Information Technology & Tourism (2019) 21:105–133
https://doi.org/10.1007/s40558-018-0135-6

ORIGINAL RESEARCH

What is the "Personality" of a tourism destination?

Mete Sertkan[1] · Julia Neidhardt[1] · Hannes Werthner[1]

Received: 1 May 2018 / Revised: 9 November 2018 / Accepted: 29 November 2018 /
Published online: 10 December 2018

Abstract
Researching online before booking a vacation can be seen as a common habit of customers nowadays. In this context, Recommender Systems (RSs) are aiming to support the customers to find the right products, but they face domain specific challenges since tourism products are typically very complex and related to emotional experiences. To counteract these challenges, comprehensive user models for capturing the preferences and personality of travellers have been introduced. One of these models is the so-called Seven-Factor Model. This paper introduces an automated way for determining the Seven-Factor representation of tourism destinations to enable a matchmaking for RSs. In particular, exploratory data analysis, cluster analysis, and regression analysis are conducted not only to find a mapping of destinations onto the Seven-Factors, but also to foster a better understanding of the relationship between destination attributes and the Seven-Factors. This work provides clear evidence that a tourism destination's Seven-Factor representation can be determined by taking its attributes into account. This is an extended version of a conference paper entitled "Mapping of tourism destinations to travel behavioural patterns" previously published in the proceedings of Information and Communication Technologies in Tourism 2018 Conference (ENTER 2018) held in Jönköping, Sweden, January 24–26, 2018.

Keywords User modelling · Tourism recommender systems · Tourism destinations · Statistical analysis · Cluster analysis · Seven-Factor Model

✉ Mete Sertkan
mete.sertkan@ec.tuwien.ac.at

Julia Neidhardt
julia.neidhardt@ec.tuwien.ac.at

Hannes Werthner
hannes.werthner@ec.tuwien.ac.at

[1] Research Division of E-Commerce, Institute of Information Systems Engineering, TU Wien, Favoritenstrasse 9-11/194, 1040 Vienna, Austria

1 Introduction

The relationship between Information and Communication Technologies (ICT) and tourism can be described as a symbiosis (Werthner and Klein 1999). Thus, the tourism landscape has been strongly affected and shaped by the rapid development of ICT during the last decades especially with the emergence of the World Wide Web (WWW). Nowadays, consumers have ubiquitous access to vast amounts of information and additionally, they are highly connected, allowing them to exchange experiences and more information among each other. However, increasing cognitive costs to process the amount and variety of information could lead to the problem of information overload. On the other side, the Web also allows a massive "informatization" of the whole tourism value chain, resulting in many novel value-generating strategies, to satisfy new consumer needs (Werthner and Ricci 2004).

According to a recent study (MediaCT 2014) people rely on online sources to get inspired where to go or how to travel. The study also shows that 65% of the leisure travellers start researching online before a travel decision. Particularly in this early phase of decision making a considerable amount of people has difficulties to explicitly express their preferences and needs (Zins 2007). Recommender Systems (RSs) aim to facilitate this decision-making. Ricci et al. (2015) define RSs as "software tools and techniques providing users with suggestions for items a user may wish to utilize" and thus addressing the problem of information overload. Particularly, profiling and personalization techniques might help in such cases, where preferences and needs are unknown or hard to express. Especially in tourism this is a big challenge, since tourism products are considered as very complex (i.e., they typically combine accommodation, transportation, activities, food, etc.), are mostly intangible and are highly associated with emotional experiences (Werthner and Ricci 2004). Considering all this, it is clear why sophisticated user models, which enhance understanding and processing of user preferences and needs, and tailored techniques, which reduce the cognitive load people are experiencing, have been and still are challenging issues of research in e-Tourism (Werthner et al. 2015).

Travel and destination decisions are usually not only based upon rational criteria but rather on implicitly given user preferences. It has been shown that a legitimate way to counteract this issue are personality based approaches, where preferences and personality are combined and used to build a comprehensive user model, which then can be exploited to recommend an item (Neidhardt and Werthner 2017). In the context of RSs personality can be seen as a user profile, which does not change with context (e.g., time, location, weather, etc.) and through different domains (e.g., movies, book, etc.) (Tkalcic and Chen 2015). A comprehensive and widely used personality model with respect to RSs is the Five-Factor Model, also known as the "Big Five" personality traits (Goldberg 1990).

Neidhardt et al. (2014, 2015) introduced a picture based approach to elicit the preferences of a user and a Seven-Factor Model to capture the respective user's profile within a travel recommender system. The Seven-Factor Model is the result of a factor analysis combining the "Big Five" personality traits (Goldberg 1990), which can be associated with the long-term behaviour of users, and 17 tourist roles proposed by Gibson and Yiannakis (2002), representing short-term preferences. These factors form the basis of a seven-dimensional vector space and are referring to travel behavioural patterns summarized as *Sun and Chill-Out*, *Knowledge and Travel*, *Independence and History*, *Culture and Indulgence*, *Social and Sport*, *Action and Fun*, and *Nature and Recreation*. RSs often tend to suffer from the so-called cold start problem (Burke 2007): RSs require historical user data or knowledge about a user's preferences and needs in order to propose appropriate items to that user. However, for a new user this information is typically missing, which is referred to as "cold start". In such cases preference elicitation can be accomplished explicitly, e.g., by asking the user a number of questions or implicitly, e.g., by observing his or her behaviour. In the picture based approach, the preferences of a user are determined by a simple picture-selection process. Thus, the user just has to select three to seven pictures out of a pre-defined set rather than tediously answering a questionnaire. The resulting user's profile comprises a score for each of the factors and thus can be seen as a point in the seven-dimensional vector space. In order to provide recommendations to a user, those items have to be determined that are closest to him or to her. Therefore, also the items have to be mapped onto the vector space, i.e., they have to obtain a score for each of the Seven-Factors. In order to build up a reasonable recommendation base more than 10,000 point of interests (POIs) such as activities, events, restaurants, sights etc. were initially mapped manually by experts. Obviously, this approach does not scale. To counter this limitation, Glatzer et al. (2018) introduced a text-mining-based method, where hotels are allocated to the Seven-Factors.

Following this line of research, the presented work aims to introduce an automated way of determining the Seven-Factor representation of tourism products. However, in contrast to Neidhardt et al. (2014, 2015) and Glatzer et al. (2018), where POIs and hotels are considered as tourism products, the focus will be on tourism destinations. Furthermore, unlike Glatzer et al. (2018), where hotels are classified to one or more factors of the Seven-Factor Model, this work aims to determine a score for each factor of the Seven-Factor Model. Similarities among tourism destinations will be analysed in order to identify latent conceptually meaningful groups, which will deliver better insights into the (dis)similarities among destinations and which, moreover, can be addressed directly by a RSs. Furthermore, the relationships between the Seven-Factors and attributes of destinations will be examined in order to map the destinations onto the Seven-Factors. This work aims to extend the analysis presented in Sertkan et al. (2018). In contrast to the previous work, a more sophisticated pre-processing of the provided data set is conducted and the data set is described and explored in more detail. In addition, the cluster analysis is extended to a considerably higher number of destinations. Furthermore, the multiple linear regression model is challenged by two non-linear and conceptually different regression methods. Finally, the outcomes are evaluated and discussed in greater depth.

The rest of the paper is organized as follows. In Sect. 2 the state of the art is presented, focusing on tourist roles, the Seven-Factor Model, and tourism recommender systems. In Sect. 3, the data set in use is described and explored. In Sect. 4 a cluster analysis aiming to identify meaningful groups is conducted and presented. In Sect. 5 regression analyses based on expert mappings are conducted, compared and evaluated. In Sect. 6 the outcomes of the conducted analyses are jointly discussed, conclusions are drawn and an outline for future work is given.

2 State of the art

This chapter provides a brief overview of related work and the state of the art in the field. First, RSs in the tourism domain are examined and finally the Seven-Factor Model is introduced and discussed in more detail.

2.1 Recommender systems in the tourism domain

Travel and tourism have always been major application domains for Web-related services (Werthner and Klein 1999). As the amount of information on the Web started to rise, the call for techniques to cope with information overload began to grow. One answer to that are RSs. From the supplier's perspective, RSs are aiming to bring the right products to the right customers, in order to increase customer experience, satisfaction, trust, and in turn profit. On the other hand, from a consumer's point of view, RSs are aiming to provide suggestions in order to support and simplify various decision-making processes. In case of tourism these decisions might be: where to go? How to travel? Where to stay? What to do? and much more. Whereas, an item, i.e., suggestion, can be a destination, a hotel, an activity, a flight etc.

During the last decades, many recommendation techniques evolved including content-based, collaborative-filtering, knowledge-based, demographic, community-based, and hybrid. Most of the listed techniques rely on user rating behaviour and were proposed for products such as movies, music, or books. However, since traveling is costly and time consuming, there are typically less rating data in the tourism domain, which leads to less accurate personalization techniques, compared to other products (Neidhardt et al. 2015). Another challenge for RSs in the tourism domain is that tourism products are complex (e.g., a bundle of accommodation, transportation, activities etc.), intangible and highly associated with emotional experiences (Werthner and Klein 1999). In order to bundle and recommend the right tourism product RSs are relying on content and knowledge (Neidhardt et al. 2015). Also, Burke and Ramezani (2011) argue that most appropriate recommendation techniques in the matter of tourism are either knowledge based and/or content-based.

This work aims to find an automated way of determining the Seven-Factor representation of tourism destinations to enable a matchmaking with user profiles (i.e., Seven-Factor representations of users). The picture based approach to RSs (Neidhardt et al. 2014, 2015) uses a gamified and user centric way to elicit the Seven-Factors of a user. Preferences and needs of a user are determined via a simple picture selection process. Users are addressed on an emotional, implicit level and do not have to state their preferences explicitly. This simple method, which can be considered as content- and knowledge-based approach, counteracts peoples difficulties in explicitly expressing their preferences and needs and helps to overcome the so-called cold-start problem.

