# Tourism Profile Measure for Data-Driven Tourism Segmentation

Guillaume Guerard[1,2*†], Quentin Gabot[1†] and Sonia Djebali[1]

[1*]Léonard de Vinci Pôle Universitaire, Research Center, Paris La Défense, 92 916, France.
[2]LI-PARAD Laboratory EA 7432, Versailles University, Versailles, 78035, France.

*Corresponding author(s). E-mail(s): guillaume.guerard@devinci.fr;
Contributing authors: quentin.gabot@edu.devinci.fr;
sonia.djebali@devinci.fr;
†These authors contributed equally to this work.

**Abstract**

The Digital Revolution has brought about profound changes in research within the tourism segmentation field. The ease of grasping tourists' behaviors is facilitated by the digital traces left on social networks. Existing studies focusing on tourists' digital traces typically apply clustering algorithms to the tourism context. This paper introduces a novel measure, named Tourism Profile Measure for determining tourism segmentation, also known as tourism profiling. The approach involves establishing a new clustering algorithm that centers on stays conducted by tourists, utilizing both the context and content of the trips. The proposed measure is then simulated and experimentally evaluated using a real dataset across various periods and diverse nationalities, particularly in the context of the French capital, Paris.

**Keywords:** Tourism Segmentation, Clustering, Similarity Measure, Machine Learning

## 1 Introduction

Tourism has experienced substantial growth since the late 1980s. In 1950, there were merely 25 million tourists, a figure that surged to 278 million by 1980, and further skyrocketed to 1.5 billion tourists in 2019. This remarkable surge in tourist numbers

has given rise to significant global consequences, including the amplification of the global gross domestic product attributed to tourism and the annual profits yielded by the tourism industry. One notable catalyst for this transformation has been the democratization of the Internet. The accessibility and immediacy of online communication have complemented the geographical proximity facilitated by the 1980s tourism boom. Consequently, tourism serves as a potent symbol of globalization, ushering in formidable challenges.

The challenges associated with tourism are predominantly logistical and economic. The symbiotic relationship between tourists as demand generators and the tourism industry as suppliers necessitates the development of various methodologies for predicting, analyzing, and synthesizing tourist behaviors [1, 2]. To illustrate, tourism management has explored aspects of tourist profiles [3, 4]. However, it is essential to recognize that the tourism management approach differs from that of computer science. As indicated in earlier research [5], a novel challenge stemming from the era of Big Data is the validation of tourism management studies through computer science methods.

Numerous studies have already cited examples of tourism's evolution, such as the United Kingdom's increasing reliance on social networks for holiday planning. Remarkably, over 50% of tourists alter their initial travel plans after engaging with social networks [6]. In recent years, tools like digital traces and user profiles on platforms such as TripAdvisor and Instagram have offered innovative approaches to address these challenges. Utilizing these new data sources has presented a substantial challenge for the field of tourism research.

This paper aims to address the analysis and synthesis of tourist behaviors by considering tourist user profiles and conducting user clustering to identify communities of tourists with similar behaviors. These communities offer essential insights into understanding standardized tourist profiles and, in turn, provide valuable information about tourist behaviors.

A precise definition of the concept of a user profile is established, which can exhibit substantial variation based on the context. In this context, a user profile is defined as a comprehensive summary encompassing user behaviors, interests, characteristics, and preferences [7]. In the specific case of this tourism study, a user profile is construed as a composite of both the user's static/demographic profile and a summary of their past travel experiences. It is vital to distinguish this user profile definition from the concept of user profiling, which pertains to the systematic collection, organization, and inference of user profile information [8].

Traditional literature often discusses two prevalent methods for user profiling: content-based filtering and collaborative-based filtering [9, 10]. However, these methods do not align with the objective, which is to partition users into clusters for the extraction of standard tourist profiles. Both content-based and collaborative-based filtering predominantly focus on predicting user behaviors, with the former relying on individual user history analysis and the latter involving the pooling of similar users' preferences.

Contrary to these conventional approaches, this paper centers on the application of clustering algorithms for user profiling [8, 11]. The primary challenge of any user

clustering method lies in the identification of clustering components that serve as suitable criteria for assigning users to clusters. In the proposed method, users' trip summaries are clustering components, as these summaries contain vital information about the trip context, including the season and duration, as well as the trip content, which encompasses the points of interest visited. While some studies in the literature consider the user's static/demographic profile as a clustering component, this kind of approach remains unsuitable in the presented context, as demographic information (gender, age, or nationality) is distinct from user behavioral aspects and should be treated as user characteristics [12–14].

Moreover, other studies delve into additional clustering components such as the order of visits to points of interest, semantic analysis of comments, or geospatial data from photos [15–18]. However, these approaches do not align with this paper, as they focus on granular details derived from user trips, neglecting the pivotal elements of trip content and context.

To gain deeper insights into tourist behaviors, a novel approach is presented in this paper: the establishment of standard tourist profiles through user clustering based on digital traces. Specifically, this study leverages the user profile concept to construct comprehensive tourist user profiles, encompassing both the user's static/demographic information and their travel history summaries. Subsequently, by utilizing these trip summaries, distances are computed between users, and perform unsupervised clustering to delineate distinct user clusters. These clusters provide a foundation for the determination of standard tourist profiles and enable various analytical explorations.

The key contributions of the presented work can be succinctly summarized as follows:

- In-depth exploration of the tourism landscape through the extraction of tourist profiles from the vast reservoir of tourism data within the Big Data context.
- Validation of results through comparisons with tourism management reports to ensure the robustness and relevance of the presented methodology.
- A versatile approach to tourist profiling adaptable to various geographical regions.

The structure of this paper unfolds as follows: Section 2 delves into the prior research in this field. Section 3 elucidates the materials and methods employed. Section 4 defines the essence of tourism and the experiences of tourists. Section 5 expounds on the details of the clustering algorithm. Section 6 provides insights into experiments done on the French capital Paris. Lastly, in Section 7, conclusions are drawn from the findings.

## 2 Related Work

The concept of market segmentation has been integral to marketing since the early 1960s [19, 20]. In the realm of tourism, segmentation serves as a strategic tool aimed at addressing the heterogeneity among tourists by categorizing them into market segments, each comprising individuals who share similarities and are distinct from members of other segments [21]. This section delves into the related work about

tourism segmentation, and its associated challenges, and provides contextual insights into each of these challenges.

## 2.1 Context

A comprehensive review of the literature reveals several gaps in tourism segmentation:

- *Lack of consistency in segmentation variables*: One notable challenge is the absence of a universally accepted set of variables for segmenting the tourism market. Different studies employ diverse criteria to identify segments, resulting in a lack of comparability across findings and a restriction on the generalizability of results.
- *Limited use of advanced analytical techniques*: Many studies still rely on simplistic descriptive statistics for identifying and profiling tourism segments, which might not capture the full complexity of the market. There is a growing need for more sophisticated techniques, such as clustering algorithms or latent class analysis, capable of pinpointing subgroups of travelers with similar preferences and behavioral patterns.
- *Over-reliance on demographic variables*: The overemphasis on demographic variables like age, gender, and income for segmenting the tourism market can be limiting. These variables may not always be the most meaningful or predictive in understanding tourist behavior. Thus, further research is needed to focus on psychographic and behavioral variables, including travel motivations, personality traits, and decision-making processes.
- *Limited attention to emerging trends and technologies*: Many tourism segmentation studies concentrate on traditional forms of tourism, like leisure or business travel, often overlooking emerging trends such as adventure tourism, sustainable tourism, or digital nomadism. Addressing these evolving trends and technologies is vital to adapt segmentation strategies effectively. Notably, ongoing work explores photographic tourism in another paper [22], while other forms of tourism will be addressed in subsequent studies.

Two principal methods exist for classifying tourists in the context of segmentation: the conceptual approach, which leads to a typology where grouping criteria are predefined (ad-hoc approach), and the data-driven approach, which results in a taxonomy. The data-driven approach is empirical by nature, involving the application of quantitative techniques to empirical datasets to derive tourist groupings (posthoc approach).

Over the past decade, the posthoc approach has gained prominence as the most popular method for tourism segmentation. This is attributed to the emergence of unsupervised learning techniques and the diversity of available models [23]. In cluster analysis, a group of tourists is subdivided into more or less homogeneous subgroups based on a similarity measure [24]. Consequently, cluster analysis shares a common framework with market segmentation, as it focuses on ensuring that the similarity among tourists within a subgroup exceeds the similarity between tourists in different subgroups.

## 2.2 Data-driven Approach Challenges

Blanco-Moreno et al. [25] conducted a comprehensive analysis of 1,152 studies spanning from 1996 to 2023, focusing on the utilization of Big Data, particularly digital traces, in tourism marketing methodologies. Surprisingly, out of this extensive pool, only 75 papers were dedicated to the exploration of tourism segmentation. This relatively low number stands in stark contrast to the abundance of research on mobility patterns, satisfaction, destination marketing, and tourism behavior within the same timeframe. The scarcity of studies in tourism segmentation can be attributed to the challenges associated with overcoming four distinct gaps in the existing literature:

1. *Acquisition of an Empirical Dataset*: The foundation of a data-driven approach relies on obtaining an empirical dataset.
2. *Selection of Relevant Segmentation Variables*: The choice of segmentation variables should be well-justified to avoid unnecessary high dimensionality and enhance the explainability of results.
3. *Selection of an Unsupervised Learning Model*: The selection of an appropriate unsupervised learning model is crucial. Various clustering algorithms, such as connectivity-based, hierarchical-based, distribution-based, and density-based, are available for segmenting tourists into consistent groups, and the choice must be justified in light of the specific characteristics of the data.
4. *Determination of a Measure of Similarity*: The choice of a metric or similarity measure and the number of clusters need to be carefully analyzed. Transparency in these choices is vital, given that clustering results are often complex to interpret. The relevance of the results can be assessed through these choices.

These four prerequisites present methodological challenges. Dolnicar et al. [26] have offered valuable recommendations to enhance the design of data-driven segmentation methods by addressing recurring issues in the literature.

The use of data should be carefully considered, and the selection of variables should always be justified. This approach aids in avoiding unnecessary high dimensionality and simplifies the interpretation of results. Since many clustering algorithms are available, their selection should be adequately justified. Different algorithms have specific characteristics, and these should be taken into account before application. Key technical issues, including the choice of the similarity metric and the number of clusters, should be thoroughly analyzed. Given that clustering method results can be challenging to interpret, transparency in these choices is essential to assess the relevance of the results.

After implementing a clustering algorithm, both the reliability and validity of the results should be demonstrated. Reliability is validated by repeatedly running the clustering algorithm and ensuring the stability of results across repetitions. Validity is tested by computing internal measures on clusters and comparing the results with findings from various tourism management studies.

While numerous studies have proposed data-driven tourism segmentation methods, only a few have attempted to tailor their approaches to the specific characteristics of tourism, as highlighted by D'Urso et al. [27]. To provide a comprehensive overview of current methods, this discussion is divided into four stages: the empirical dataset,

the segmentation variables, the unsupervised learning model, and the measure of similarity.

## 2.3 The empirical dataset

The initial challenge in implementing a data-driven tourism segmentation method is the acquisition of an empirical dataset. As Gauch et al. [28] point out, data can be gathered using either implicit or explicit methods.

The explicit method necessitates actively soliciting information from tourists through surveys, registration processes, and forms. It is particularly useful for obtaining static information about tourists, such as demographic and geographic details. However, this method is not without its challenges. It depends on the willingness of tourists to respond to questions, and it can consume substantial time and resources for deployment [8]. Notably, tourism management and marketing studies have historically relied on explicit data acquisition methods and continue to do so [29].

In contrast, the Digital Revolution, coupled with the rise of social networks, has introduced a novel approach to data collection, referred to as the implicit method [30]. Implicit data acquisition involves the use of intelligent agents or data mining techniques to analyze tourist activities and is well-suited for capturing dynamic information about tourists, including their behaviors. This method is commonly employed in computer science approaches [31]. A hybrid approach, which combines both implicit and explicit methods, can be adopted to gather static and dynamic information effectively [32]. In the context of efficiently segmenting tourists, the acquisition of data should ideally follow a hybrid approach.

