.

Furthermore, in this work, we evaluated the predictions on a monthly basis. We selected this granularity of time since the official ground-truth data was available and aggregated at this level. However some preliminary experiments suggest that there is a strong possibility of successfully applying the same methodology on a finer granularity of time such as weekly, daily and even hourly. This aspect will be further explored in future work.

In future work, we intend to continue improving accuracy, mainly of highly accurate predictions (MAPE < 10), by evaluating the contents and sentiments of the reviews of each attraction. We intend to apply text analysis techniques such as Temporal, Semantic and Hierarchical Topic Modeling [Viegas et al., 2018, 2020b] and Sentiment Analysis[Viegas et al., 2020a; Canuto et al., 2016] in order to extract useful information from visitors daily reviews and their possible visiting behaviour trends. We also intend to better comprehend the results of our clustering exercise aiming at improving them. If successful it could make it simpler and more practical to use our solutions in the real life of business owners. Clustering could also produce more robust forecasting models for touristic places with low availability of visitation census, due to multiple reasons such as high costs of surveys or difficulty to collect data in remote places.

Bibliography

Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *WI-IAT*, volume 1, pages 492--499.

Bollier, D., Firestone, C. M., et al. (2010). *The promise and peril of big data*. Aspen Institute, Communications and Society Program Washington, DC.

Borras, J., Moreno, A., and Valls, A. (2014). Intelligent tourism recommender systems: A survey. *Expert Systems with Applications*, 41(16):7370--7389.

Box, G. E. and Jenkins, G. M. (1976). Time series analysis: Forecasting and control san francisco. *Calif: Holden-Day*.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5--32. ISSN 1573-0565.

Butler, R. (2001). Seasonality in tourism: Issues and implications. *Seasonality in tourism*, pages 5--21.

Cankurt, S. and Subasi, A. (2015). Developing tourism demand forecasting models using machine learning techniques with trend, seasonal, and cyclic components. *Balkan Journal of Eletrical and Computer Engineering*, Vol.3, No.1.

Canuto, S. D., Gonçalves, M. A., and Benevenuto, F. (2016). Exploiting new sentiment-based meta-level features for effective sentiment analysis. In Bennett, P. N., Josifovski, V., Neville, J., and Radlinski, F., editors, *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*, pages 53--62. ACM.

Chang, J., Wall, G., and Chu, S.-T. T. (2006). Novelty seeking at aboriginal attractions. *Annals of Tourism Research*, 33(3):729--747.

Chang Jui Lin, Hsueh Fang Chen, T. S. L. (2011). Forecasting tourism demand using time series, artificial neural networks and multivariate adaptive regression splines: Evidence from taiwan. *Int. J. Business Administration*, 2.

Chetty, P. (2011). Advantages of demand forecast for the tourism industry. *projectguru*.

Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *KDD*, pages 1082--1090.

Chunara, R., Andrews, J. R., and Brownstein, J. S. (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *Amer. J. of Tropical Medicine and Hygiene*, 86(1):39--45.

Clark, V., Dunn, O., and Mickey, R. (1974). *Applied statistics, analysis of variance and regression*. Wiley.

Cuccia, T. and Rizzo, I. (2011). Tourism seasonality in cultural destinations: Empirical evidence from sicily. *Tourism Management*, 32(3):589--595.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V. (1996). Support vector regression machines. In *NIPS*, pages 155--161.

Ferrari, L., Rosi, A., Mamei, M., and Zambonelli, F. (2011). Extracting urban patterns from location-based social networks. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 9--16. ACM.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1:3--32.

Fisichelli, N. A., Schuurman, G. W., Monahan, W. B., and Ziesler, P. S. (2015). Protected area tourism in a changing climate: Will visitation at us national parks warm up or overheat? *PLoS ONE*, 10(6).

Frechtling, D. (2012). *Forecasting tourism demand*. Routledge.

Gautheir, T. D. (2001). Detecting trends using spearman's rank correlation coefficient. *Environmental forensics*, 2(4):359--362.

Georgiev, P., Noulas, A., and Mascolo, C. (2014). Where businesses thrive: Predicting the impact of the olympic games on local retailers through location-based services data. *ICWSM*.

Goeldner, C. R. and Ritchie, J. R. B. (2006). Tourism principles, practices, philosophies. *John Wiley and Sons Inc., New Jersey*.

Hasan, S., Zhan, X., and Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *KDD*, page 6.

Hengyun Li, H. S. and Li, L. (2016). A dynamic panel data analysis of climate and tourism demand: Additional evidence. *Journal of Travel Research*, I-14.

Herr, D. G. (1986). On the history of anova in unbalanced, factorial designs: The first 30 years. *The American Statistician*, 40(4):265–270.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.

Höpken, W., Eberle, T., Fuchs, M., and Lexhagen, M. (2018). Google trends data for analysing tourists' online search behaviour and improving demand forecasting: the case of åre, sweden. *Information Technology & Tourism*. ISSN 1943-4294.