According to Garcia et al. (2011) tourism RSs can be distinguished into two types: one focusing on destination selection the other on activities that can be performed at a certain destination. The presented work can be considered as part of the first group since it considers tourism destinations as recommendation items. In contrast to Neidhardt et al. (2014, 2015), where the focus lies on Point of Interests (POIs), e.g., activities, events, restaurants, sights. Much research has already been conducted targeting destination recommender systems (Fesenmaier et al. 2006; Borràs et al. 2014), but they are mainly focusing on distinct regions or POIs in a destination. There are few, moreover, that focus on personality traits and motifs of a user, e.g., Braunhofer et al. (2014).

2.2 Seven-factor model

In the last decades, much research has been conducted in order to identify and categorize tourist roles, describing the relation between a person's travel behaviour and his or her preferences, interest, and needs. In this context, Gibson and Yiannakis (2002) introduced a well-established classification framework, distinguishing 17 different tourist roles to capture short-term preferences of tourists, i.e., preferences, which might change depending on the context (e.g., seasonality like summer or winter, special occasions, single or group, etc.). Gretzel et al. (2006) demonstrated that tourist roles can be used in order to recommend touristic activities and, in turn, destinations. It has also been shown that tourist roles can be related to personality traits. Delić et al. (2016) provide significant evidence that there are relations between the well-established "Big-Five" personality traits (Goldberg 1990) and the 17 tourist roles (Gibson and Yiannakis 2002). Matthews et al. (2003) argue that "large-scale reviews and large single studies offer overwhelming evidence for the stability of personality traits over many years". Thus, they can facilitate the prediction of the long-term behaviour of a person (Woszczynski et al. 2002).

Both the "Big Five" personality traits and the 17 tourist roles are well-established frameworks and have been subject and bases to many empirical and behavioural studies. Thus, there are existing standardized methods to assess and measure both of them. Neidhardt et al. (2014, 2015) conducted an online and offline survey, with 30 questions addressing the tourist roles and 20 questions addressing the personality traits. About 1000 participants completed the questionnaires. Upon the collected data, they conducted a factor analysis in order to reduce the 22 dimensions (the "Big Five" personality traits plus 17 tourist roles) and summarize them in fewer dimensions. The factor analysis resulted in seven independent factors, which are able to capture different travel behavioural patterns. They are summarized in the following:

Sun and Chill-Out a neurotic sun lover, who likes warm weather and sun bathing and does not like cold, rainy or crowded places;

Knowledge and Travel an open minded, educational and well-organized mass tourist, who likes travelling in groups and gaining knowledge, rather than being lazy;

Independence and History an independent mass tourist, who is searching for the meaning of life, is interested in history and tradition, and likes to travel independently, rather than organized tours and travels;

Culture and Indulgence an extroverted, culture and history loving high-class tourist, who is also a connoisseur of good food and wine;

Social and Sports an open minded sportive traveller, who loves to socialize with locals and does not like areas of intense tourism;

Action and Fun a jet setting thrill seeker, who loves action, party, and exclusiveness and avoids quiet and peaceful places;

Nature and Recreation a nature and silence lover, who wants to escape from everyday life and avoids crowded places and large cities.

These factors are easier to process cognitively as well as computationally compared to the original 22 dimensions. However, dimensionality reduction not only decreased computational and cognitive cost, but also lead to a better understanding and more insights. Neidhardt and Werthner (2017) showed that based on different demographic characteristics different user groups can be well distinguished within this model.

3 The data

Beirman (2003) refer to a tourism destination as “a country, state, region, city or town which is marketed or markets itself as a place for tourists to visit”. In this work destinations are defined in a similar way, except that the range is wider, i.e., from a hamlet with a population smaller than 100 to a metropolis with a population larger than one million. The data is provided as a SQL-dump by a German e-Tourism company and consists of more than 30,000 destinations all around the world.

Figure 1 shows the structure of the tables in the SQL-dump and the relations among them. Destinations are described through 22 geographical attributes and 27 motivational ratings.

Motivational ratings lie in the interval [0,1] and describe the degree of suitability for a particular motif. The following 27 motifs are listed: *nightlife, wellness, shopping, nature and landscape, image and flair, culture, sightseeing, entertainment, mobility, price level, accommodations, gastronomy, beach and swimming, golf, scuba diving, kite and windsurfing, hiking, cycling, horseback riding, winter sports, sports, family, peacefulness, surfing, sailing, gays, mountain biking*. The motivational ratings are determined by the e-Tourism company by considering factors such as infrastructure, climate, user opinions, number of services, image, and marketing.

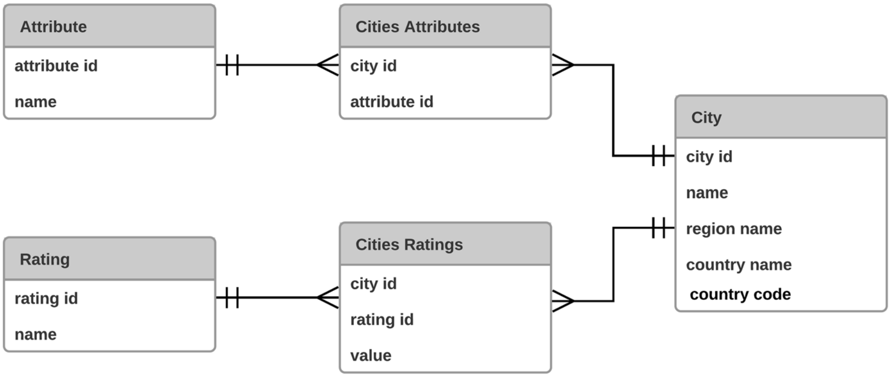

Fig. 1 Entity-relationship model of the provided destinations SQL-dump

However, not all details are disclosed and thus it is not known how exactly the scores are determined.

Geographical attributes are given in binary format and describe the presence or absence of a particular geographical attribute. The following 22 attributes are listed: *sea, mountain, lake, island, sandy beach, metropolis, forest, river, desert, old town, pebble beach, sand and pebble beach, hill, swamp, volcano, fjord, flat decaying sandy beach, beach promenade, wine-growing, heath, health resort, winter sports resort*.

All possible attributes and ratings are persisted in the tables *Attribute* and *Rating*. As previously mentioned, there are over 30K tourism destinations. They are persisted in the table *City*. This table contains an identifier for each destination and textual descriptions to capture destination name, region name, country name, and country code in ISO 3166-1 alpha-2 format, for example AT for Austria. In the table *Cities Attributes* tuples of geographical attributes and tourism destinations are recorded, e.g. (Vienna, *old town*). Similarly, the table *Cities Ratings* persists the motivational ratings of a tourism destination with corresponding (rating) value, e.g. (Vienna, *culture*, 0.99). A major drawback of such a structure is that a tourism destination does not necessarily have an entry for each rating or attribute. Thus, in many cases it is not clear if a destination does in fact not have such attribute or rating, or the data is missing. This ambiguity leads to many "missing values" and in turn to a sparse data set.

In Sertkan et al. (2018) missing values were treated naively by replacing them by zero. In contrast, the presented work uses a more sophisticated strategy to treat missing values. First, destination attributes, which are denoted by the data provider as experimental, are deleted. Afterwards, destination attributes with similar meaning are combined, e.g. *sandy beach*, *flat decaying sandy beach*, *pebble beach*, *sand and pebble beach* are combined to the attribute *beach*. Then, destinations with many missing values are discarded. Since there are seven destination features (i.e., *price level*, *gastronomy*, *sports*, *accommodations*, *shopping*, *nightlife*, and *entertainment*) which are mostly non-missing (i.e., in about 90% of the cases), only destinations with minimum ten non-missing features are kept in the data set in order to have at least three more aspects of a tourism destination. This leads to a more concise data set with 16,950 destinations and 38 attributes (i.e., 26 motivational ratings and 12 geographical attributes). Finally, a model-based procedure is defined (i.e., naive imputation for geographical attributes and SOFT-IMPUTE (Mazumder et al. 2010) for motivational ratings) in order to intelligently replace the remaining missing values (i.e., 62% of the overall cells).

Almost all countries are represented in the database, but the majority (65%) of destinations are located in the USA, Germany, France, Italy, Spain, Great Britain, Austria, Greece, Switzerland, and Sweden. This can also be observed in Fig. 2, where the distribution of tourism destinations over countries is presented as a heat map.

Summary statistics of the distributions of the different motivational ratings for the tourism destinations are listed in Table 1. The motivational ratings *nature and landscape, peacefulness, hiking, cycling,* and *mountain biking* have similar and high average values of 0.61–0.71. Those ratings are not only similar, but they can also be considered as nature and recreation related. On the other end, some specific water sports related ratings (i.e., *sailing, diving, kite- and windsurfing,* and *surfing*) have low average values (0.21–0.29).

In Table 2 the relative frequencies of the different geographical attributes are listed. It can be observed that 21–25% of destinations are either on an island and/or at the seaside and/or near a beach. However, it is noteworthy that in the data set 43% of destinations do not have any geographical attribute listed at all.

In order to analyse similarities among all destination attributes (i.e., motivational ratings and geographical attributes) a correlation matrix comprising all pairwise Pearson correlation coefficients is calculated. To get a better understanding and overview, the correlation matrix is visualized as a clustered heat map (see Fig. 3).