This means that when aiming to segment tourists effectively, it is prudent to obtain data using a hybrid method that combines the strengths of implicit data acquisition for capturing dynamic aspects and explicit data acquisition for securing static information.

In the recent study by McKercher et al. [33], an insightful analysis delves into the comparison and contrast of various segmentation techniques—geographic-based, motivation-oriented, demographic-focused, behavior-centric, and hybrid models—with the aim of determining which method best captures the nuances of tourist behaviors. The hybrid approach emerges as the most effective, albeit with the prerequisite of both a priori and a posteriori data from tourists. The authors observe that while geographic and demographic-based methods yield the most diverse patterns, this diversity poses challenges during analysis. As an integral recommendation when selecting variables, they emphasize the importance of exercising caution regarding the level of detail associated with each variable. Hence, the proposed method must guarantee the availability and quality of such variable.

Once the empirical dataset is at hand, the critical task of selecting segmentation variables arises. The choice of these variables is inherently tied to the specific objectives of the segmentation method. For instance, Abbasi et al. [15] proposed segmenting tourists based on sentiment analysis of their reviews, while Rodriguez et al. [18] suggested a segmentation method based on tourist localization. In such cases, the selection of segmentation variables naturally varies to align with the intended goals. Consequently, it is impractical to propose a universal set of segmentation variables

applicable to all situations. Nonetheless, as recommended by Dolnicar et al. [26], a clear justification should underpin the selection of these variables.

Beyond justification, another concern regarding segmentation variables pertains to their types. The type of variables used has a direct impact on the clustering process. As observed by D'Urso [27], the majority of methods have favored numerical variables as segmentation variables, potentially leading to a loss of information contained within categorical variables. In practice, many have sought to address this by converting categorical variables into numerical ones through encoding [32]. However, it's worth noting that such a practice is not recommended in the context of clustering, except for ordinal variables, as arbitrary numerical values can distort the distance computation between instances [34].

Based on our extensive literature review, we assert that while the selection of variables plays a pivotal role, the manner in which data is represented is equally critical to guarantee relevance and readability throughout the unsupervised learning process. The proposed approach not only considers the variables themselves but also addresses the efficient representation of these variables within a database.

## 2.4 Unsupervised learning model

Clustering algorithms can be either model-based [35] or non-model-based [36]. Past data-driven tourism segmentation studies have proposed both model-based [37] and non-model-based methods [18]. Model-based methods are almost always part of a recommender system, as they use a rating feature to construct a 3-dimensional matrix (user-rating object) to perform collaborative filtering. The scope of this study differs from the desire to base a segmentation on actual behavioral features such as POIs visited.

Non-model-based models used are, most of the time, either K-Means or Hierarchical algorithms [27]. The choice of an unsupervised learning model is partly tied to this of segmentation variables, as seen previously. Indeed, each model may treat data types differently from another. As pointed out by D'Urso et al. [27], data-driven tourism segmentation studies have rarely used mixed data clustering algorithms to perform their segmentation. This is a direct throwback to those discussed in Section 2.1 about segmentation variables. To tackle this issue, mixed data clustering algorithms have been proposed such as: partitional algorithms [38], hierarchical algorithms [39], model-based [40] and neural network-based [41] and are classified by Ahmad et al. [42].

This paper refrains from an exhaustive exploration of each method and its respective merits and drawbacks. Instead, the emphasis of the approach lies in addressing how to manage data, conduct analyses, and extract valuable information. The discussion within this paper does not center on debating unsupervised algorithms. However, it does provide documentation and rationale for the selected choice.

## 2.5 Main contributions

In this research, the primary focus is to address the unique challenges associated with metric and similarity measures, to introduce a specialized metric tailored for the domain of tourism segmentation. To the best of our knowledge, this study marks

the pioneering effort in adopting a metric measure specifically designed to align with the nuances of the tourism context. Implementing a hierarchical clustering algorithm alongside the innovative metric aims to reveal significant tourism segments by extracting insights from tourists' experiences. The method adheres to the fourfold recommendations outlined earlier, ensuring robustness and coherence in the obtained results. Additionally, a comprehensive assessment is conducted, comparing clusters with findings from tourism management studies to strengthen the validation of outcomes.

The research outlined in this paper introduces a data-driven tourism segmentation approach characterized by the following distinctive features:

1. *Hybrid Data Collection Method*: The methodology employs a hybrid data collection approach, seamlessly integrating the strengths of implicit and explicit data acquisition methods. This approach demonstrates exceptional proficiency in managing extensive datasets while capturing both static and dynamic information relevant to tourism.
2. *Relevant Segmentation Variables for the Tourism Context*:Extensive data analysis forms the foundation for identifying and defining the most pertinent segmentation variables tailored to the intricacies of the tourism domain. A curated selection of variables, with a focus on the most contextually significant ones, is meticulously detailed in this paper.
3. *Incorporation of Non-Demographic Variables*: Departing from the convention of predominantly relying on demographic variables, the presented method takes a pioneering step by embracing a diverse array of non-demographic variables. These encompass data sources such as social media interactions, online reviews, and location-based data, providing a richer and more nuanced perspective on traveler preferences and behavioral patterns.
4. *Identification of Emerging Segments*: One of the distinguishing capabilities of the presented model is its agility in identifying emerging tourism segments. By being adaptable to various datasets, it can uncover new trends and emerging patterns that have developed over time. The presented method conducts thorough information retrieval within each segment, thereby shedding light on the specific type of tourism it represents and its distinctive characteristics.

This paper elaborates on the methodology, elucidating its various components and highlighting its advantages in addressing gaps within the field of tourism segmentation. Pursuing these objectives, the research aims to make a substantial contribution to the field of tourism segmentation, presenting a novel approach that can better capture the diversity and evolving nature of tourism experiences.

# 3 Materials

In this article, tourism segmentation is analyzed through the digital traces left by tourists' use of social networks. Digital traces refer to the digital data left by tourists on these social networks. Tourism data contains information about the tourists, the places they have visited, and their interactions. A tourist is identified by his *personal*

*data* such as his age, his nationality, and his gender. The tourist visits places resulting in a set of stays. A place is characterized by a name, a coordinate (longitude and latitude), a type (hotel, restaurant, attraction), and reviews given by tourists. Each place was aligned with administrative areas (GADM)[1]. Since the tourist leaves digital traces at various places, it is necessary to define the time frame of a stay before analyzing its context and content. This section presents how to transform the raw digital traces into useful materials.

## 3.1 Static and dynamic features

The data acquisition process falls beyond the scope of this article. It is worth noting that a raw dataset comprising digital traces is not inherently suitable. Rather than relying directly on digital traces, a transformation process results in the creation of two distinct data structures, inspired by the concept of a user profile as presented by Ping et al. [43].

**Tourist's Static Data**: This category pertains to information regarding the tourist's identity, essentially answering the question, "Who is the tourist?" It encompasses personal details such as age, nationality, and gender. Given that this information is readily available within digital traces, a dataset about tourists' static data is created. It's pertinent to note that, for this study, the tourist's static data is defined as encompassing age, nationality, and gender. Future studies may incorporate additional variables like salary or marital status, but they are not considered in the current research.

**Tourist's Dynamic Data**: This category revolves around the actions and movements of the tourist, essentially addressing the question, "What does the tourist do?" Dynamic data encapsulates a tourist's movements and activities during their travels, with "travel" defined as the timespan that extends from their departure from their home to their return. The terms "movement," "activities," and "while traveling" may seem inherently ambiguous. This leads us to introduce a crucial concept: the "stay."

The concept of a "stay" allows us to delineate the boundaries and context of a tourist's actions more precisely. By doing so, the following questions are addressed: Should a tourist's movements and activities be analyzed independently from any spatiotemporal criteria? Is the generic notion of "travel" adequately relevant? These questions have prompted the development of the concept of a "stay."

## 3.2 Tourist's Stay

Digital traces left by a tourist are valuable as they offer insights into the places they have visited. These places are characterized by several attributes, including a name, geographical coordinates (longitude and latitude), a category (e.g., hotel, restaurant, or attraction, depending on the platform), and reviews submitted by tourists.

Upon collecting these digital traces, the tourist's travel paths are constructed. However, the concept of a "travel" may be overly broad, making it challenging to discern a tourist's specific tendencies. For instance, a tourist who visits an urban center and then a rural area is likely to exhibit different behaviors due to the disparate reasons

[1] GADM:https://gadm.org/index.html

for visiting these places. To address this issue, the concept of "stay" is introduced. In this framework, digital traces are now redefined as reviews authored by tourists about the places they have visited.

A tourist's travel can be viewed as a sequence of stays, where each stay represents a chronological sequence of visited places, typically denoted by reviews and/or photos. More precisely, a stay is a consecutive span of days during which the tourist posts at least one review per day within the same geographical area. The geographical area is mostly defined by the limitation of a study, i.e. depending on the selection of administrative divisions. If the tourist does not post a daily review or submits a review for a place located in a different area, the current stay is considered terminated, and a new stay begins when the tourist posts a fresh review. This temporal and spatial condition is pivotal for maintaining consistency within each stay, ensuring that a stay does not encompass reviews made under different circumstances.

However, it's worth considering whether it is realistic to split a stay simply because a tourist takes a day off from posting. In reality, a tourist may refrain from posting, either written reviews or photos, for a brief period while still being within the same stay. In such cases, it is reasonable to merge two stays if the gap between them does not exceed 7 days (while still adhering to the spatial condition). This approach acknowledges the practicality that a stay is composed of reviews that are, at most, 7 days apart, in line with findings by Gossling [44].

The presented method consists of merging two stays to form a single one if the following conditions are met (in addition to the spatial condition):

$$\Delta B \leqslant \Delta S_i \,\&\, \Delta B \leqslant \Delta S_j \,\&\, \Delta B \leq 7 \tag{1}$$

Where $\Delta B$ represents the time between the two stays, $\Delta S_i$ and $\Delta S_j$ respectively represent the $i^{th}$ and $j^{th}$ stay's duration.

Having defined the notion of *stay*, a dataset of stays is built as shown in Figure 1, where every set of a tourist stays corresponds to his dynamic data.

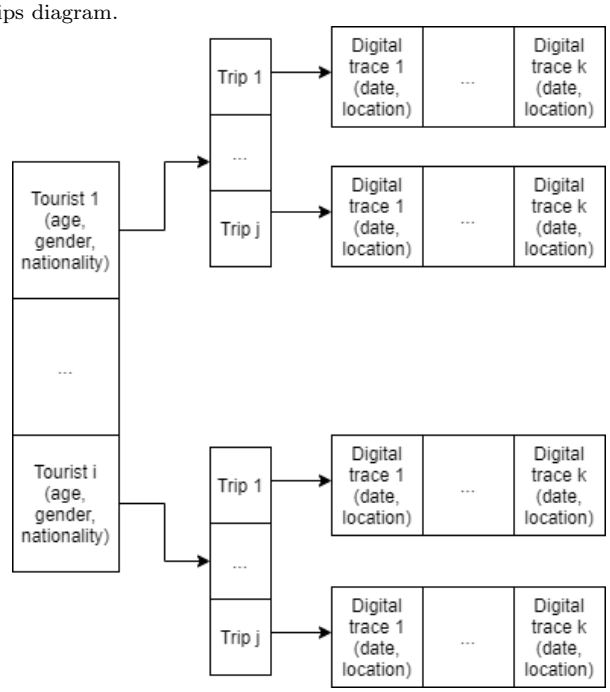
## 3.3 Dynamic Data Overview

Tourist behaviors and decision-making are significantly shaped by a set of external factors collectively referred to as "contextual factors" [45]. These contextual factors encompass the broader environment in which a tourist operates, including considerations such as the season, weather conditions, duration of visits, social dynamics, and more. Unlike certain data, such as places and nationalities, contextual factors are not explicitly embedded within digital traces. Therefore, it becomes imperative to enrich the data to harness the potential of these contextual factors.