Hossain, N., Hu, T., Feizi, R., White, A. M., Luo, J., and Kautz, H. (2016). Inferring fine-grained details on user activities and home location from social media: Detecting drinking-while-tweeting patterns in communities. *arXiv preprint arXiv:1603.03181*.

Huang, X., Zhang, L., and Ding, Y. (2017). The baidu index: Uses in predicting tourism flows–a case study of the forbidden city. *Tourism management*, 58:301--306.

Hylleberg, S. (1992). *Modelling seasonality*. Oxford University Press.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679--688.

Iebeling Kaastra, M. B. (1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing - Elsevier*, 10:215–236.

Jain, R. e. a. (1991). A test of goodness of fit. *The Art of Computer Systems Performance Analysis: techniques for experimental design, measurement, simulation, and modeling*, 49(268):765--769.

Jin, X. and Han, J. (2010). *K-Means Clustering*, pages 563--564. Springer US, Boston, MA.

Jolliffe, I. and Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374:20150202.

Kamel, N., Atiya, A. F., El Gayar, N., and El-Shishiny, H. (2008). Tourism demand forecasting using machine learning methods. *ICGST International Journal on Artificial Intelligence and Machine Learning*, 8:1--7.

Khadivi, P. and Ramakrishnan, N. (2016). Wikipedia in the tourism industry: Forecasting demand and modeling usage behavior. In *ICWSM*, pages 4016--4021.

Khatibi, A., Belém, F., da Silva, A. P. C., Almeida, J. M., and Gonçalves, M. A. (2019). Fine-grained tourism prediction: Impact of social and environmental features. *Information Processing & Management*, page 102057.

Khatibi, A., Belém, F., Silva, A. P., Shasha, D. E., and Gonçalves, M. A. (2018). Improving tourism prediction models using climate and social media data: A fine-grained approach. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 636--639. AAAI Press.

Khatibi, A., Silva, P. C. d. A., Almeida, M. J., and Gonçalves, A. M. (2020). Fisetio: A fine-grained, structured and enriched tourism dataset for indoor and outdoor attractions. *Data in Brief - Information Processing & Management*.

Koenig-Lewis, N. and Bischoff, E. E. (2005). Seasonality research: The state of the art. *International Journal of Tourism Research*, 7(4-5):201--219.

Koutras, A., Panagopoulos, A., and Nikas, I. A. (2017). Forecasting tourism demand using linear and nonlinear prediction models. *Academica Turistica-Tourism and Innovation*, 9(1).

Kulendran, N. and Wong, K. K. (2005). Modeling seasonality in tourism forecasting. *Journal of Travel Research*, 44(2):163--170.

Leask, A., Fyall, A., and Barron, P. (2014). Generation y: An agenda for future visitor attraction research. *International Journal of Tourism Research*, 16(5):462--471.

Lewis, C. D. (1982). *Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting*. Butterworth-Heinemann.

Li, H., Song, H., and Li, L. (2017). A dynamic panel data analysis of climate and tourism demand: Additional evidence. *Journal of Travel Research*, 56(2):158--171.

Li, N. and Chen, G. (2009). Analysis of a location-based social network. In *Computational Science and Engineering*, volume 4, pages 263--270.

Li, Y. and Cao, H. (2018). Prediction for tourism flow based on lstm neural network. *Procedia Computer Science*, 129:277--283.

Lim, K. H., Chan, J., Leckie, C., and Karunasekera, S. (2018). Personalized trip recommendation for tourists based on user interests, points of interest visit durations and visit recency. *Knowledge and Information Systems*, 54(2):375--406.

Liu, Z. and Park, S. (2015). What makes a useful online review? implication for travel product websites. *Tourism Management*, 47:140--151.

Maditinos, Z. and Vassiliadis, C. (2008). Crises and disasters in tourism industry: happen locally, affect globally. In *MIBES*, pages 67--76.

Martín, M. B. G. (2005). Weather, climate and tourism a geographical perspective. *Annals of tourism research*, 32(3):571--591.

Moro, S. and Rita, P. (2016). Forecasting tomorrow's tourist. *Worldwide Hospitality and Tourism Themes*, 8(6):643--653.

Nyamen Tato, A. A. and Nkambou, R. (2018). Improving adam optimizer. *Workshop track -ICLR*.

Pai, P.-F., Hung, K.-C., and Lin, K.-P. (2014). Tourism demand forecasting using novel hybrid system. *Expert Systems with applications*, 41(8):3691--3702.

Pan, B., Xiang, Z., Law, R., and Fesenmaier, D. R. (2011). The dynamics of search engine marketing for tourist destinations. *Journal of Travel Research*, 50(4):365--377.

Pantano, E. and Di Pietro, L. (2013). From e-tourism to f-tourism: emerging issues from negative tourists' online reviews. *Journal of Hospitality and Tourism Technology*, 4(3):211--227.