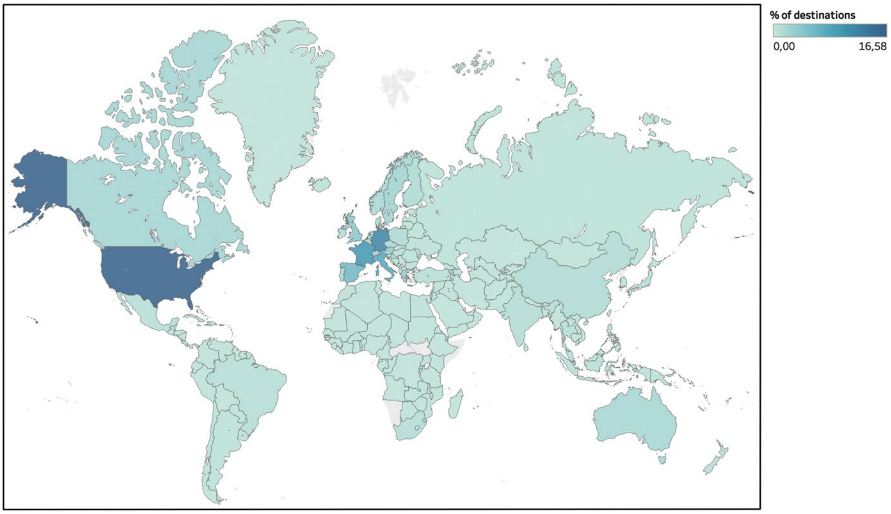

Fig. 2 Distribution of tourism destinations over countries

Table 1 Summary statistics of the motivational ratings

	Mean	Std	Min	Median	Max
Nightlife	0.51	0.15	0.05	0.50	0.99
Wellness	0.37	0.16	0.05	0.32	1.00
Shopping	0.52	0.14	0.02	0.51	1.00
Nature_landscape	0.71	0.16	0.10	0.75	1.00
Image_flair	0.64	0.14	0.09	0.62	1.00
Culture	0.51	0.15	0.03	0.46	1.00
Sightseeing	0.40	0.17	0.05	0.35	1.00
Entertainment	0.31	0.21	0.01	0.26	0.98
Mobility	0.45	0.15	0.13	0.41	1.00
Pricelevel	0.57	0.15	0.10	0.56	1.00
Accommodations	0.41	0.19	0.05	0.38	0.97
Gastronomy	0.40	0.20	0.05	0.36	0.98
Beach_swimming	0.45	0.21	0.01	0.36	0.98
Golf	0.39	0.13	0.01	0.36	1.00
Diving	0.28	0.16	0.01	0.22	1.00
Kite_windsurfing	0.26	0.15	0.01	0.19	1.00
Hiking	0.68	0.17	0.12	0.69	0.98
Cycling	0.63	0.15	0.09	0.66	0.96
Horsebackriding	0.31	0.11	0.01	0.28	1.00
Wintersports	0.27	0.12	0.03	0.26	0.93
Sports	0.39	0.19	0.05	0.36	0.98
Family	0.44	0.14	0.04	0.41	1.00
Peacefulness	0.70	0.18	0.06	0.74	1.03
Surfing	0.21	0.14	0.01	0.15	1.00
Sailing	0.29	0.15	0.01	0.23	1.00
Mountainbiking	0.62	0.18	0.09	0.59	0.99

Table 2 Frequencies of geographical attributes

	Relative frequency
Island	0.25
Sea	0.22
Beach	0.21
Mountains	0.09
Forest	0.07
Old_town	0.07
Hill	0.07
Volcano	0.06
Lake	0.04
Health_resort	0.04
Metropolis	0.03
River	0.03

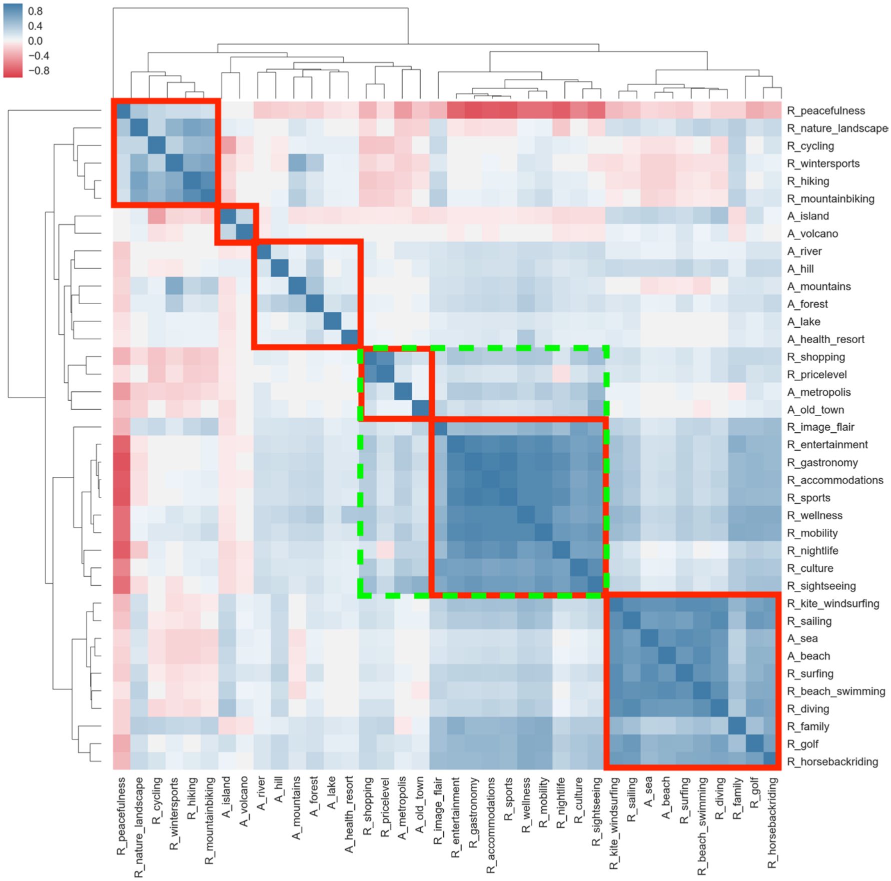

Fig. 3 Clustered correlation heat map of all destination attributes

Following meaningful groups are identified (marked by red rectangular):

- Group one comprises attributes with a relationship to recreational travelling. *Peacefulness*, *nature & landscape, mountain biking, hiking, cycling* and *winter sports* are forming this group.
- Group two consist of *island and volcano*, which might be a sign for nature and adventure.
- In group three there are typical attributes of destinations at the countryside, namely *river, hill, mountains, forest, lake* and *health resort*.
- Group four comprises attributes related to city trips and metropolitan areas. Those attributes are *shopping, price level, metropolis, and old town*.
- Group five contains attributes related to mass tourism, such as *image and flair, entertainment, gastronomy, accommodations, sports, wellness, mobility, nightlife, culture,* and *sightseeing*. The attributes within this group are strongly positively correlated.

- Group six mainly comprises attributes related to water sports and beach vacations. Attributes within this group are *kite and windsurfing, sailing, sea, beach, surfing, beach and swimming, diving, family, golf,* and *horseback riding*. The last three attributes are a bit detached from the other attributes within this group, which is also reflected by the low correlation coefficients.

Overall, one can observe a contrast between attributes related to mass tourism (green rectangular) and attributes related to recreational destinations (first group), especially in the case of motivational rating *peacefulness*. However, the correlation analysis delivers first insights of a latent structure in the given data set and furthermore the results can be used in order to substitute or merge highly correlated attributes.

Out of all destinations, 561 destinations were chosen randomly and mapped manually to the Seven-Factors by experts. These experts were members of an Austrian e-Tourism company using an implementation of the picture based approach. Thus, they were familiar with both characteristics of tourism destinations and the Seven-Factor Model. For the 561 destinations, three experts assigned first individually a score for each factor using the scale 0–0.25–0.50–0.75–1. The higher the score the more suitable, in the expert's opinion, the destination for that specific factor. After the individual mappings, a final mapping was determined in a joint discussion. The majority of destinations in the sample are located in Germany, USA, France, Greece, Great Britain, Italy, Denmark, Spain, Austria, and Netherlands (62%), which is similar to the distribution in the whole data set. However, after the essential missing value treatment, also the expert sample got smaller in size, namely from N = 561 to N = 350 (62%).

In Fig. 4 the rating behaviour of the experts can be examined, i.e., the score distribution for each of the Seven-Factors are shown. For example, in case of the factor *Sun and Chill-Out* 30% of the destinations scored with 0, 17.1% with 0.25, 15.4% with 0.5, 10.6% with 0.75, and 26.9% with 1. Thus, the majority of destinations (56.9%) scored with 0 or 1 in the factor *Sun and Chill-Out*. Whereas, the majority of destinations (55.7%) in case of *Knowledge and Travel* scored with either 0 or 0.25 and a few with 1 (10.6%), similar to the distribution in the factor *Action and Fun*. Almost half of the destinations have a score in the "lower middles," i.e., 0.5 (28.3%) and 0.25 (21.4%), in the factor *Culture and Indulgence*. This is similar to the distribution in the factor *Nature and Recreation*, where the majority of destinations have scores in the "upper middles," i.e., 0.5 (24%) and 0.75 (27.4%). An extreme case of this "upper middles" can be seen in factor *Social and Sports*, where almost all destinations (87.1%) scored with either 0.5 (53.4%) or with 0.75 (33.7%). The only factor where an approximately normal distribution (bell shape) of scores can be observed is *Independence and History*.