To achieve this objective, a series of stays for each tourist is constructed based on their digital traces. For each stay, both "context" and "content" are defined, as delineated below:

**Context**: This encapsulates the passive information about the stay, essentially answering the question of how the tourist conducted their stay. The context comprises the array of contextual factors that influenced the tourist's experience.

**Fig. 1** Tourist trips diagram.




**Content**: In contrast, the content pertains to the active information concerning what the tourist engaged in during their stay. This delves into the specific actions and activities undertaken by the tourist.

It's important to acknowledge that while contextual factors contribute to the "context" of a stay, there may be computational challenges if the list of reviewed places remains static (utilizing the entire list of reviewed places would be resource-intensive in terms of time and memory). To circumvent this, a classification ontology is introduced that facilitates the definition of the "content" of a stay, ensuring a more efficient and structured representation of a tourist's activities and experiences.

## 3.4 Context of a stay

Within the framework of each tourist's stay, a set of contextual factors plays a pivotal role. These contextual factors fall into two categories: "push factors" and "pull factors."

Push factors are elements that motivate tourists to leave their home country and embark on a journey. These factors encompass natural motivations, such as a desire to experience a different climate, as well as institutional motivations, which may include considerations like school vacations. In contrast, pull factors are the elements that entice tourists and are associated with the destination area [46]. These factors encompass attributes like the climate, cultural events, or sports seasons that make a particular location appealing to tourists.

In the investigation of tourism segmentation, focus will be placed on incentive and attraction factors derived from metadata within reviews and information about places. Key indicators, namely the season and the length of stay, will be utilized to study these factors.

### 3.4.1 Seasonality

Seasonality plays a significant role in both incentivizing and attracting tourists to various destinations. However, determining the season of a tourist's country of origin can be a complex task for two primary reasons.

Firstly, one of the challenges arises from the limited information available about the specific origin of each tourist. While data on the nationalities of tourists are available, precise information regarding their exact place of residence are missing. For instance, a tourist with French nationality may actually reside in South Africa, in which case their true place of origin is South Africa rather than France.

Secondly, comparing the season of a tourist's home country with that of the country they are visiting is often an impractical endeavor. It can be challenging to align these two seasons accurately.

In light of these complexities, this approach will primarily consider the season of the places visited by tourists. This information can be relatively straightforward to deduce from the dates of the beginning and end of a stay, along with the destination country. This more granular approach allows us to work with data that is readily available and can provide valuable insights into the seasonality of a tourist's travels.

### 3.4.2 Duration of a stay

The duration of a stay is calculated during the creation of a stay. The duration is equal to the date difference between the first review of the stay and the last review of the same stay.

## 3.5 Content of a stay

The content of a tourist's stay is constructed from the diverse array of locations they visit during their travels. However, the challenge lies in how to effectively manage and handle the vast and varied locations found within touristic areas. This issue involves determining the most suitable approach for handling the abundance of different places and points of interest that tourists may explore during their stays.

### 3.5.1 Locations limitations

At this stage, locations are primarily defined by a set of parameters, including coordinates, type of location (e.g., hotel, restaurant, attraction), and name. However, in practice, the name parameter is often the most readily available and frequently the sole parameter used to define locations. Relying solely on the name parameter can result in a significant loss of valuable information. To address this issue, it's essential to explore the utilization of additional parameters, such as the type of location, which is available in the database. Moreover, this parameter can be further extended into a comprehensive location ontology.

Ontologies for classifying locations have been proposed in various studies [14, 47]. Drawing inspiration from these prior works, our own ontology will be created. In the subsequent sections of the paper, locations will be primarily defined by their classes, enabling a more effective means of capturing similarities between different trips.

Examining locations on an individual basis can lead to a problem of sparsity. For instance, in studies that deal with tourism recommender systems, a users/locations matrix is often employed, and techniques like Latent Profile Analysis are applied to it [10, 48]. However, as some research has highlighted, this matrix approach can be limiting in the tourism context. Tourists tend to visit only a subset of the numerous locations available in a given place, resulting in a highly sparse users/locations matrix. This sparsity issue can negatively impact the quality of the results.

By employing location classification, two significant challenges are addressed: the limited meaningfulness of certain parameters, such as the name, and the issue of sparsity. The following section will delve into the precise implementation process of this approach.

### 3.5.2 Locations ontology

The ontology designed for locations is centered around the type of location. The ontology structure comes from studies who proposed their point of view [14, 47, 49]. The ontology is centered around two levels; the first level is composed of six main concepts : "Cultural Heritage", "Cultural Buildings", "Food & Services", "Entertainment", "Viewpoints", "Nature". Those concepts represent almost every possible type of location while remaining general; the second level is composed of several sub-categories for each concept.

**Table 1** Ontology of places.

| **Category** | **Subcategory** |
|---|---|
| Heritage | Monuments, Parks and Gardens,<br>Urbanism (neighborhoods, bridges, cemeteries, streets) |
| Cultural Buildings | Art galleries and Museums, Holy sites and Places of worship,<br>Historic buildings, Theaters and Auditorium |
| Food and Services | Shops, Restaurants and Bars, Gastronomy, Hotels |
| Entertainment | Music buildings (concerts, discotheques), Cinemas,<br>Amusement park, Sports |
| Viewpoints | (no sub-categories) |
| Nature | Woods, Watering place (river, lake), Beaches, Mountains |

The categorization of tourist places is a meticulous task that requires expert human intervention to accurately reflect the complexity of each location. Each tourist place must be assigned to at least one category and one subcategory within the provided ontology. It's important to note that a single place can belong to multiple categories and subcategories simultaneously. For instance, "The Cathedral of Notre-Dame" may be classified under the "Heritage" category as well as the "Cultural building" category.

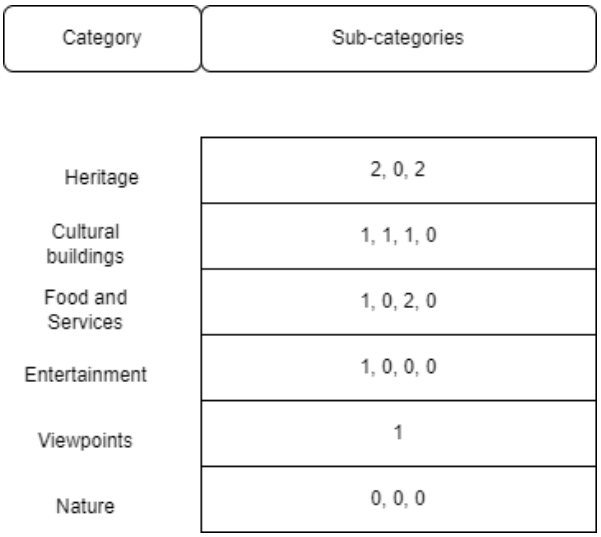
**Table 2** Example of Ontology of places.

| Locations | Category1 | Category 2 | Category 3 |
|---|---|---|---|
| Notre-Dame | Heritage<br>-Monuments | Cultural<br>- Holy sites<br>- Historic Building | |
| Louvre Museum | Heritage<br>- Urbanism | Cultural<br>- Museums | |
| Galleries Lafayette | Heritage<br>- Urbanism | Services<br>- Shops<br>- Gastronomy | Viewpoints |
| Moulin Rouge | Heritage<br>- Monuments | Service<br>- Gastronomy | Entertainment<br>- Music |

To facilitate the aggregation of location categories for trips, a classification vector is introduced. This vector is comprised of six sub-vectors, each corresponding to one of the main concepts within the ontology. The length of each sub-vector is determined by the number of sub-categories associated with the respective main concept. As a result, each trip's content is represented by a classification vector in which the location categories visited during the trip are transcribed.

Here's an example to illustrate this concept: Imagine a tourist's stay during which they visit "The Cathedral of Notre-Dame," "The Louvre Museum," "Galleries Lafayette," and dine at "Moulin Rouge." These places are associated with various categories and subcategories, as detailed in Table 2. The resulting classification vector, which is a summation of counts across subcategories, is depicted in Figure 2.

**Fig. 2** Classification vector.



## 3.6 Personal information

By considering tourists' digital traces, a dataset is built where each entry represents a digital trace left by a traveler, over 8 features: *user ID*, *date creation*, *name*, *longitude*, *latitude*, *age*, *nationality* and *gender*. Nevertheless, this dataset is not suited to depict tourists' behavior to segment them most efficiently. Rather than focusing on tourists' actions (here depicted by their digital traces), the presented study focuses on tourists themselves (which include their actions).

In this section, the data will be presented, as well as its context and its peculiarities. By taking advantage of the Digital Revolution paradigm, it is nowadays possible to consolidate a dataset based on digital traces left by tourists during their travels; such as reviews, photos, or posts. Those digital traces contain various metadata, such as:

- the *user ID* of the tourist who posted the digital trace;
- the *date creation* when the digital trace was posted;
- the *place ID* of the place (also called point of interest or POI in the literature) concerned by the digital trace;
- the place's geographical coordinates: *longitude* and *latitude*;
- the *age* of the tourist who posted the digital trace;
- the *nationality* of the tourist who posted the digital trace;
- the *gender* of the tourist who posted the digital trace.

## 3.7 Dataset

To summarize the proposed approach (see Figure 3), a tourist profile comprises two key components:

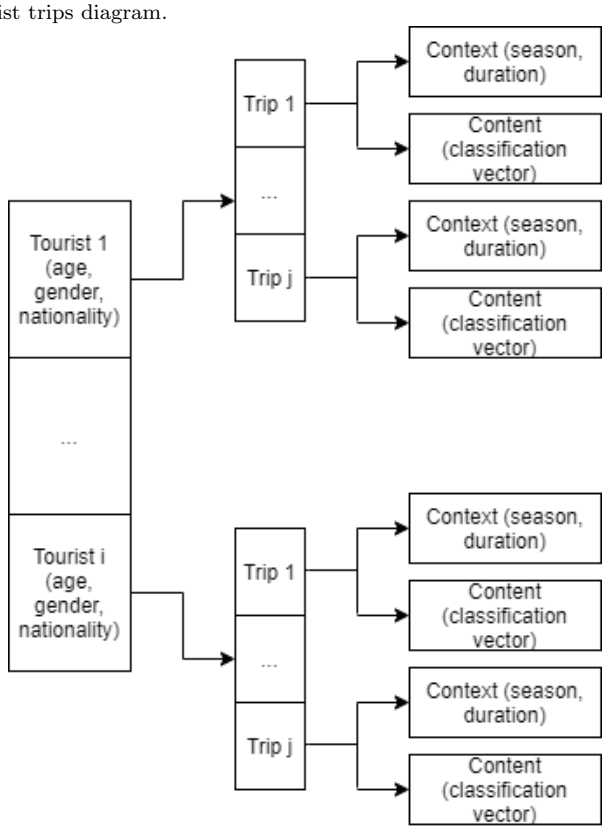
- Static Profile: This component contains information about the tourist's identity, answering the question of "who is the tourist?" It includes demographic details such as age, nationality, and gender.
- Dynamic Profile: This component contains information about the tourist's actions, answering the question of "what does the tourist do?" The dynamic profile is built around the concept of a "stay", which represents a chronological sequence of places visited during a single travel experience. The dynamic profile is itself composed of the Context and the Content of each stay.

## 3.8 Data consistency

In the study, specific constraints are applied to the data to ensure the quality and reliability of the information being analyzed:

1. **Minimum Threshold for User Activity**: Inclusion in the analysis requires users to have a minimum number of reviews, ensuring active engagement with the database. Specifically, the set minimum threshold is four comments in this paper.
2. **Consideration of Major Locations**: Given the intensive processing involved, the proposed method focuses on major locations, considering a subset that is significant and relevant to the analysis. A threshold based on the location's support (the number of times an item appears in a dataset) is defined. The threshold is fixed

**Fig. 3** Final tourist trips diagram.




according to the recommendation of Philippe Fournier-Viger, a renowned researcher in data mining: $(x * e^{(-0.4x-0.2)}) + 0.2$ where $x$ is the number of elements in the dataset.