Parmezan, A. R. S. (2016). *Predição de séries temporais por similaridade*. PhD dissertation, Universidade de São Paulo, São Carlos.

Perdana, D. H. F. et al. (2014). Trip guidance: A linked data based mobile tourists guide. *Advanced Science Letters*, 20(1):75--79.

Petrevska, B. (2013). Empirical analysis of seasonality patter ns in tourism. *Journal of Process Management. New Technologies*, 1(2):87--95.

Richards, G. (2002). Tourism attraction systems: Exploring cultural behavior. *Annals of tourism research*, 29(4):1048--1064.

Rosselló, J. and Sansó, A. (2017). Yearly, monthly and weekly seasonality of tourism demand: A decomposition analysis. *Tourism Management*, 60:379--389.

Sander, J. (2010). *Density-Based Clustering*, pages 270--273. Springer US, Boston, MA.

Shokoohi-Yekta, M., Chen, Y., Campana, B., Hu, B., Zakaria, J., and Keogh, E. (2015). Discovery of meaningful rules in time series. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1085--1094.

Song, H. and Liu, H. (2017). Predicting tourist demand using big data. In *Analytics in smart tourism design*, pages 13--29. Springer.

Specht, D. F. (1991). A general regression neural network. *IEEE transactions on neural networks*, 2(6):568--576.

Spencer A. Wood, Anne D. Guerry, J. M. S. and Lacayo, M. (2013). Using social media to quantify nature-based tourism and recreation. *Scientific Report*, 3.

Stainton, D. H. (2021). Types of tourist attractions - understanding tourism.

Steenjacobsen, J. K. (2001). Nomadic tourism and fleeting place encounters: exploring different aspects of sightseeing. *Scandinavian Journal of Hospitality and Tourism*, 1(2):99--112.

Valverde-Rebaza, J. C., Roche, M., Poncelet, P., and de Andrade Lopes, A. (2018). The role of location and social strength for friendship prediction in location-based social networks. *Inf. Process. Manage.*, 54(4):475--489.

Vasconcelos, M., Almeida, J. M., and Gonçalves, M. A. (2015). Predicting the popularity of micro-reviews: a foursquare case study. *Information Sciences*, 325:355--374.

Vecchio, P. D., Mele, G., Ndou, V., and Secundo, G. (2018). Creating value from social big data: Implications for smart tourism destinations. *Inf. Process. Manage.*, 54(5):847--860.

Viegas, F., Alvim, M. S., Canuto, S. D., Rosa, T., Gonçalves, M. A., and da Rocha, L. C. (2020a). Exploiting semantic relationships for unsupervised expansion of sentiment lexicons. *Inf. Syst.*, 94:101606.

Viegas, F., Cunha, W., Gomes, C., Pereira, A., da Rocha, L. C., and Gonçalves, M. A. (2020b). Cluhtm - semantic hierarchical topic modeling based on cluwords. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8138--8150. Association for Computational Linguistics.

Viegas, F., Luiz, W., Gomes, C., Khatibi, A., Canuto, S. D., Mourão, F., Salles, T., da Rocha, L. C., and Gonçalves, M. A. (2018). Semantically-enhanced topic modeling. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 893--902. ACM.

Volchek, K., Liu, A., Song, H., and Buhalis, D. (2018). Forecasting tourist arrivals at attractions: Search engine empowered methodologies. *Tourism Economics*, page 1354816618811558.

Wang, C.-H. (2004). Predicting tourism demand using fuzzy time series and hybrid grey theory. *Tourism Management-Elsevier*, 25:367–374.

Wang, F., Franco-Penya, H.-H., Kelleher, J., Pugh, J., and Ross, R. (2017). An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity.

Webber, A. G. (2001). Exchange rate volatility and cointegration in tourism demand. *Journal of Travel research*, 39(4):398–405.

Westcott, M. and Wendy Anderson, E. (2019). Introduction to tourism and hospitality in bc-introduction to tourism and hospitality in bc.

Xiaoxuan, L., Qi, W., Geng, P., and Benfu, L. (2016). Tourism forecasting by search engine data with noise-processing. *African Journal of Business Management*, 10(6):114--130.

Yeh, C.-C. M., Kavantzas, N., and Keogh, E. (2017). Matrix profile iv: using weakly labeled time series to predict outcomes. *Proceedings of the VLDB Endowment*, 10(12):1802--1812.

Zhang, C. and Zhang, J. (2011). Neural network ensemble for chinese inbound tourism demand prediction [j]. *Scientia Geographica Sinica*, 10:009.

Zhu, Y., Imamura, M., Nikovski, D., and Keogh, E. (2019). Introducing time series chains: a new primitive for time series data mining. *Knowledge and Information Systems*, 60(2):1135--1161.

Zhu, Y., Imamura, M., Nikovski, D., and Keogh, E. J. (2018). Time series chains: A novel tool for time series data mining. In *IJCAI*, pages 5414--5418.