4 Cluster analysis

Identifying conceptually meaningful groups of destinations with shared common characteristics will help to further understand the data and its structure, which may contribute to a more generalizable solution. Furthermore, clustering is considered as

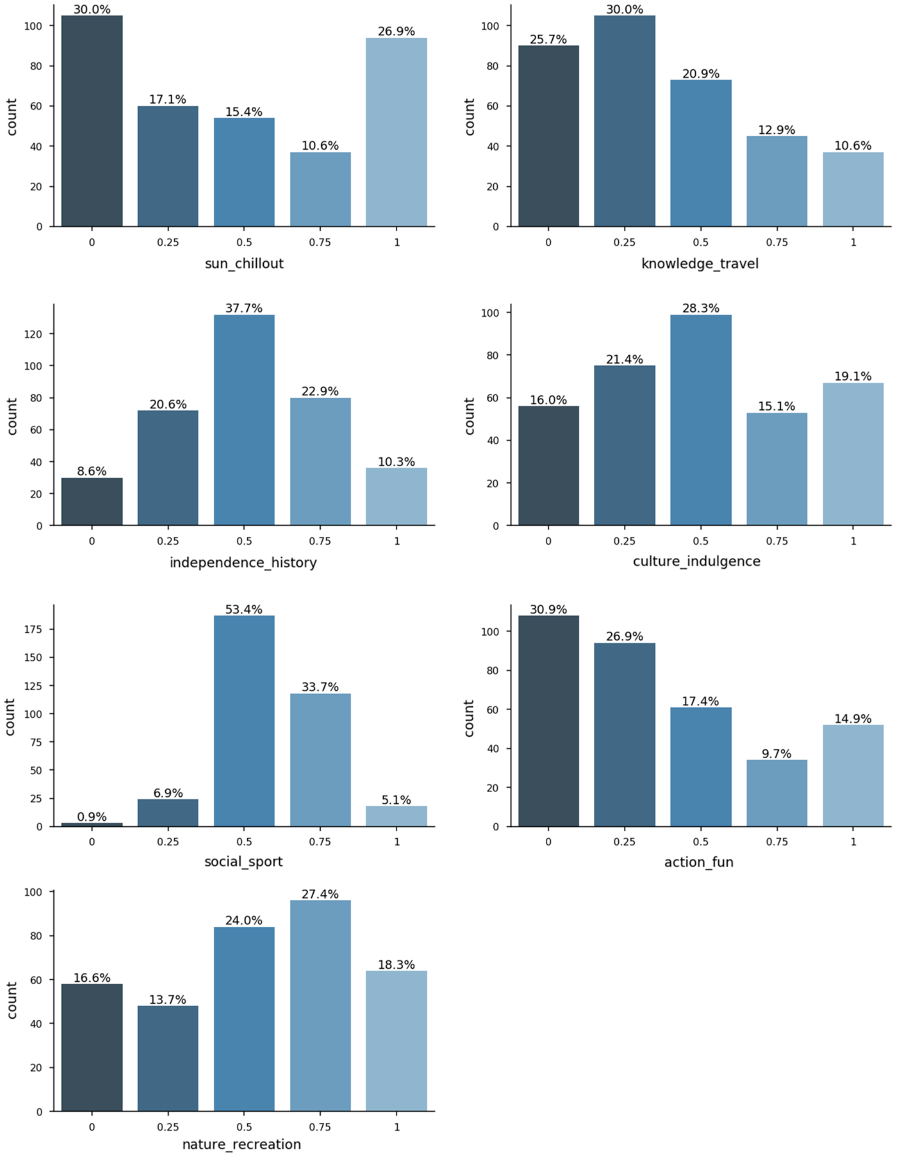

Fig. 4 Distribution of Seven-Factor scores in the expert sample

an unsupervised learning method. Hence, it does not rely on a priori labelled data sets and thus on resource intensive (e.g., time, costs, etc.) expert knowledge. Therefore, it can be used as an initial approach for recommending destinations if no further knowledge about the destinations is available. The cluster analysis comprises 16,950 destinations (i.e., the data set after pre-processing). Partitional clustering techniques are considered, where most prominent ones are K-means and K-medoids.

Since the data comprises binary attributes, using the Euclidean distance and thus centroids (both are essentials of the K-means algorithm) are not meaningful. Therefore, K-medoids is applied. A medoid corresponds per definition to an actual data point, which is considered as the most representative point for the cluster (Tan et al. 2006). Specifically, Partitioning Around Medoids (PAM) (Kaufman and Rousseeuw 1990), the most common K-medoids algorithm, is used. Since the data consists of two different data types, i.e., binary (geographical attributes) and continuous (motivational ratings), the Gower distance (appropriate for mixed datatypes) (Gower 1971) is used as distance metric.

In order to find an appropriate number of clusters, the internal evaluation metric silhouette width (i.e., silhouette coefficient) (Rousseeuw 1987) is used for assessment. Figure 5 shows the average silhouette width within different cluster sizes. Based on the average silhouette width two, three, four and six cluster solutions are considered, but for the sake of interpretability and diversity a six-cluster solution is chosen. For example, the four-cluster solution distinguishes between beach resorts and non-beach-resorts. On the other hand, the six-cluster solution enables to distinguish between mass touristic beach resorts, recreational beach resorts, and non-beach-resorts. This adds diversity to the cluster solution and facilitates finding associations with the Seven-Factors. Next, the resulting clusters are examined in detail. The number of destinations in each cluster is provided at the beginning of each paragraph.

C1 (N = 1940) The medoid of cluster C1 is Paralia, a small city in Greece. Paralia means in Greek beach and as the name already suggests, the city is located directly at the beach. It is a popular and vibrant seaside resort with many nightlife and shopping opportunities. Interestingly, 93% of the destinations in C1 are located at the sea and 92% directly at the beach, whereas globally only about 20% of destinations are

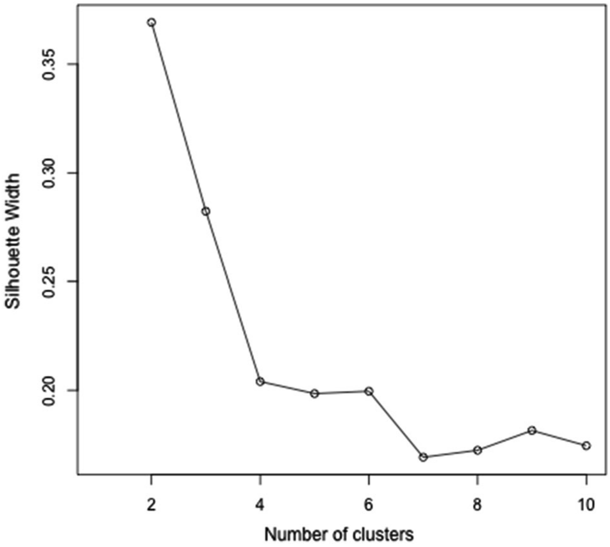

Fig. 5 Average silhouette width in different k numbers of clusters

located at the sea or beach. Also, the rating *beach and swimming* has a high mean value of 0.81. Additionally, ratings *gastronomy, nightlife, sports, accommodations*, and *culture* are showing an increased average value (0.62–0.66). To conclude, destinations in C1 are mainly located on the beach, vibrant and lively, and also attractive for various sports.

C2 (N = 2177) The medoid of cluster C2 is Gubbio, a city located on the lowest slope of Mt. Ingino in Italy. Its origins are ancient and reach back to the Bronze Age. Thus, many cultural and sightseeing activities are provided. Features *image and flair, hiking, culture, gastronomy, nightlife, mobility, accommodations, sports, sightseeing*, and *entertainment* are showing increased mean values in C2 (0.61–0.77). Interestingly, 17% of the destinations in C2 are metropoles, which is about six times more considering the whole data set. Plus, only 1% of the destinations in C2 are located at the sea or beach. Hence, destinations in C2 can be considered as mainly vibrant cities or metropoles not located at the beach, offering many nightlife, cultural, sightseeing, gastronomy, and entertainment opportunities.

C3 (N = 1774) The medoid of cluster C3 is Aghios Markos, a small, peaceful village in the nature on the island of Corfu, Greece. In C3, 90% of the destination are located at the sea, 88% at the beach, and 70% on an island. Whereas, in the whole data set only 20% of destinations are located at the sea or beach and 25% on an island. Ratings *beach and swimming, nature and landscape,* and *peacefulness* have an increased mean value of 0.77–0.79. Furthermore, there is only one metropolis in C3. Therefore, destinations in C3 can be seen as small and peaceful towns at the seaside, probably on an island, with a few sports opportunities and not many tourists.

C4 (N = 5576) The medoid of cluster C4 is Montbriò del Camp, a small, peaceful village in Catalonia, Spain. The average value of motivational rating *peacefulness* in C4 is 0.81. Also, ratings *nature and landscape, hiking, cycling* and *mountain biking* have an increased mean value of 0.62-0.67. Interestingly, none of the 5576 destinations is located on an island or is a metropolis. Furthermore, the mean values of all other attributes are relatively low for destinations in C4. Hence, destinations of C4 can be considered as small and peaceful villages, probably in the nature, and more or less good for hiking, cycling, and mountain biking.

C5 (N = 1877) The medoid of cluster C5 is Reynoldston, a small, peaceful village in Wales, Great Britain. Interestingly, all destinations within this cluster are located on an island, only 2% are at the beach, and there is only one metropolis. Further, only ratings *peacefulness* (0.76), *nature and landscape* (0.71), and *hiking* (0.64) are showing an increased mean, all other destination attributes have a relatively low average value. C5 is quite similar to C4, except that destinations of C5 are only located on islands, where destinations of C4 are not. Thus, destinations of C5 can be considered as mainly small, peaceful villages, located on an island and in the nature, with some recreational sports offers.

C6 (N = 3606) The medoid of cluster C6 is Irun, a city in Spain at the border to France and on the Atlantic coast. It offers some cultural and sightseeing activities, but also some sports and recreational activities in the nature. Following ratings have an increased average values (0.63–0.71) within this cluster: *nature and landscape, peacefulness, image and flair, mountain biking, cycling, nightlife,* and *culture*. In C6, 12% of the destinations are located near a mountain, which is about three times more compared to the whole data set. Only 1% of the destinations are considered as metropoles and only 1% are located at the beach or sea. Thus, destinations within C6 can be considered as small cities, probably in the nature, with recreational, cultural, and entertainment offers, but none of them dominating.

In summary, it can be said that there is an underlying structure of the data leading to six conceptually meaningful groups of destinations. For a better understanding, these groups or clusters can be simplified and summarized as follows: C1—*vibrant beach resorts*, C2—*energetic cities*, C3—*tranquil seaside resorts*, C4—*peaceful towns*, C5—*idyllic island villages*, C6—*ordinary towns*. Furthermore, the identified clustering is showing a clear contrast between tranquil and vibrant, land and island, and seaside and inland.