3. **Discarding Users Without a Referenced Nationality**: Users without a referenced nationality are excluded from the study. This decision is based on the understanding that such users do not contribute relevant information to creating standard tourist profiles, considering nationality as a key demographic factor.

These constraints are applied to ensure that the data used in the study is both manageable for analytical methods and relevant for the creation of tourist standard profiles. A high-quality dataset well-suited to the objectives is achieved.

# 4 Methods

The flowchart in Figure 4 provides an overview of the proposed research methodology. The dataset primarily consists of stays, each defined by its context and content. The main objectives of this paper are to gain insights into tourism behaviors and segment tourists based on their behavior. To achieve these objectives, a two-step approach is designed building on the data processing steps described in the Materials section.

Here is a breakdown of the presented approach:

1.
2. Clustering Algorithm: In this step, the introduction of a new metric termed the "Tourist Profile Measure" (TPM) and the utilization of a clustering algorithm facilitate the assignment of stays to clusters based on their similarity.
3. Tourist Segmentation: Drawing from the clusters of stays, segments of tourists with similar travel behavior are created, facilitating the retrieval of valuable information from these segments.

After providing a detailed explanation of the stays dataset, let us present the clustering algorithm stage. This will include an introduction to the presented metric, the TPM, and an explanation of the measures used to evaluate the quality of clustering results.

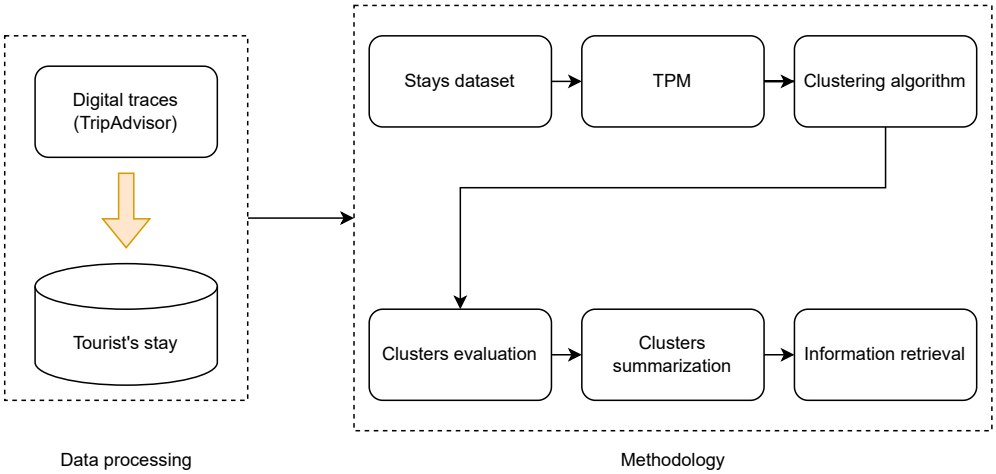



**Fig. 4** Flowchart of the method.

## 4.1 Tourist Profile Measure (TPM)

To obtain meaningful clusters from the stays dataset, it's essential to have a metric that quantifies the closeness or distance between stays. Any method needs a measure on the similarity or dissimilarity between pairs of stays within the same cluster. Moreover, the metric should be capable of handling mixed data and leveraging tourists' experiences. With these considerations in mind, a new metric is proposed, called the TPM to calculate the distance between two stays, taking into account both their context and content. TPM is designed to meet the fundamental properties of a distance metric, as outlined in the literature:

1. Positivity: $distance(x, y) \geq 0$ for all stays $x$ and $y$.
2. Symmetry: $distance(x, y) = distance(y, x)$, ensuring the distance is symmetric.

In addition to these basic properties, TPM should also satisfy two critical properties to be considered a metric:

3. Triangle Property: $distance(x, y) \leq distance(x, z) + distance(z, y)$, which ensures that the distance between three points doesn't violate the triangle inequality.
4. Reflexivity: $distance(x, y) = 0$ if and only if $x = y$, signifying that the distance from a stay to itself is always zero.

TPM calculates the distance between two stays, considering both their context and content. Given two stays, denoted as $S_a$ and $S_b$, the TPM is computed using the following formula:

$$\begin{aligned} TPM_{\alpha,\beta}(S_a, S_b) = \alpha * distance_{context}(S_a, S_b) \\ +\beta * distance_{content}(S_a, S_b) \end{aligned} \tag{2}$$

The two hyper-parameters, $\alpha$ and $\beta$, are real numbers in the range of [0, 1], and they must satisfy the constraint $\alpha + \beta = 1$. These parameters are used to control the influence of the context and content components in the distance calculation. While it's evident that both context and content are essential for describing a stay, no specific proportions are assumed and leave that to be determined empirically. The context distance and the content distance are defined as follows:

$$\begin{aligned} distance_{context}(S_a, S_b) = distance_{duration}(S_a, S_b) \\ +distance_{season}(S_a, S_b) \end{aligned} \tag{3}$$

$$distance_{content}(S_a, S_b) = distance_{classification}(S_a, S_b) \tag{4}$$

where the season, the duration and the classification vector are defined by their respective distance.

### 4.1.1 Distance context

Let us now detail the context distance presented in Equation (3). This distance is composed of the duration distance and season distance.

***Duration distance.***

The duration distance is computed using two methods: a distribution distance and the Euclidean distance.

For the distribution distance, given a trip $S_a$ with a duration $a$ and a trip $S_b$ with a duration $b$, their distance duration is defined as :

$$distance_{duration}(S_a, S_b) = |p(a) - p(b)| \tag{5}$$

where $p(.)$ is the normal cumulative distribution function of each trip's duration.

The Euclidean distance difference have also been implemented, followed by a normalization of the result:

$$distance_{duration}(S_a, S_b) = \sqrt{(a-b)^2} \tag{6}$$

***Season distance***

The season distance is based on the calendar and represents the difference between the seasons of two stays. A graph to represent the cyclic nature of seasons is used, where each season is a node, and nodes are connected if the corresponding seasons are consecutive.

Given two stays, $S_a$ and $S_b$, with their respective seasons, and a graph $G$, the season distance between these stays is calculated as follows:

$$distance_{season}(S_a, S_b) = \begin{cases} 0 \text{ if } S_a \text{ and } S_b \text{ are the same node} \\ 0.5 \text{ if } S_a \text{ and } S_b \text{ are adjacent nodes} \\ 1 \text{ if } S_a \text{ and } S_b \text{ are distant nodes} \end{cases} \tag{7}$$

Adjacent nodes are determined using the adjacency matrix of the graph $G$. This approach effectively captures the seasonality difference between stays.

## 4.2 Content distance

The content distance, as detailed in Equation (4), is based on the classification vector that represents the proportion of places' categories visited by a tourist during their stay. Three types of distances are computed: Euclidean distance, cosine distance, and Manhattan distance.

To compute the Euclidean distance between two classification vectors $V_a$ and $V_b$ of stays $S_a$ and $S_b$, the squared differences of each sub-vector is summed and then take the square root of the result:

$$distance_{classification}(S_a, S_b) = \sum_{i=0}^{6} \sqrt{(V_{a_i} - V_{b_i})^2} \tag{8}$$

This distance is then normalized to ensure it does not bias the TPM computation.

The cosine distance compares the distribution of two vectors, not their magnitude, which is suitable for comparing behavior patterns. It calculates the cosine similarity between two sub-vectors of the classification vectors:

$$distance_{content}(S_a, S_b) = \sum_{i=0}^{n} 1 - cosine(Vec_{a_i}, Vec_{b_i}) \tag{9}$$

Where $Vec_{a_i}$ is the $i$-th sub-vector of stay $S_a$, and cosine similarity is defined as:

$$cosine(Vec_a, Vec_b) = \frac{Vec_a . Vec_b}{\|Vec_a\| \|Vec_b\|} \tag{10}$$

The Manhattan distance, also known as the Absolute-value norm, is simply the sum of the absolute differences between each element of the two classification vectors.

## 4.3 Clustering

The clustering process involves several steps. Firstly, the distance matrix $A$ is computed based on the stays in the dataset $S$. $A$ is a square matrix of size $n \times n$, where $n$ is the number of stays. Each entry $a_{ij}$ in $A$ represents the value of the TPM between the $i$-th and the $j$-th stay. $A$ must satisfy several properties to be considered a valid metric distance matrix, as outlined by Hakimi [50]:

- $a_{ij} = a_{ji}, \forall i, j = 1, 2, ..., n$ (symmetry)
- $a_{ii} = 0, \forall i = 1, 2, ..., n$ (hollow matrix)
- $a_{ij} > 0,\ if\ i \neq j\ \forall i, j = 1, 2, ..., n$ (positivity for off-diagonal entries)
- $a_{ik} \leq a_{ij} + a_{jk}, \forall i, j, k = 1, 2, ..., n$ (triangle inequality)

Once the distance matrix $A$ is computed, it serves as the input to a clustering algorithm, which groups the stays into clusters based on their similarity. Finally, the resulting clusters are analyzed to assess the quality of the clustering.

### 4.3.1 Clustering models

To detect communities among the trips, a clustering algorithm is required. There are various clustering algorithms available, including partition-based methods like K-Means, hierarchy-based methods like AGNES, density-based methods like DBSCAN, and model-based methods like Gaussian Mixture Models (GMM) as outlined in a comprehensive review by Xu et al. [23]. However, the choice of the clustering algorithm is not the central focus of this paper, and it will be explored in more detail in a future study to optimize cluster detection.

In this paper, the AGNES algorithm (Agglomerative Nesting Hierarchical Clustering) is used due to its consistency, but other algorithms could have been chosen as well. AGNES is based on the successive merging of clusters using a linkage method [51]. Initially, each trip in the distance matrix is considered a separate cluster. At each step, clusters are merged based on proximity criteria, with proximity being determined by the linkage method. Several linkage methods are available for hierarchical clustering, including average, single, complete, ward, weighted, centroid, and median [52]. The ward method is chosen to determine cluster proximity and facilitate the merging of clusters in the hierarchical clustering process. The ward method minimizes the total within-cluster variance, making it suitable for discovering compact clusters. It's important to note that each linkage method has its strengths and weaknesses, and a comprehensive method comparison will be conducted in a future paper.

To use the ward method, it's crucial to ensure that the distance used to compute the distance matrix is indeed Euclidean. This is because the ward method minimizes clusters formed based on Euclidean distance. The K-Means algorithm, for instance, is recommended for Euclidean distance applications, while K-Medoids is recommended for non-Euclidean distance applications. The distance matrix is positive-definite, allowing us to use the ward method.

Using an unsupervised algorithm enables us to determine clusters that contain trips with similar content and context. It's worth noting that trips are clustered independently of their original tourist, meaning trips made by the same tourist can be in the same or different clusters.

The unsupervised hierarchical clustering algorithm takes two inputs: the distance matrix and the desired number of clusters. Five indices are used (Dunn, Silhouette, Frey, McClain, C-Index) to optimize the number of clusters based on inter-cluster distance and intra-cluster density.

Internal measures for evaluating cluster partitions are important to assess the compactness, connectedness, and separation of the clusters. Several internal measures have been developed, such as the Silhouette index, Dunn index, Davies-Bouldin index, and more. When dealing with higher dimensions, the silhouette score is quite useful to validate the working of clustering algorithm.

The Silhouette index is a useful measure as it only requires a distance matrix to function, unlike many other internal measures that need the dataset itself. Since the presented method, based on TPM, generates a distance matrix, the choice of internal measures is somewhat limited by this factor. The Silhouette index needs to be optimized to determine the optimal number of clusters, denoted as k, and to evaluate a clustering method. This approach allows us to compare different clustering methods effectively.

## 4.4 Tourists Segmentation

Once the clusters have been determined by the algorithm, a summary for each of them is computed. To enhance these summaries, static information of every tourist are used, whose trips are present in a given cluster, including their demographic information. Additionally, both content and context details of every trip within the cluster are included. This comprehensive approach allows us to create a summary for each cluster, akin to the original tourist profile. These cluster summaries contain valuable information, including:

- statistics about trip duration (mean, median and standard variation);
- statistics about cluster construction (mean of digital traces per trip, cluster size);
- season distribution per cluster;
- nationality, age, gender distribution per cluster;
- location classification distribution per cluster.

These cluster summaries serve as the culmination of this method, synthesizing all the information contained in tourist stays with similar content and context. They are analyzed to extract valuable insights about tourists' behaviors, helping us create standardized tourist profiles.

In addition to the individual cluster summaries, a summary of the entire dataset is also computed. These summaries provide a comprehensive understanding of tourists' behaviors and can be compared with existing knowledge to challenge existing boundaries in the field.

# 5 Experiments and Results

## 5.1 TripAdvisor Dataset

The proposed approach is applied to the city of Paris, the capital of France. Choosing a relevant case study is crucial to validate the relevance of the proposed approach, and Paris, as one of the most attractive cities in the world, offers an ideal setting for this purpose. Paris consistently ranks among the topmost visited cities globally. Another reason for selecting Paris is the potential diversity in tourist profiles. This diversity stems from the city's multitude of attractions and visitors from various backgrounds.

The data for this study are derived from digital traces left by tourists during their trips, gathered from the booking site TripAdvisor. The data covers the period from 2015 to 2018. Each tourist is associated with a user who left a comment at a Parisian location, and the user's trip is represented by the sequence of comments left during their visit to Paris. The database contains $4,222,838$ comments distributed among $1,571,362$ tourists, resulting in an average of about 2.7 comments per tourist.

The TripAdvisor comments provide essential information including the user's ID (iduser), the date the comment was posted (dateCreation), the name of the location related to the comment (name), and its geographical coordinates (longitude & latitude). Some data like nationality, sex, and age are not revealed. To make a consistent database, we choose to drop users without a nationality (68% of loss), since sex and age are not relevant in the case study, we keep them if mentioned (22% missing value).

"Considering the prevalence of fake data on the internet, we treat nationality as a reliable attribute, given that TripAdvisor's tourism market analysis aligns closely with poll-based tourism market analysis. The distribution of nationalities is illustrated in Figure 5 for the raw database, Figure 6 per unique user, and Figure 7 for the stay database. A notable disparity exists between the two databases, with the distribution of all reviews and unique users being relatively similar. For instance, French individuals tend to contribute to trips with a substantial number of reviews, which, in the stay database, results in a lower overall percentage of French contributions. Additionally, data cleaning processes, including the application of a reviews threshold, consideration of only main locations, and ensuring nationality consistency, lead to the removal of 62% of the reviews."

Concerning age and sex, the statistics remain close (less than 2 percent of difference) between the raw database and the stay database:

- Sex: 54% men, 46% women.
- Age: 5%18-24yo, 22% 25-34yo, 42% 35-49yo, 30% 50-64yo.

In the following experiments, digital traces posted on the well-known social network TripAdvisor during the year 2018 are used. After basic data cleaning operations, 337,325 reviews are retained in the initial dataset.

From the initial dataset of digital traces, tourists' dataset and the stays' dataset are extracted. The stays' dataset comprises 4,427 stays. For the distance calculation, Euclidean distance provides the best results for both the Duration distance and the Content distance. However, before obtaining a distance matrix, the optimal values for the hyperparameters $\alpha$ and $\beta$ are tested.

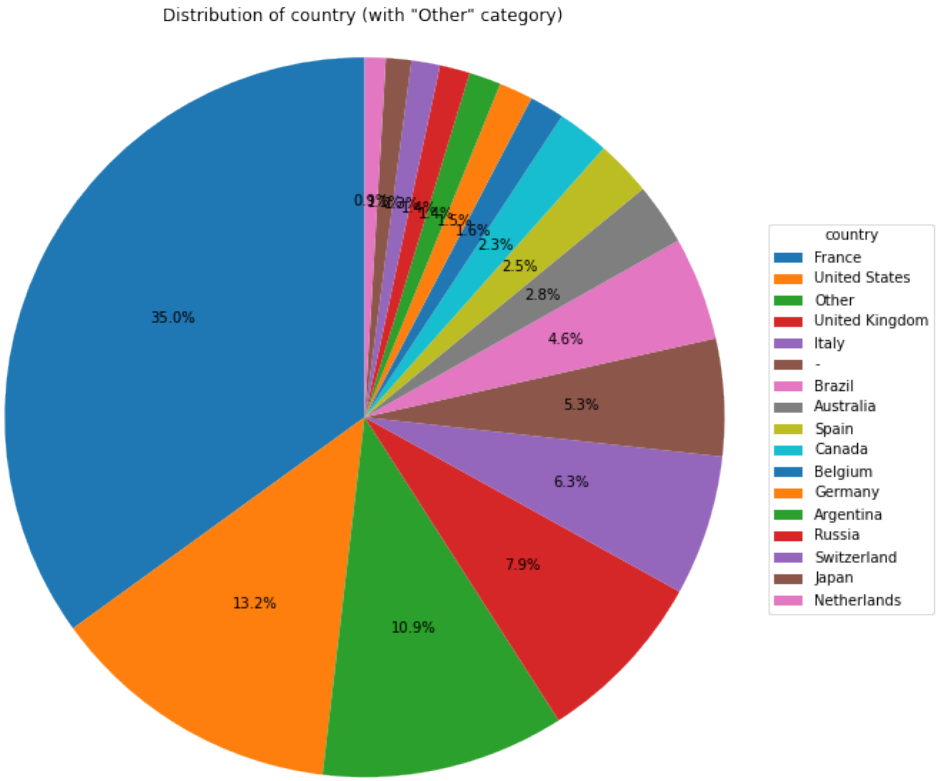



**Fig. 5** Countries distribution in the review database. "-" corresponds to missing value.

Let's consider a set of couples, where $\alpha$ and $\beta$ both range from 0 to 1 with a step of 0.1. This step size provides a good balance between exploring variations and computational efficiency. Instead of computing for all possible combinations of $\alpha$ and $\beta$, only the couples that yield sufficiently distinct results are selected. To identify these couples, the entanglement factor between the dendrograms produced for each pair of hyperparameters is computed. The entanglement factor ranges from 0 (no entanglement) to 1 (full entanglement) and indicates the quality of alignment between two dendrograms. Lower entanglement coefficients indicate better alignment. In Figure 8, one can observe three distinct plateaus.

The couples from the three plateaus as well as the two extreme values are retained. These couples are: $(1.0, 0.0)$, $(0.8, 0.2)$, $(0.5, 0.5)$, $(0.2, 0.8)$, and $(0.0, 1.0)$.

## 5.2 Results Comparison

To justify the relevance of the proposed approach, a comparison is made with a naive approach, as illustrated in Figure 9. The naive approach involves using the initial dataset composed of digital traces, combined with various clustering algorithms

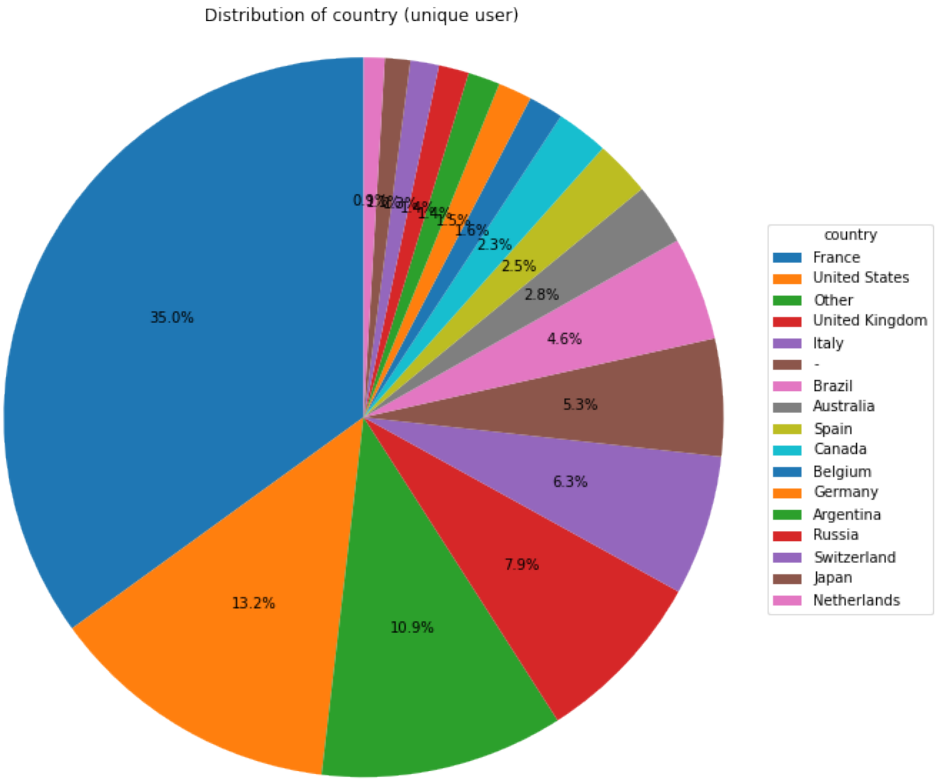



**Fig. 6** Countries distribution in the review database (unique user). "-" corresponds to missing value.

(K-Means, AGNES, and Spectral). However, it's important to note that these two approaches differ significantly. The naive approach relies on the initial dataset of digital traces, while our approach is based on the stays' dataset. This fundamental difference makes it challenging to directly compare the results, as shown in Figure 10 with a UMAP projection.

The comparison of approaches relies on three key indicators: the maximum Silhouette index value, the number of clusters determined from this value, and the number of validated segments. While the maximum Silhouette index value and the number of clusters have been discussed, let's delve into the concept of validated segments.

The goal of tourism segmentation is to uncover meaningful segments within the tourist data. It's essential to assess which method performs better in identifying these segments. To do this, the segments discovered by the proposed approach are compared to segments identified in tourism management studies. This process involves manual matching and validation by a tourism expert, and it's based on existing resources from

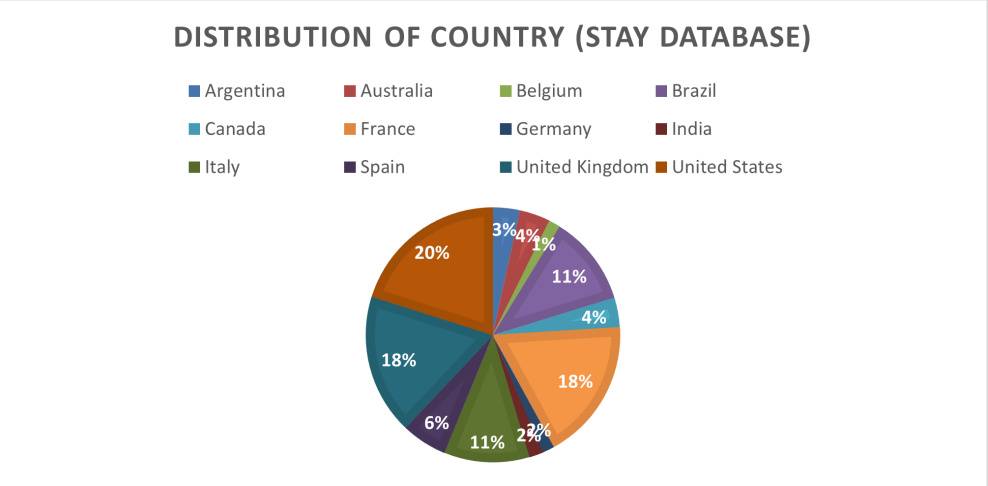



**Fig. 7** Countries distribution after the creation of the stay database.

regional tourism committees[2]. Please note that automating this validation process is a potential avenue for future work.

The number of validated segments is a crucial measure for evaluating a tourism segmentation method, and its significance surpasses that of the other two indicators. It helps determine how well the identified segments align with real-world tourism management insights. It is important to note that a cluster may validate one or more segments from the regional tourism committees.