The silhouette plot in Fig. 6 displays the silhouette coefficients of each destination in a cluster in an ordered way. The red dashed line shows the average silhouette width of 0.2, which assisted to find the right cluster size. The silhouette plot enables a visual assessment of the relative quality of the developed clustering. A negative silhouette coefficient indicates an incorrect assignment of a destination to a cluster and a very low silhouette coefficient points out that a destination is located in-between two clusters. Hence, almost none of the destinations in C4 and C5 are incorrectly assigned, but some might be located between two clusters. Whereas, in all other clusters there are falsely assigned destinations, especially in C2 and C6.

Furthermore, the proposed clustering can be analysed in the light of the provided expert mapping (N = 350). Table 3 lists the mean Seven-Factor scores and corresponding standard deviations (SD) of the expert sample as a whole and in different clusters. Note, cluster means, which are significantly different to the average factor scores of the rest of the sample, are in bold. Significance is tested

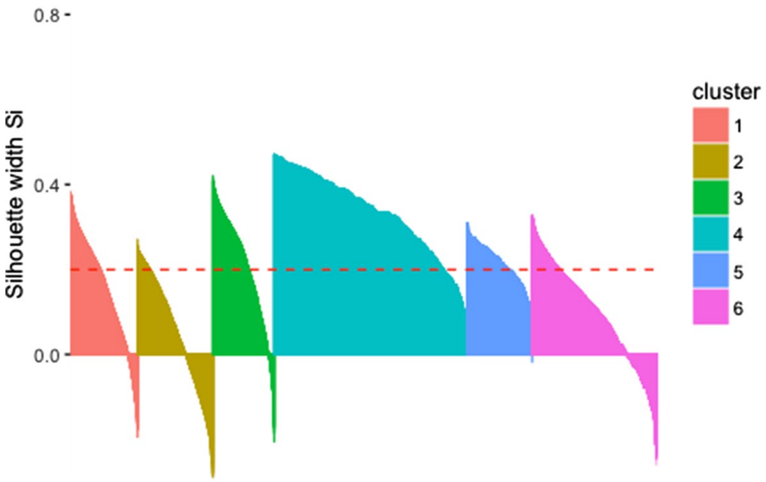

Fig. 6 Silhouette plot of the six-cluster solution

Table 3 Average factor scores and standard deviations (SD) of the expert sample (N = 350) as a whole and in different clusters

	Sample	C1	C2	C3	C4	C5	C6
Sun and Chill-Out							
Mean	0.37	**0.71**	**0.18**	**0.96**	**0.22**	0.39	**0.30**
SD	0.38	0.31	0.26	0.11	0.25	0.43	0.33
Knowledge and Travel							
Mean	0.30	**0.46**	**0.65**	**0.24**	**0.14**	**0.18**	0.30
SD	0.30	0.31	0.30	0.21	0.21	0.23	0.24
Independence and History							
Mean	0.45	**0.58**	**0.73**	**0.42**	**0.31**	**0.32**	0.45
SD	0.28	0.23	0.22	0.21	0.24	0.28	0.24
Culture and Indulgence							
Mean	0.40	**0.58**	**0.76**	0.41	**0.23**	**0.24**	0.44
SD	0.34	0.30	0.26	0.22	0.26	0.30	0.30
Social and Sports							
Mean	0.57	**0.64**	0.57	**0.50**	0.59	0.58	0.58
SD	0.19	0.16	0.19	0.16	0.18	0.14	0.20
Action and Fun							
Mean	0.26	**0.59**	**0.56**	0.30	**0.05**	**0.12**	**0.19**
SD	0.33	0.34	0.34	0.19	0.15	0.17	0.20
Nature and Recreation							
Mean	0.54	**0.35**	**0.31**	**0.76**	**0.82**	**0.79**	**0.72**
SD	0.34	0.28	0.30	0.18	0.17	0.22	0.25

Cluster means, which are significantly different to the average factor scores of the rest of the sample, are in bold

with the Mann-Whitney-U-test (Hollander et al. 2013) at a significance level of 0.05.

The factor *Sun and Chill-Out* shows, as expected, an increased mean value in cluster C1—*vibrant beach resorts* and C3—*tranquil seaside resorts*. The factors *Knowledge and Travel*, *Independence and History*, and *Culture and Indulgence* have an increased average score in clusters C1—*vibrant beach resorts* and C2—*energetic cities*, where in contrast to the other clusters more sightseeing, cultural, entertainment, and nightlife activities are offered. Furthermore, the mentioned factors have a decreased mean value (in comparison to the sample mean) in the clusters with more tranquil and peaceful destinations (C3, C4, C5), where the probability of activities fitting the characteristic aspects of the three mentioned factors is low. In comparison to the sample mean, the factor *Social and Sports* shows an increased score in cluster C1—*vibrant beach resorts* and a decreased score in C3—*tranquil seaside resorts*, but overall it does not show much diversity. In other words, the cluster means are similar (i.e, 0.50–0.64) throughout the clusters and in comparison to the sample mean. The factor *Action and Fun* has an increased mean value in destinations, which are considered as vibrant and energetic (i.e., destinations in clusters C1 and C2), and a decreased mean value in destinations, which are considered as peaceful, idyllic, or ordinary (i.e., destinations in clusters C4, C5, and C6). Such average Seven-Factor score distribution is reasonable and in line with the characteristics of the factor. Finally and obviously, the factor *Nature and Recreation* has an increased score in tranquil, peaceful, ordinary, and idyllic places (i.e., destinations in C3, C4, C5, and C6), rather than in destinations with mass tourism characteristics (i.e., destinations in C1 and C2).

5 Regression analysis

The objective of the work is to automatically map destinations onto the seven-dimensional vector space of travel behavioural patterns based on attributes of the destinations. However, additionally the relationships between the Seven-Factors and these attributes should be understood. James et al. (2013c) suggest to choose linear models over more complex ones if inference and interpretability is the goal. Taking this into account, a multiple linear regression model (James et al. 2013b) with stepwise variable selection (James et al. 2013a) is applied. All Seven-Factors are considered as independent from each other, since they are obtained from factor analysis. Therefore, they can be treated separately by fitting a model for each travel behavioural pattern, which takes the attributes of a destination as input and returns the factor score in the interval [0, 1] as output. The expert sample is split into a training and test set in a ratio of 80/20. Model performance is assessed by R^2, the proportion of variance explained, and root mean square error (RMSE), the standard deviation of the residuals / prediction errors (Gunawardana and Shani 2009; James et al. 2013c). Furthermore, a performance evaluation is conducted, in order to compare the performance of the linear model against following two more complex models: K-Nearest-Neighbour regression (KNN) (James et al. 2013b) and Random Forest regression (RF) (James et al. 2013d). Finally, the outcomes are evaluated by assessing the performance against a baseline and by examining the distributions of the predicted factors.

The resulting multiple linear regression models comprise both motivational ratings and geographical attributes. After the variable selection 15 out of 26 motivational ratings and seven out of twelve geographical attributes in total are used. Table 4 summarizes the outcomes of the regression analysis. Motivational ratings *sightseeing, peacefulness, nightlife, culture, nature and landscape,* and *shopping* appear in more than one model. Also, geographical attributes *health resort* and *sea* are used in several models. In the following, the models are discussed in more detail.

F1—Sun and Chill-Out The geographical attributes *sea*, *health resort*, and especially the motivational rating *beach and swim* have a significant positive impact on the factor *Sun and Chill-Out*. Those attributes can be interpreted as indicators for sun and relaxation. On the other side, the motivational rating *nightlife* has a significantly strong, negative impact, which can be associated with crowded places and mass tourism.

Table 4 Results of the regression analysis conducted on the expert sample (N = 350)

	F1	F2	F3	F4	F5	F6	F7
(Intercept)	0.41***	– 0.18***	0.08	– 0.09	0.28**	0.02	0.61***
Sightseeing	–	1.02***	0.39***	0.30*	– 0.29***	–	– 0.27***
Nightlife	– 0.76***	–	–	–	–	0.41***	– 0.57***
Peacefulness	–	–	–	–	0.27**	– 0.53***	0.51***
Health resort	0.27***	–	–	–	–	– 0.11***	0.09**
Sea	0.23***	– 0.12***	–	–	–	0.17***	–
Culture	–	–	0.61***	0.47***	–	–	–
Nature and landscape	–	–	– 0.07*	– 0.27**	–	–	–
Shopping	–	–	–	–	–	0.42***	– 0.22**
Winter sports	–	– 0.24*	–	–	–	0.54***	–
Beach and swimming	0.73***	–	–	–	–	–	–
Family	–	–	–	–	–	– 0.33***	–
Hiking	–	–	–	–	–	–	0.35***
Image and flair	–	–	–	0.48***	–	–	–
Kite and windsurfing	–	–	–	–	-	0.10***	–
Mobility	–	0.26**	–	–	–	–	–
Sports	–	–	–	–	0.85***	–	–
Wellness	–	–	–	–	– 0.31**	–	–
Metropolis	–	–	–	–	–	0.19***	–
Mountains	–	–	–	–	0.09***	–	–
Old town	–	–	–	0.17***	–	–	–
Beach	–	–	–	–	–	–	– 0.05*

Significance level is coded as follows: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

F2—Knowledge and Travel The motivational rating *sightseeing* has a significant, strongly positive relation with the factor *Knowledge and Travel*. Hence, it is capturing the knowledge part of the factor. Whereas, the motivational rating *mobility* is also positively related to the factor and one can say it captures the travel part. On the other hand, the geographical attributes *sea* and *winter sports resort* are significantly negatively related with the factor. This is reasonable, since in such areas usually the tourism focus does not lie on gaining knowledge.

F3—Independence and History The motivational ratings *culture* and *sightseeing* are significantly, positively related to the factor *Independence and History*. Those attributes can be seen as the main motivation of travellers with interests in history and tradition. Whereas, the motivational rating *nature and landscape* has a significant negative impact on the factor. Since cultural and historical interests are short coming in nature and recreation related destinations, such negative association is reasonable.