**Table 3** Models comparison.

| Methods | Max Silhouette value | N° of clusters | N° of validated segments |
|---|---|---|---|
| $TPM_{1.0,0.0}$ | 0.319 | 9 | 12 |
| $TPM_{0.8,0.2}$ | 0.220 | 15 | 14 |
| $TPM_{0.6,0.4}$ | 0.419 | 4 | 4 |
| $TPM_{0.4,0.6}$ | 0.612 | 4 | 4 |
| $TPM_{0.2,0.8}$ | 0.762 | 4 | 0 |
| $TPM_{0.0,1.0}$ | 0.893 | 4 | 0 |
| K-Means | 0.180 | 8 | 5 |
| AGNES | 0.180 | 8 | 4 |
| Spectral | 0.180 | 8 | 5 |

The findings presented in Table 3 are indeed insightful. The decrease in the Silhouette index value as the hyperparameter $\alpha$ (related to the content of the stay) decreases suggests that the content of a stay has a more significant impact on the compactness of clusters than its context. This observation aligns with the understanding that the content of a stay, which is related to the types of locations visited, plays a crucial role in shaping tourists' behavior.

[2]https://pro.visitparisregion.com/chiffres-du-tourisme/profil-clientele-tourisme

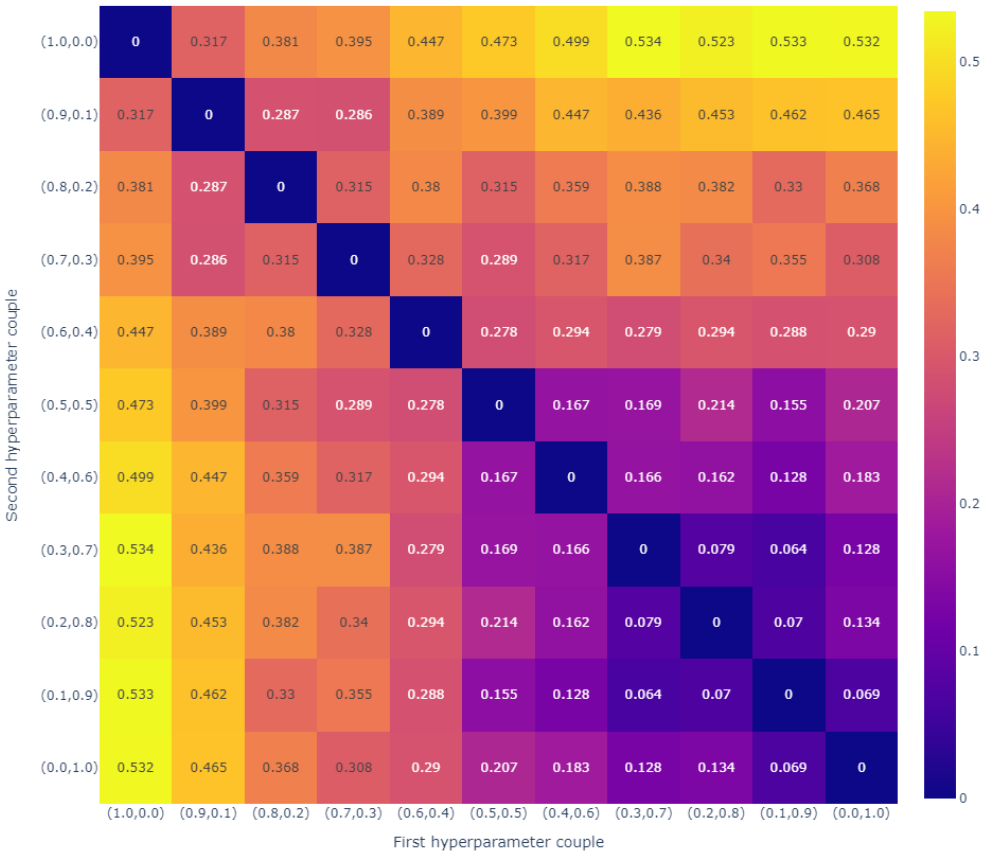



**Fig. 8** Entanglement heatmap.

It's worth noting that while the naive approach (benchmark algorithms) may have better Silhouette index values, the number of validated segments indicates that, under specific circumstances, the proposed approach outperforms the naive approach. This is a valuable finding, as it suggests that it can reveal segments that are more aligned with tourism management insights, even if the Silhouette index is not the highest.

The negative correlation between the Silhouette index value and the number of validated segments is an interesting observation. It implies that achieving high compactness in clusters may lead to a reduction in the diversity of segments, which might not always align with real-world tourism characteristics. This correlation indeed deserves further exploration in future studies.

From Table 3, it appears that the optimal hyperparameter couple is $(0.8, 0.2)$, which balances the influence of content and context. Additionally, it's important to highlight that these hyperparameters should be determined for each specific case study to optimize results.

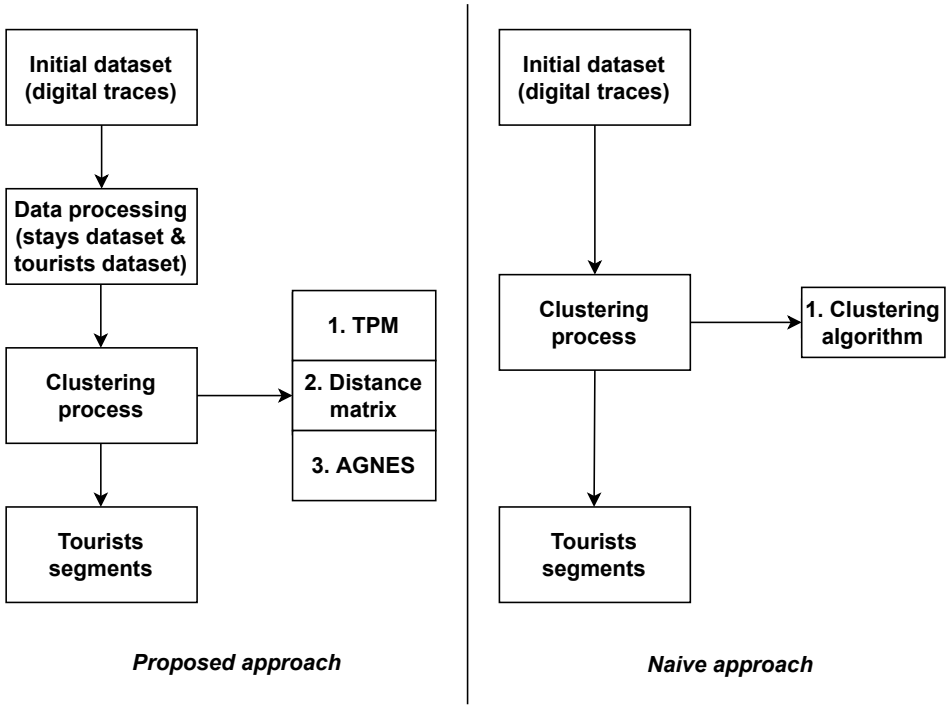



**Fig. 9** Approaches comparison: proposed approach (left) against naive approach (right).

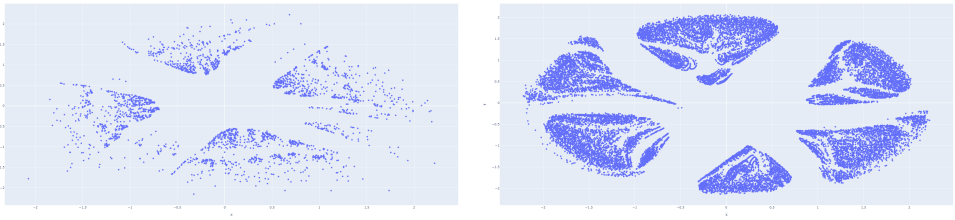


**Fig. 10** Dataset projection comparison: stays' dataset (left) against initial dataset (right).

## 5.3 Segments Validation

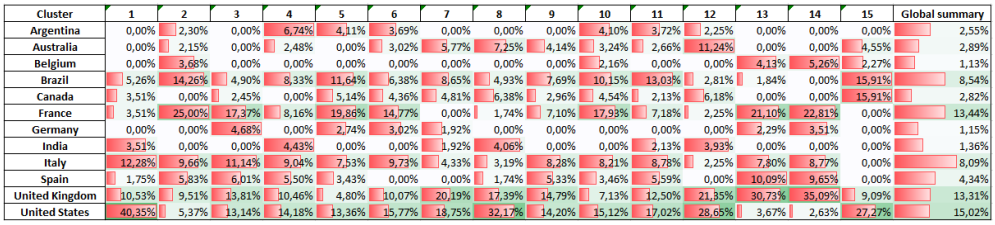
By summarizing clusters, the standard profile associated with each of them can be depicted. As detailed results for every model are not feasible, a method is sought to determine the model's performance in finding tourism segmentation. The count of segments determined by tourism management studies found by models in their respective clusters is conducted. This counting is based on tourism management studies found in the resources of the regional tourism committee data of Paris. For the explanation,

**Table 4** Statistics about clusters obtained by the $TPM_{0.8,0.2}$ setting

| Cluster | Average duration + std | N° of stays + % | Reviews per stay means + Std |
|---|---|---|---|
| Dataset | 0.93 + 2 | 4′427 + 100 | 5.453 + 2.423 |
| 1 | 2.088 + 3.1 | 57 + 1.3 | 16.386 + 5.4 |
| 2 | 0.707 + 1.7 | 652 + 14.7 | 4.04 + 2 |
| 3 | 0.606 + 1.4 | 449 + 10.1 | 3.503 + 1.6 |
| 4 | 1.321 + 2.7 | 564 + 12.7 | 8.149 + 3.9 |
| 5 | 1.301 + 2.6 | 292 + 6.6 | 6.562 + 3.6 |
| 6 | 0.674 + 1.6 | 298 + 6.7 | 4.171 + 1.9 |
| 7 | 0.548 + 1.3 | 208 + 4.7 | 6.202 + 2.6 |
| 8 | 1.107 + 2.3 | 345 + 7.8 | 6.388 + 2.7 |
| 9 | 1.503 + 2.9 | 169 + 3.8 | 6.29 + 3.6 |
| 10 | 0.801 + 1.9 | 463 + 10.5 | 4.019 + 1.6 |
| 11 | 0.779 + 1.9 | 376 + 8.5 | 5.67 + 2.5 |
| 12 | 0.736 + 2.3 | 178 + 4 | 5.893 + 2.3 |
| 13 | 0.78 + 1.8 | 218 + 4.9 | 2.596 + 0.9 |
| 14 | 0.868 + 1.8 | 114 + 2.6 | 4.64 + 1.2 |
| 15 | 2.841 + 2.7 | 44 + 1 | 12.545 + 3 |

the hyperparameters $(0.8, 0.2)$ are considered. All presented segments have been validated thanks to tourism management studies[3]; no new segments will be introduced here. For a more in-depth analysis of the results, please refer to the GitHub[4]. All the founded segments have also been validated by a tourist expert which want to remains anonymous.