F4—Culture and Indulgence The motivational ratings *sightseeing*, *culture*, *image and flair*, and geographical attribute *old town* are significantly, positively related to the factor *Culture and Indulgence*. Those ratings can be interpreted as the main motivation of a culture and history interested high class tourist. On the other side, the motivational rating *nature and landscape* has a significant, negative impact on the factor. Again, this might show that destinations branded with a nature and landscape motif have shortcomings in cultural tourism.

F5—Social and Sports The motivational rating *sports* has a strong, significant, positive impact on the factor *Social and Sports*, which is obvious. Also, the motivational rating *peacefulness* and the geographical attribute *mountains* have a significant positive relation to the factor. Since the factor *Social and Sports* factor avoids crowded areas and locations of mass tourism, and prefers more tranquil places, positive associations of both attributes with the factor are reasonable. On the other hand, the motivational rating *sightseeing* has a significant, negative impact on the factor. It can be seen as an indicator of crowded areas and mass tourism. Surprisingly, the motivational rating *wellness* is significantly, negatively associated with the factor *Social and Sports*. This is caused by an unsound sample, as 55% of the destinations in the expert sample have a larger wellness rating (> 0.5) and are located at the beach and 25% are metropoles, which is far less in the whole data set.

F6—Action and Fun The motivational ratings *peacefulness*, *family*, and the geographical attribute *health resort* have a significant, negative impact on the factor. This fits perfectly to the character traits of the factor *Action and Fun*. Whereas, the motivational ratings *nightlife*, *winter sports*, *shopping*, and the geographical attributes *sea* and *metropolis* are significantly positively related to the factor. Those can be interpreted as attributes of energetic, vibrant and action loaded places, which are main aspects of destinations for thrill seeking and action loving travellers.

F7—Nature and Recreation The motivational rating *peacefulness*, *hiking*, and the geographical attribute *health resort* are significantly positively related to the factor *Nature and Recreation*, which is obvious and does not need further explanation. On the other side, the motivational ratings *nightlife*, *sightseeing*, *shopping*, and the geographical attribute *beach* have a significant, negative impact on the factor. Those attributes can be interpreted as signs of mass tourism and crowded areas. Hence, a negative association on a recreational and escapist traveller is reasonable.

The resulting models are evaluated by assessing both in-sample and out-of-sample performance. In other words, the performance measures are determined using both training set (in-sample) and test set (out-of-sample). Obviously, out of sample performance plays a bigger role, because it delivers an approximation to the question, how the model will perform using unseen data. Still, in sample performance also provides some crucial insights. For example, it might give some hint, whether the developed models are overfitting. Furthermore, resulting linear regression models are compared to an appropriate baseline function f_0 in order to show, whether the resulting models actually did learn something. Since the target variables, i.e. the Seven-Factors, are continuous, a simple mean function is chosen as baseline.

Table 5 Comparison of performance measures of baseline function (f_0), multiple linear regression (MLR), KNN regression (KNN), and Random Forest regression (RF)

	f_0	MLR	KNN	RF
Sun and Chill-Out				
R^2_{train}	0.00	0.68	0.78	0.94
R^2_{test}	0.00	0.62	0.61	0.64
$RMSE_{train}$	0.40	0.23	0.19	0.10
$RMSE_{test}$	0.40	0.25	0.25	0.24
Knowledge and Travel				
R^2_{train}	0.00	0.72	0.62	0.85
R^2_{test}	0.00	0.71	0.64	0.70
$RMSE_{train}$	0.32	0.17	0.20	0.12
$RMSE_{test}$	0.33	0.18	0.20	0.18
Independence and History				
R^2_{train}	0.00	0.65	0.46	0.71
R^2_{test}	0.00	0.59	0.58	0.62
$RMSE_{train}$	0.27	0.17	0.20	0.14
$RMSE_{test}$	0.28	0.17	0.18	0.17
Culture and Indulgence				
R^2_{train}	0.00	0.69	0.99	0.79
R^2_{test}	0.00	0.61	0.58	0.67
$RMSE_{train}$	0.33	0.20	0.03	0.15
$RMSE_{test}$	0.35	0.21	0.22	0.20
Social and Sports				
R^2_{train}	0.00	0.28	0.22	0.54
R^2_{test}	0.00	0.22	0.06	0.16
$RMSE_{train}$	0.18	0.15	0.16	0.12
$RMSE_{test}$	0.19	0.17	0.18	0.17
Action and Fun				
R^2_{train}	0.00	0.73	1.00	0.88
R^2_{test}	0.00	0.68	0.63	0.70
$RMSE_{train}$	0.35	0.18	0.01	0.12
$RMSE_{test}$	0.36	0.20	0.21	0.19
Nature and Recreation				
R^2_{train}	0.00	0.80	1.00	0.92
R^2_{test}	0.00	0.77	0.69	0.75
$RMSE_{train}$	0.33	0.15	0.02	0.10
$RMSE_{test}$	0.34	0.17	0.19	0.17

All models have been applied on the expert sample (80% for training and 20% for testing)

Additionally, two more complex and non-linear models, namely KNN and RF, are fitted, in order to challenge the performance of the simple linear model.

In Table 5 the training and test performance of f_0, MLR, KNN, and RF are listed. Note that f_0 is always a constant function. Thus, it does not explain any variance in the factor scores. Therefore, R^2_{train} and R^2_{test} of f_0 are always zero. Training and test performance of the MLR models is close together, which shows that this model is not much overfitting. Whereas the RF models and especially the KNN models are overfitting the training set, i.e. the training performance is much better than the test performance. For example, an extreme case is the KNN model for factor *Action and Fun*, where R^2_{train} is 1.00 (100% of the variance in the factor is explained) and R^2_{test} is 0.63 and also $RMSE_{train}$ is 0.01 (almost perfect) and $RMSE_{test}$ is 0.21. Although both models are well tuned, the overfitting can be a sign of too few training data, but it also shows a potential for enhancement if more data is used.

Overall, the out-of-sample performance of all three models MLR, KNN, and RF are pretty close. Hence, one can expect that they will perform similar if confronted with unseen data. The overall performance of all three models (MLR, KNN, RF) are always better than the simple mean function f_0, which indicates that the models must have learned something out of the data. The difference is in most cases clear to observe, except in factor *Social and Sports*. Here, the $RMSE_{test}$ of f_0 is 0.19 and MLR, KNN, and RF have a $RMSE_{test}$ of 0.17-0.18. This is caused by an uneven distribution of the expert mapping, where 87% of the destinations have scored with 0.5 or 0.75. Hence, a constant prediction of 0.58, like f_0 does, is performing pretty well, but it also means that there is less information to learn from. On the other hand, the models are performing the best in factor *Nature and Recreation*, where $RMSE_{test}$ is 50% smaller than the baseline. The out of sample performance of the KNN model is always a tick worse than the MLR and RF model, whereas there is almost no difference in the performance of RF and MLR. Thus, discarding the KNN model and choosing the MLR model over RF is reasonable since they are performing similar but the MLR model is much simpler to fit and easier to interpret.

MLR delivers promising results by explaining 59–77% of the variance of factor scores in the test set. Except the model for *Social and Sports*, where only 22% of the variance is explained. Such poor performance is caused by an uneven distribution of the factor scores in the expert mapping.

In contrast to the previous analysis, where the focus is predictive performance, now the distribution of the predicted factor scores is analysed. In detail, the factor score distribution of the expert mapping is compared to the distribution behaviour of the predicted factor scores. In order to do so, the MLR model of each factor is fed with the complete data set as input. Then the resulting distribution in factor scores is compared to the one in the expert mapping, which is shown in Fig. 7. This comparison will foster a better understanding of the generalization power of the developed models.

Sun and Chill-Out Here, 49% of the destinations in the complete set are scoring with 0.25. This is not observable in the expert sample. The expert sample shows an increased amount of destinations with score 0 (30%) and 1 (27%), whereas 43% of destinations score either with 0.25, 0.5 or 0.75. A similar but damped behaviour can be observed in the predicted factor scores of the complete set (setting aside the peak at score 0.25).

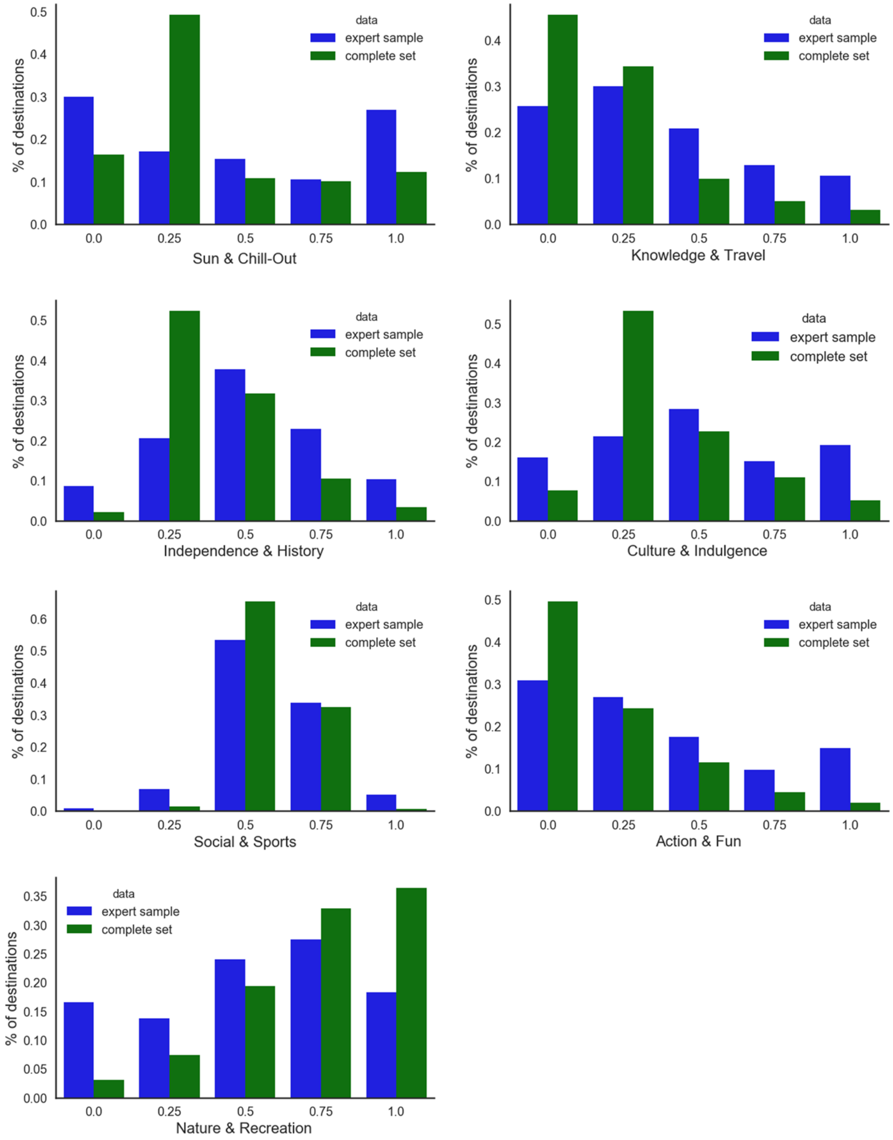

Fig. 7 A comparison of factor score distributions in the expert sample versus the distributions of predicted factor scores of the complete data set

Knowledge and Travel Taking into account the expert sample, the majority of destinations score either with 0 or 0.25 and with increasing factor score the amount of destinations decays. A similar behaviour can be observed in the predicted factor scores of the complete set.

Independence and History Considering the predicted factors of the complete set, once again one can see a peak at score 0.25 like previously in *Sun and Chill-Out*. Besides that, the distribution has more or less a normal shape (bell), similar to the factor score distribution of the expert mapping.

Culture and Indulgence Looking at the predicted factors of the complete set, there is again a peak at score 0.25 (57%), which is not observable in the expert mapping. At score 0.5 and 0.75 the percentage of destinations in the expert sample (28% and 15%) are relatively close to the ones in the complete set (23% and 11%). On the other hand, this is not the case for scores 0 or 1, where the difference is much higher.

Social and Sports The vast majority of destinations score either with 0.5 or 0.75 in both, whereas only few destinations score with 0, 0.25 or 1.

Action and Fun Considering the expert mapping, one can see that the majority of destination score either with 0 or 0.25 and that this amount is decaying the higher the score gets. A similar behaviour can be observed by looking at the predicted factors of the complete set.

Nature and Recreation In the predicted scores of the complete set the amount of destinations is increasing with increasing factor scores. This cannot be observed in the expert mapping. Still, in both, the expert sample and the complete set, most of the destinations are scoring with 0.5 or more.

To sum up, there are some differences in the distributions of factor scores between the manually labelled expert sample and the predicted scores of the complete set. But overall, both show similar trends in the distributions. This shows that the build multiple linear regression models are mimicking the experts quite good. Hence, one can expect a sufficient generalization.

6 Conclusions and future work

This chapter reflects on the outcomes of this study and highlights some managerial implications. Furthermore, limitations of the approach are discussed and a future outline is presented.

6.1 Discussion

In general, one can distinguish between knowledge-poor recommendation techniques (i.e., only basic data like user ratings for items in use) and knowledge-dependent recommendation techniques (i.e., ontological description of users and items, constraints, or social relations and activities of users in use) (Ricci et al. 2015). The picture based approach (Neidhardt et al. 2014, 2015), where both users and items are described through the Seven-Factor Model, belongs to the latter group. Primarily, this work's aim is to identify and explain associations between destination attributes and the Seven-Factor Model to enable an automated mapping of destinations onto the Seven-Factors. Ultimately, such a mapping should enable a match-making between users and destinations. To reach the stated goals an exploratory data analysis, a cluster analysis, and a regression analysis were conducted.

Since the overall aim is not only to project destinations into the seven-dimensional vector space of travel behavioural patterns, but also to explain which attributes of destinations are more important for this purpose, a multiple linear regression (MLR) analysis with step wise variable selection was conducted. Seven models were established, one for each factor of the Seven-Factor Model. The resulting models are providing strong evidence that there is a significant relation between selected destination attributes and the Seven-Factors. Table 6 lists all independent variables (destination attributes) of the fitted MLR models of the Seven-Factors. Note that a minus sign indicates a negative impact on the corresponding factor. For example, the model of the factor *Sun and Chill-Out* consists out of indicators of sun and beach as expected, but there are also indicators of crowdedness, which have a negative impact on the factor. Such model structure is in line with the characteristics of *Sun and Chill-Out*, where crowdedness and mass tourism are negatively associated with the factor.

Furthermore, the developed linear models were challenged by two conceptually different non-linear models, namely random forest regression (RF) and K-Nearest-Neighbour regression (KNN). Additionally, the predictive performances of all three models MLR, RF, and KNN were compared to a baseline function (in this case simple mean of each factor). The evaluation showed that all three models were always better than the baseline function, which indicated that they had learned something out of the data. Although the RF models were excelling in the training phase of the models their performance in the test phase dropped drastically. Thus, the RF models were clearly overfitting the training data. Also, fine-tuning and regularization did not bring any performance gain. Thus, the big difference in training and test performance might be caused by too small samples. Consequently, bigger samples will be aimed in the follow up studies. However, it has been demonstrated that the

Table 6 Used destination attributes in the resulting multiple linear regression (MLR) models

Factor	MLR coefficients
Sun and Chill-Out	*– nightlife, beach and swimming, health resort, sea*
Knowledge and Travel	*sightseeing, mobility, – winter sports resort, - sea*
Independence and History	*culture, sightseeing, – nature and landscape*
Culture and Indulgence	*image and flair, culture, sightseeing, – nature and landscape, old town*
Social and Sports	*sports, – wellness, – sightseeing, peacefulness, mountains*
Action and Fun	*winter sports, – peacefulness, shopping, nightlife, – family, metropolis, sea, – health resort, kite and windsurfing*
Nature and Recreation	*– nightlife, peacefulness, hiking, – sightseeing, – shopping, health resort, – beach*

A minus sign indicates a negative impact on the corresponding factor performance of the MLR model is similar to the performance of the RF model and both are outperforming the KNN model. In the end the MLR model was chosen over the RF model since MLR is simpler to fit and easier to interpret than RF.

Overall, all travel behavioural patterns are well described (59–77% of the variance) by the resulting models, except for the factor *Social and Sports*, where only 22% of the variance can be explained. This is caused by an uneven distribution of scores of the factor *Social and Sports* in the expert sample. Thus, statistically sounder samples will be targeted in future work. Furthermore, the fitted MLR models show a RMSE between 0.17 and 0.25 (in a scale of 0–1). The interpretation of a RMSE value is totally domain dependent. Unfortunately, there is no reference to look at or a rule of thumb to follow, which indicates whether such RMSE is sufficient enough or not. For example, in critical decisions like in medicine a RMSE of 0.01 (in a scale of 0–1) might still be too much. But, for simple recommendation items like for example movies in the Netflix prize case (Bell and Koren 2007) a RMSE of about 0.9 (in a scale of 1–5) might be sufficient. Tourism products are usually more complex and more expensive than movies and thus one might need a lower RMSE than about 20% of the scale. Note that the models here are trained and tested on a small dataset evaluated by experts, which is substantially different to the setting of the Netflix prize. Ultimately, the decision whether the fitted models are applicable or not has to be evaluated within a RSs and will be addressed in follow up studies.

As already mentioned, the MLR models were obtained by using just a small subset of the given data set (i.e., 2%). In order to get a better understanding of the generalization performance of the fitted models with respect to the whole data set the Seven-Factor score distribution of the expert sample was compared to the predicted Seven-Factor score distribution of the whole data set. There was some discrepancy between both distributions but overall, they showed similar trends and thus one can expect a sufficient generalization within the given data set. Most differences were experienced at a score of 0 and 0.25, which might indicate the difficulty in differentiating between "bad" and "very bad" (i.e., between 0.25 and 0). A finer scale than the currently used one (i.e., 0–0.25–0.50–0.75–1) might counter this issue.

Anyway, one should keep in mind that the fitted models are based on a proprietary data set. Although the introduced approach can easily be replicated with other datasets, it brings the disadvantage that for each new data set new models have to be trained and tested (since commonly data sets in the tourism landscape are heterogeneous and not standardized). To counter this generalization issue, a comprehensive tourism product model (partly based on existing ontological representations) will be developed in order to harmonize heterogeneous data sources. Afterwards, analyses conducted in this study will be applied on the harmonized data to obtain a more general solution. First steps in this direction have already been taken.

In addition to the regression analysis (i.e., a supervised learning method, where pre-labelled data is needed) a cluster analysis (i.e., an unsupervised learning method) was conducted. The goal was to find latent, conceptually meaningful structures within the given data set without the need of prior expert knowledge. Six conceptually meaningful clusters were identified, namely *vibrant beach resorts,* *energetic cities, tranquil seaside resorts, peaceful towns, idyllic island villages,* and *ordinary towns*.

The resulting cluster solution is supported by the exploratory data analysis, where destination attributes could be group based on their pairwise correlation (see Fig. 3). The identified groups of attributes are covering following aspects of tourism destinations: *recreational*, *island*, *countryside*, *urban area*, *mass tourism*, and *seaside*. Considering both the six clusters and the groups of destination attributes one can observe a clear contrast of *vibrant* to *tranquil*, *land* to *island*, *seaside* to *inland (urban area)*. Also, the expert sample supports to some extent the six cluster solution, where a reasonable average factor score distribution over the clusters (see Table 3) could be observed. Only in case of the factor *Social and Sports* no clear associations with the clusters could be drawn. The ambiguity of *Social and Sports* is due to an uneven distribution of factor scores in the expert mapping, where 87% of the destinations scored either with 0.5 or 0.75.