For instance, as seen in Table 3, the setting has 14 validated segments. The Table 4 presents the statistics about clusters for the $TPM_{0.8,0.2}$ setting. Please note that the average duration do not only into account the date of reviews, not the real length of stay. Similarly, the information about the tourists demographics and places visited per cluster is available in clusters summaries (see Figure 11, 12, 13, 14, 15, 16). Since age and sex are not consistent information, i.e. TripAdvisor cannot guarantee of true or not, those data are not used to determine segments.



| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | Global summary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Argentina | 0,00% | 2,30% | 0,00% | 6,74% | 4,11% | 3,69% | 0,00% | 0,00% | 0,00% | 4,10% | 3,72% | 2,25% | 0,00% | 0,00% | 0,00% | 2,55% |
| Australia | 0,00% | 2,15% | 0,00% | 2,48% | 0,00% | 3,02% | 5,77% | 7,25% | 4,14% | 3,24% | 2,66% | 11,24% | 0,00% | 0,00% | 4,55% | 2,89% |
| Belgium | 0,00% | 3,68% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 2,16% | 0,00% | 0,00% | 4,13% | 5,26% | 2,27% | 1,13% |
| Brazil | 5,26% | 14,26% | 4,90% | 8,33% | 11,64% | 6,38% | 8,65% | 4,93% | 7,69% | 10,15% | 13,03% | 2,81% | 1,84% | 0,00% | 15,91% | 8,54% |
| Canada | 3,51% | 0,00% | 2,45% | 0,00% | 5,14% | 4,36% | 4,81% | 6,38% | 2,96% | 4,54% | 2,13% | 6,18% | 0,00% | 0,00% | 15,91% | 2,82% |
| France | 3,51% | 25,00% | 17,37% | 8,16% | 19,86% | 14,77% | 0,00% | 1,74% | 7,10% | 17,93% | 7,18% | 2,25% | 21,10% | 22,81% | 0,00% | 13,44% |
| Germany | 0,00% | 0,00% | 4,68% | 0,00% | 2,74% | 3,02% | 1,92% | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 2,29% | 3,51% | 0,00% | 1,15% |
| India | 3,51% | 0,00% | 0,00% | 4,43% | 0,00% | 0,00% | 1,92% | 4,06% | 0,00% | 0,00% | 2,13% | 3,93% | 0,00% | 0,00% | 0,00% | 1,36% |
| Italy | 12,28% | 9,66% | 11,14% | 9,04% | 7,53% | 9,73% | 4,33% | 3,19% | 8,28% | 8,21% | 8,78% | 2,25% | 7,80% | 8,77% | 0,00% | 8,09% |
| Spain | 1,75% | 5,83% | 6,01% | 5,50% | 3,43% | 0,00% | 0,00% | 1,74% | 5,33% | 3,46% | 5,59% | 0,00% | 10,09% | 9,65% | 0,00% | 4,34% |
| United Kingdom | 10,53% | 9,51% | 13,81% | 10,46% | 4,80% | 10,07% | 20,19% | 17,39% | 14,79% | 7,13% | 12,50% | 21,35% | 30,73% | 35,09% | 9,09% | 13,31% |
| United States | 40,35% | 5,37% | 13,14% | 14,18% | 13,36% | 15,77% | 18,75% | 32,17% | 14,20% | 15,12% | 17,02% | 28,65% | 3,67% | 2,63% | 27,27% | 15,02% |

**Fig. 11** Nationalities distribution per cluster for the $TPM_{0.8,0.2}$ setting. In red bar, the relative percent for a given nationality; in shade of green, the relative percent for a given cluster.

Let's analyze the clusters to determine segments.

[3] https://pro.visitparisregion.com/chiffres-du-tourisme/profil-clientele-tourisme
[4] https://github.com/SmartGridandCity/TourismProfileMeasure

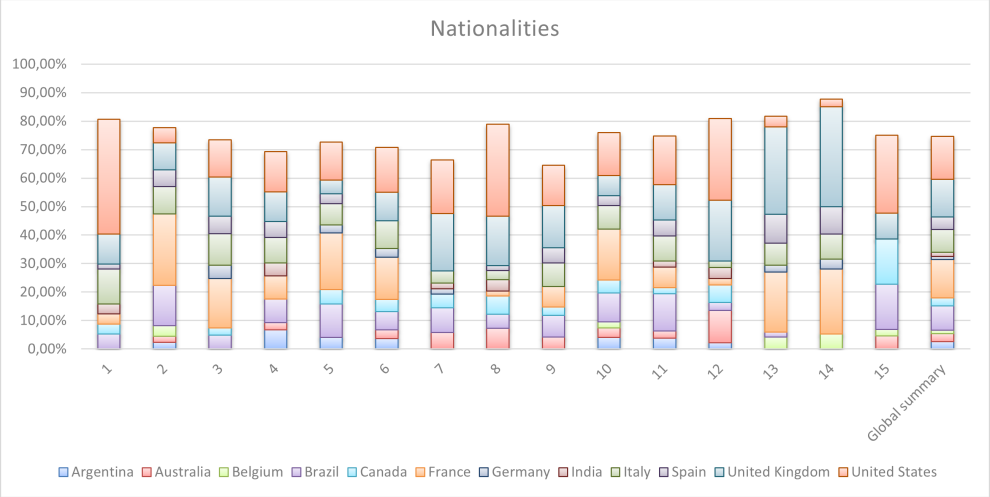



**Fig. 12** Nationalities distribution per cluster for the $TPM_{0.8,0.2}$ setting.

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Monument | 10,19% | 6,83% | 0,13% | 10,09% | 4,21% | 2,62% | 19,72% | 22,64% | 9,20% | 5,22% | 16,88% | 22,53% | 0,34% | 0,18% | 8,77% | 8,93% |
| Park/garden | 4,46% | 4,38% | 7,29% | 3,82% | 5,20% | 5,79% | 1,84% | 2,89% | 3,07% | 6,76% | 4,94% | 2,73% | 1,03% | 0,18% | 4,11% | 4,42% |
| Urbanism | 16,24% | 13,47% | 12,51% | 27,46% | 12,12% | 14,67% | 7,37% | 7,35% | 15,99% | 17,31% | 12,96% | 6,79% | 3,43% | 0,92% | 8,77% | 13,74% |
| Art gallerie/museum | 8,60% | 10,19% | 8,36% | 5,80% | 7,33% | 13,48% | 10,13% | 11,14% | 5,48% | 10,54% | 7,41% | 13,01% | 0,51% | 0,18% | 14,31% | 8,57% |
| Holy site/place of worship | 13,06% | 17,69% | 17,35% | 13,11% | 14,56% | 22,36% | 14,28% | 12,22% | 14,68% | 21,94% | 18,60% | 10,18% | 1,03% | 0,55% | 20,22% | 15,40% |
| Historic building | 26,01% | 26,71% | 24,89% | 19,62% | 35,15% | 32,83% | 25,48% | 24,18% | 21,28% | 30,67% | 26,15% | 24,78% | 1,71% | 0,74% | 33,81% | 24,58% |
| Theater/auditorium | 5,95% | 0,04% | 0,06% | 1,02% | 15,19% | 0,08% | 0,08% | 0,05% | 1,12% | 0,05% | 0,05% | 0,09% | 0,17% | 0,18% | 0,18% | 1,30% |
| Shop | 0,21% | 0,04% | 0,06% | 0,04% | 0,10% | 0,08% | 0,38% | 0,09% | 0,19% | 0,05% | 0,14% | 0,09% | 0,17% | 20,96% | 0,18% | 0,64% |
| Restaurant/bar | 1,06% | 4,34% | 5,66% | 0,41% | 1,35% | 3,09% | 0,23% | 0,23% | 1,12% | 2,08% | 0,70% | 0,28% | 9,42% | 6,99% | 0,36% | 2,58% |
| Gastronomy | 0,74% | 1,28% | 1,13% | 0,26% | 0,62% | 0,87% | 0,15% | 0,05% | 0,84% | 0,80% | 0,28% | 0,09% | 0,51% | 0,55% | 0,36% | 0,64% |
| Hotel | 0,64% | 4,23% | 7,67% | 0,61% | 0,68% | 1,51% | 1,54% | 0,90% | 1,39% | 1,33% | 0,98% | 0,94% | 21,40% | 10,11% | 0,54% | 3,41% |
| Wood | 0,11% | 0,04% | 0,06% | 0,02% | 0,05% | 0,08% | 0,08% | 0,05% | 0,09% | 0,05% | 0,05% | 0,09% | 0,17% | 0,18% | 0,18% | 0,06% |
| Watering place | 4,03% | 0,04% | 0,06% | 12,24% | 0,52% | 0,08% | 0,08% | 0,05% | 2,97% | 0,05% | 0,05% | 0,09% | 0,17% | 0,18% | 0,72% | 1,82% |
| Beach | 0,11% | 0,04% | 0,06% | 0,02% | 0,05% | 0,08% | 0,08% | 0,05% | 0,09% | 0,05% | 0,05% | 0,09% | 0,17% | 0,18% | 0,18% | 0,06% |
| Mountain | 0,11% | 0,04% | 0,06% | 0,02% | 0,05% | 0,08% | 0,08% | 0,05% | 0,09% | 0,05% | 0,05% | 0,09% | 0,17% | 0,18% | 0,18% | 0,06% |
| Music building | 1,49% | 0,04% | 0,06% | 0,07% | 0,73% | 0,08% | 0,08% | 0,77% | 15,71% | 0,05% | 0,09% | 0,09% | 0,17% | 0,18% | 0,54% | 0,79% |
| Cinema | 0,11% | 0,04% | 0,06% | 0,02% | 0,05% | 0,08% | 0,08% | 0,05% | 0,09% | 0,05% | 0,05% | 0,09% | 0,17% | 0,18% | 0,18% | 0,06% |
| Amusement park/aquarium | 0,53% | 10,52% | 14,33% | 0,96% | 0,88% | 1,43% | 2,23% | 1,44% | 2,32% | 1,70% | 2,42% | 0,85% | 58,90% | 56,99% | 0,36% | 8,38% |
| Sport | 0,21% | 0,04% | 0,06% | 0,07% | 0,05% | 0,08% | 0,08% | 0,05% | 0,09% | 0,53% | 0,05% | 0,09% | 0,17% | 0,18% | 0,18% | 0,12% |
| Viewpoint | 6,16% | 0,04% | 0,13% | 4,34% | 1,09% | 0,63% | 16,04% | 15,79% | 4,18% | 0,69% | 3,11% | 16,97% | 0,17% | 0,18% | 5,90% | 4,42% |

**Fig. 13** Places classification distribution per cluster for the $TPM_{0.8,0.2}$ setting. In red bar, the relative percent for a given location type; in shade of green, the relative percent for a given cluster.

**Cultural Enthusiasts, History Enthusiasts (Cluster 6, 10 and 15)**: These clusters exhibit similar characteristics in terms of tourists' length of stay and the nationality of tourists, with a significant proportion of English and American tourists. The primary attractions for tourists in these clusters are museums, historic buildings, and monuments. These tourists typically visit Paris during autumn and summer, showing a preference for cultural experiences. These clusters represent the cultural appeal of the French capital. Not that those cluster show the highest score in cultural places, but in general most of clusters show high values in those locations.

**Amusement Park Aficionados, Local Tourism (Cluster 13 and 14)**: These clusters share common traits, such as a high presence of amusement parks, hotels, and restaurants. These segments are strongly associated with visits to attractions like