The benefit of the cluster solution is that it works totally unsupervised. In other words, clusters can be obtained without the need of a previous manual mapping of experts and only by considering the given data set. Further, the resulting clusters could be addressed directly by RSs, where they can be used as an initial approach for recommending destinations if no further data is available. But the biggest limitation of the clustering approach is mutual exclusivity, i.e. destinations are members of only one cluster. For example, Rio De Janeiro is a huge city offering many cultural, historical, nightlife and entertainment activities, but it is also located at the seaside with famous beaches like Copacabana and Ipanema. Therefore, assigning Rio to just one cluster for example *energetic cities* would not consider customers interested in beach and swimming adequately. Furthermore, people tend to have a combination of travel preferences rather than just one interest (Gretzel et al. 2006; Neidhardt et al. 2014, 2015), where an explicit assignment would have shortcomings.

Finally, the exploratory data analysis and also the developed models showed that the used data set is not able to cover all characteristic aspects of the factors of the Seven-Factor Model. Especially, there were no features indicating independence, the passion for knowledge gain, indulgence, or socialization with locals. This is also pointed out by Glatzer et al. (2018), where they argue that this shortcoming might be the cause of some performance loss in their models. Other data sources might be able to cover the characteristic aspects of the factors of the Seven-Factor Model better and will be used to enhance the models.

6.2 Managerial implications

One of the main goals of the presented work is to automatically describe tourism products within a recommender system. The proposed approach projects tourism destinations into the seven-dimensional vector space of travel behavioural patterns, i.e. it determines the Seven-Factor representation of tourism destinations based on their attributes. This approach, moreover, can easily be adapted to other tourism products.

A main benefit of representing tourism products with just seven factors is low processing costs, for computers as well as for humans (i.e., cognitive costs). In other words, it facilitates the communication of tourism products to customers, between customers, to other companies, and also between co-workers. In addition, from a computational point of view, a matchmaking with just seven factors instead of tens or hundreds of attributes saves process and storage costs and thus leads to low response times and in turn to a better customer experience and satisfaction.

Manually mapping tourism products to the Seven-Factors is resource intensive (i.e., time and expert knowledge). As opposed to this, the presented work proposes an automatic and scalable solution for destinations, which can easily be adapted to other tourism products. This will not only save time and costs with respect to mapping tourism products onto the Seven-Factors, but also with respect to maintaining already mapped tourism products (e.g., a game changing new attraction opens near an already mapped destination). Such an easiness of determination of the Seven-Factors of tourism products can also be seen as an incentive to add more and more tourism products into the recommendation base and in turn to increase the variety.

From the consumers point of view, a representation of tourism products with just seven factors does not only mean easier communication, lower cognitive costs, and better customer experience, but it also increases the recognition potential of the products as well as of the recommender system (i.e., travel agency brand). Furthermore, recommending products, which are not only dependent on previous interactions with the RSs (e.g., collaborative filtering) and which also considers current needs, can lead to a broader diversity and in turn to a better experience and satisfaction.

From a marketing point of view, an automated clustering or classification of existent or new products can facilitate the campaign and package building. One could, for example, design a marketing campaign just for *tranquil seaside resorts*. Furthermore, an automatic package building can be achieved, where destinations, accommodations, activities and more are clustered, packaged and recommended based on the Seven-Factors. For example, the destination *Las Vegas*, the hotel *The Stratosphere Casino, Hotel and Tower*, and the activity base jumping could be grouped and offered to a traveller with a high *Action and Fun* score.

6.3 Limitations and future research

Summing up, the main limitations of this work are caused by the data that is used and the expert sample. In future, higher performance and more robustness of the models will be achieved with statistically sounder and bigger samples. Currently, the fitted models are limited to tourism destinations and their respective attributes (i.e., to a proprietary attribute structure). In order to break this data source dependency and derive a more general solution, future research will consider diverse, heterogeneous data sources and a way to harmonize them.

Right now, moreover, the evaluation of the models is limited to the test set (i.e., 20% of the expert sample, which was not used for model fitting). However, as a next step we plan to evaluate the applicability of the developed models within a running RSs setting.

Finally, since tourism products are considered as rather complex (i.e., they typically combine accommodation, transportation, activities, food, etc.) (Werthner and Ricci 2004) it is also planned to develop an aggregation model to deal with packages of tourism products.

Acknowledgements Open access funding provided by TU Wien (TUW).

References

Beirman D (2003) Restoring tourism destinations in crisis: a strategic marketing approach. In: CAUTHE 2003: riding the wave of tourism and hospitality research, p 1146

Bell RM, Koren Y (2007) Lessons from the netflix prize challenge. Acm Sigkdd Explor Newslett 9(2):75–79

Borràs J, Moreno A, Valls A (2014) Intelligent tourism recommender systems: a survey. Expert Syst Appl 41(16):7370–7389

Braunhofer M, Elahi M, Ricci F (2014) Usability assessment of a context-aware and personality-based mobile recommender system. In: International conference on electronic commerce and web technologies, Springer, Berlin, pp 77–88

Burke R (2007) Hybrid web recommender systems. In: The adaptive web, Springer, Berlin, pp 377–408

Burke R, Ramezani M (2011) Matching recommendation technologies and domains. In: Recommender systems handbook, Springer, Berlin, pp 367–386

Delić A, Neidhardt J, Werthner H (2016) Are sun lovers nervous? In: ENTER 2016: Volume 7 research notes, e-Review of tourism research

Fesenmaier DR, Wöber KW, Werthner H (2006) Destination recommendation systems: behavioral foundations and applications. CABI, Wallingford

Garcia I, Sebastia L, Onaindia E (2011) On the design of individual and group recommender systems for tourism. Expert Syst Appl 38(6):7683–7692

Gibson H, Yiannakis A (2002) Tourist roles: needs and the lifecourse. Ann Tour Res 29(2):358–383

Glatzer L, Neidhardt J, Werthner H (2018) Automated assignment of hotel descriptions to travel behavioural patterns. In: Information and communication technologies in tourism 2018, Springer, Berlin, pp 409–421

Goldberg LR (1990) An alternative "description of personality": the big-five factor structure. J Pers Soc Psychol 59(6):1216

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–871

Gretzel U, Mitsche N, Hwang YH, Fesenmaier DR et al (2006) Travel personality testing for destination recommendation systems. Destin Recomm Syst Behav Found Appl. CABI, Oxfordshire, pp 121–136

Gunawardana A, Shani G (2009) A survey of accuracy evaluation metrics of recommendation tasks. J Mach Learn Res 10(Dec):2935–2962

Hollander M, Wolfe DA, Chicken E (2013) Nonparametric statistical methods, vol 751. Wiley, Hoboken

James G, Witten D, Hastie T, Tibshirani R (2013a) Linear model selection and regularization. In: An introduction to statistical learning, Springer, Berlin, pp 203–264

James G, Witten D, Hastie T, Tibshirani R (2013b) Linear regression. In: An introduction to statistical learning. Springer, Berlin, pp 59–126

James G, Witten D, Hastie T, Tibshirani R (2013c) Statistical learning. In: An introduction to statistical learning. Springer, Berlin, pp 15–57

James G, Witten D, Hastie T, Tibshirani R (2013d) Tree-based methods. In: An introduction to statistical learning. Springer, Berlin, pp 303–335

Kaufman L, Rousseeuw PJ (1990) Partitioning around medoids (program pam). In: Finding groups in data, Wiley, Hoboken, pp 68–125

Matthews G, Deary IJ, Whiteman MC (2003) Personality traits. Cambridge University Press, Cambridge

Mazumder R, Hastie T, Tibshirani R (2010) Spectral regularization algorithms for learning large incomplete matrices. J Mach Learn Res 11(Aug):2287–2322

MediaCT I (2014) The 2014 traveler's road to decision. https://www.thinkwithgoogle.com/_qs/documents/918/2014-travelers-road-to-decision_research_studies.pdf. Accessed 14 July 2017 **(online)**

Neidhardt J, Werthner H (2017) Travellers and their joint characteristics within the seven-factor model. In: Information and communication technologies in tourism 2017, Springer, Berlin, pp 503–515

Neidhardt J, Schuster R, Seyfang L, Werthner H (2014) Eliciting the users' unknown preferences. In: Proceedings of the 8th ACM conference on recommender systems, ACM, New York, RecSys '14, pp 309–312. https://doi.org/10.1145/2645710.2645767

Neidhardt J, Seyfang L, Schuster R, Werthner H (2015) A picture-based approach to recommender systems. Inf Technol Tour 15(1):49–69

Ricci F, Rokach L, Shapira B (2015) Recommender systems: introduction and challenges. In: Recommender systems handbook, Springer, Berlin, pp 1–34

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

Sertkan M, Neidhardt J, Werthner H (2018) Mapping of tourism destinations to travel behavioural patterns. In: Information and communication technologies in tourism 2018, Springer, Berlin, pp 422–434

Tan P-N, Steinbach M, Kumar V (2006) Cluster analysis: basic concepts and algorithms. In: Introduction to data mining, Chapter 8, 6th edn. Peason Addison Wesley, Boston, pp 486–568

Tkalcic M, Chen L (2015) Personality and recommender systems. In: Recommender systems handbook, Springer, Berlin, pp 715–739

Werthner H, Klein S (1999) Information technology and tourism: a challenging ralationship. Springer, Wien

Werthner H, Ricci F (2004) E-commerce and tourism. Commun ACM 47(12):101–105. https://doi.org/10.1145/1035134.1035141

Werthner H, Alzua-Sorzabal A, Cantoni L, Dickinger A, Gretzel U, Jannach D, Neidhardt J, Pröll B, Ricci F, Scaglione M (2015) Future research issues in it and tourism. Inf Technol Tour 15(1):1–15

Woszczynski AB, Roth PL, Segars AH (2002) Exploring the theoretical foundations of playfulness in computer interactions. Comput Hum Behav 18(4):369–388

Zins AH (2007) Exploring travel information search behavior beyond common frontiers. Inf Technol Tour 9(3–1):149–164