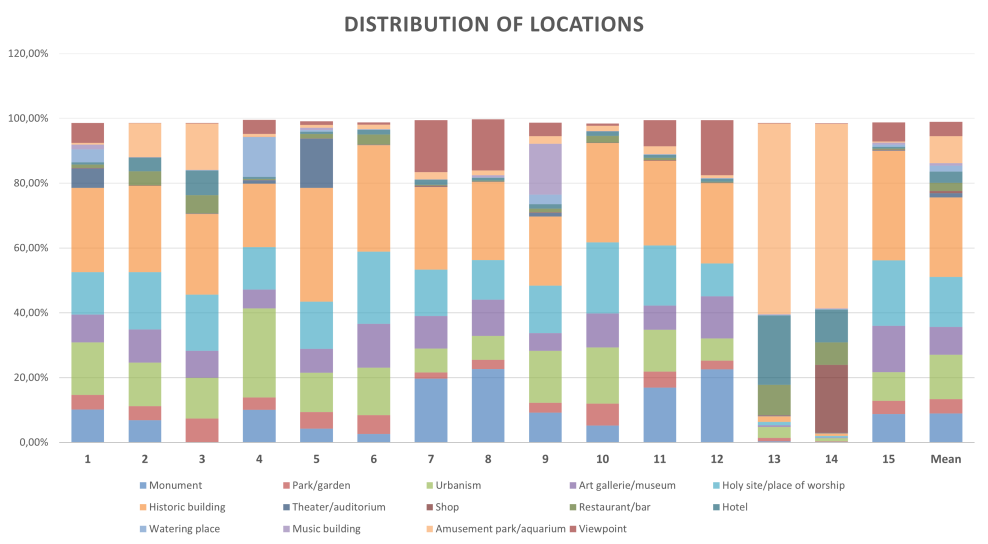
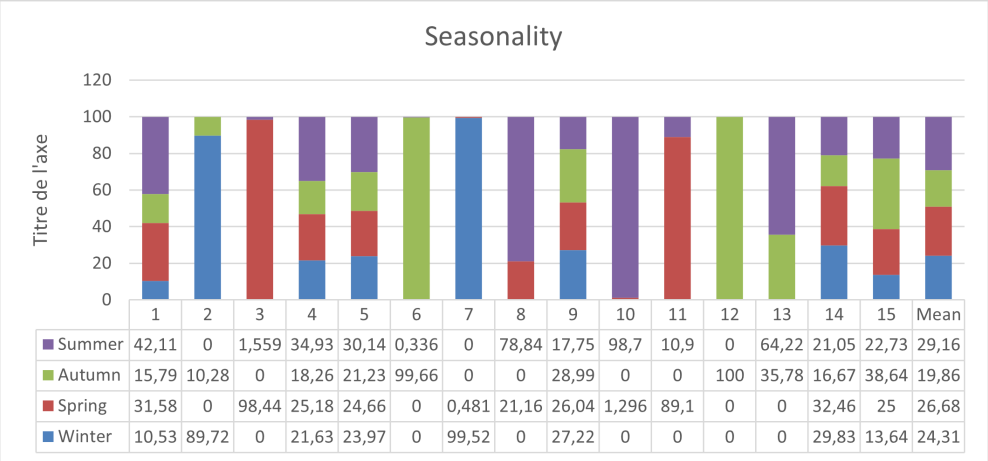



**Fig. 14** Places classification distribution per cluster for the $TPM_{0.8,0.2}$ setting.




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Summer | 42,11 | 0 | 1,559 | 34,93 | 30,14 | 0,336 | 0 | 78,84 | 17,75 | 98,7 | 10,9 | 0 | 64,22 | 21,05 | 22,73 | 29,16 |
| Autumn | 15,79 | 10,28 | 0 | 18,26 | 21,23 | 99,66 | 0 | 0 | 28,99 | 0 | 0 | 100 | 35,78 | 16,67 | 38,64 | 19,86 |
| Spring | 31,58 | 0 | 98,44 | 25,18 | 24,66 | 0 | 0,481 | 21,16 | 26,04 | 1,296 | 89,1 | 0 | 0 | 32,46 | 25 | 26,68 |
| Winter | 10,53 | 89,72 | 0 | 21,63 | 23,97 | 0 | 99,52 | 0 | 27,22 | 0 | 0 | 0 | 0 | 29,83 | 13,64 | 24,31 |

**Fig. 15** Season distribution per cluster for the $TPM_{0.8,0.2}$ setting.

Disneyland Paris and Parc Astérix, among others. The tourists in these clusters are predominantly from France and neighboring countries. An interesting observation is that cluster 11 corresponds to the winter season, suggesting that special events in these amusement parks during the winter draw a distinct tourist crowd.

**Shopping Tourists (Cluster 14)**: Enjoying exploring local markets, shopping districts, and unique boutiques, in Paris, Luxury Travelers and Local Tourists are also here to do some shopping.

| Cluster | man | woman | 50-64 | 35-49 | 25-34 | 18-24 |
|---|---|---|---|---|---|---|
| 1 | 68,421 | 31,579 | 33,333 | 47,368 | 15,789 | 3,509 |
| 2 | 51,534 | 48,466 | 26,994 | 42,791 | 23,313 | 6,442 |
| 3 | 49,22 | 50,78 | 31,626 | 37,194 | 24,722 | 5,791 |
| 4 | 56,56 | 43,44 | 32,624 | 40,78 | 20,745 | 5,496 |
| 5 | 52,397 | 47,603 | 33,562 | 40,411 | 22,945 | 2,397 |
| 6 | 53,02 | 46,98 | 41,611 | 32,55 | 21,812 | 4,027 |
| 7 | 52,885 | 47,115 | 27,885 | 40,385 | 23,077 | 8,173 |
| 8 | 52,754 | 47,246 | 25,507 | 46,957 | 21,449 | 5,797 |
| 9 | 52,663 | 47,337 | 29,586 | 43,787 | 23,077 | 3,55 |
| 10 | 56,803 | 43,197 | 36,069 | 40,821 | 19,006 | 4,104 |
| 11 | 57,181 | 42,819 | 24,734 | 43,617 | 25 | 6,383 |
| 12 | 49,438 | 50,562 | 34,27 | 38,202 | 24,157 | 3,371 |
| 13 | 57,339 | 42,661 | 24,312 | 49,541 | 22,018 | 3,67 |
| 14 | 58,772 | 41,228 | 18,421 | 49,123 | 25,439 | 5,263 |
| 15 | 47,727 | 52,273 | 36,364 | 38,636 | 18,182 | 6,818 |
| Mean | 53,896 | 46,104 | 30,495 | 41,563 | 22,408 | 5,173 |

**Fig. 16** Age and sex distribution per cluster for the $TPM_{0.8,0.2}$ setting. In shade of green, the relative percent for each demographic category.

**Leisure Travelers, Photography Enthusiasts (Cluster 7, 8, and 12)**: Tourists in these clusters travel for relaxation and enjoyment, often seeking recreational activities, sightseeing, and cultural experiences. They travel to Paris during any season. They are mostly foreigners and are focus on main attractions of Paris. Those tourists are also focused on capturing picturesque landscapes, architecture, and cultural moments.

**Luxury Travelers, Food and Culinary Travelers, Recreational Tourists (Cluster 2 and 3)**: Those tourists seek high-end experiences (hotel/spa, theaters), luxurious accommodations, and fine dining. In Paris, luxury travelers are also amusement park aficionados since they are very expensive. Luxury travelers are mostly local tourists or neighbourhood countries during the new season trends, i.e. when shops propose new items (spring). They also came during winter sales.

**Nightlife Seekers (Cluster 5 and 9)**: Those tourists look for theatre and music building where they can enjoy Cabaret ambiance like Moulin Rouge or Crazy Horse. Those tourists travel to Paris during the whole year from all the world.

**Nature and Parks Explorers, Wandering Tourists (Cluster 3, 5, 6, and 10)**: Enjoying the green spaces of Paris during all seasons excepted winter, including Luxembourg Gardens and Parc des Buttes-Chaumont. They include Wandering Tourists who want to discover all the beauty of the city. Wandering tourists also want to discover traditional dishes. The close relation between those segments is due to the proximity between traditional restaurants (more than 9,400) and parks (more than 420) in Paris. Parks Explorers who do not seek Culinary experiences during summer, when it's almost impossible to stroll in the city.

**Romantic Getaway (Clusters 4)**: Couples seeking the romantic atmosphere of Paris, enjoying the city; its main monuments, urbanism and the famous watering places. Marriage proposal in Paris are made during all the year. Local people and

neighbourhood countries are well represented in these clusters since it's an easy trip to do the proposal.

Clusters 1 and 11 are very close to the global summary, which can be designated as Basic Tourism.

As seen in the clustering results, tourists in Paris may have various profile, mostly independent of the nationality and the season. This kind of results is specific to Paris and cannot be extrapolated to another city of region. Each results must be analysed carefully and independently to any other results. For instance, results on thee Hauts-de-France region are mostly related to the nationality and age of tourists. The corresponding results are discribed in the Github.

# 6 Discussions

## 6.1 Filling the gaps

This discussion highlights several key points that merit further consideration and exploration. Let's consider the gaps cited in the Context:

- *Impact of Context and Content Hyper-Parameters*: The observation that the content of a stay is more correlated with demographic and geographic information than the context of the stay is intriguing. This suggests that focusing on the content of stays may be more effective in identifying tourist segments. However, it's essential to validate this assumption through more extensive studies to establish its generalizability beyond the specific case of Paris.
- *Need for a More Meaningful Indicator Measure*: The significance of the number of validated segments as an indicator measure is highlighted. Automating this measure would streamline its application and contribute to the efficiency of future research. Developing robust, automated validation techniques for tourism segments could be a valuable direction for future work.
- *The Intriguing Correlation Between the Silhouette Index and the Number of Clusters*: The close-to-zero correlation between the Silhouette index and the number of clusters is a thought-provoking discovery. This correlation suggests that focusing solely on compact clusters might not always lead to meaningful tourist segments. A deeper investigation into the implications of this correlation and its impact on the quality of segmentation is warranted.
- *Applications to the Tourism Industry*: The potential applications to the tourism industry are promising. Using tourism segmentation to improve tourist experiences and align with local expectations can be valuable for both tourists and local economies. Further research could explore practical implementations and assess the impact of such applications.

Overall, this study has raised important questions and provided valuable insights into the complexities of tourist segmentation. It paves the way for future research to build upon these findings and address the challenges and opportunities in this domain.

## 6.2 Hyperparametrisation

When considering the two hyperparameters, $\alpha$ and $\beta$, it's important to acknowledge that optimizing them is not a straightforward task. The reason lies in the fact that their values will determine different perspectives on the studied tourist population.

In the case of $(1.0, 0.0)$, where demographic values take precedence, it aligns with the conventional approach to tourism segmentation found in the field of tourism management. This approach aims to identify distinct population groups and understand their behaviors. Conversely, the case of $(0.0, 1.0)$ is closely tied to data-driven analysis, where the dynamic data of tourists are the primary focus for segmentation.

To explore the need for varying hyperparameters, employing an entanglement matrix can be highly beneficial. This matrix helps navigate the spectrum of possibilities, allowing experts to extract valuable insights from each cluster. It facilitates comparisons with overall data statistics and enables an examination of variations in statistics between clusters.

Thanks to the method and metric presented, we can effectively manage big data in tourism for tourism segmentation. This approach provides interpretable results and addresses significant challenges in dealing with big data and segmentation.

It has the potential to mitigate inconsistencies in segmentation variables and ensures that the variables used are more predictive of travel behavior and preferences. Depending on each segment, the results are compared within segments or the overall data to identify which characteristics are distinctive and deviate significantly from mainstream behaviors. This method can be instrumental in uncovering cross-cultural differences in travel behavior and preferences and in developing segmentation models tailored to specific cultural contexts.

## 6.3 Discovering of new segment trends

The proposed method can be harnessed to incorporate non-demographic variables, such as social media data, online reviews, and location-based data, into tourism segmentation models. This integration can yield a more nuanced comprehension of traveler preferences and behavior patterns. The flexibility provided by the hyperparameter allows for fine-tuning the impact of each parameter on the creation of segments.

Furthermore, it serves as a tool to detect emerging trends and technologies within the tourism market. It facilitates the development of segmentation models that adapt to these trends. For instance, it can identify travelers interested in sustainable tourism or digital nomadism, enabling the formulation of targeted marketing strategies for these specific segments.

The method supports the development of more sophisticated segmentation models capable of identifying subgroups of travelers with similar preferences and behavior patterns. For instance, the clustering algorithm can group travelers based on their travel motivations or personality traits by adjusting hyperparameters and employing classification instead of clustering.

The method has the potential to enhance the accuracy and predictive capabilities of tourism segmentation models, even in uncharted areas. It offers a more intricate comprehension of the diverse and evolving tourism market.

Here are some emerging trends:

- **Wellness and Health Tourism**: Health and wellness tourism is on the rise, with travelers prioritizing relaxation, fitness, and mental well-being during their trips. This segment seeks destinations offering spa treatments, yoga retreats, and health-focused activities. This kind of tourism can be found in the Alps, Normandy, and Provence regions. The hotels' category is overrepresented (where the spas belong), with few to zero other category..
- 
- **Adventure and Extreme Tourism**: Adventure seekers are looking for adrenaline-pumping experiences such as extreme sports, trekking, and adventure travel. Segments within this trend include thrill-seekers, hikers, and those interested in extreme sports. This kind of tourism can be found in the Auvergne, Rhône and Alpes regions. The woods and moutains category is overrepresented, with few to zero other category.
- 
- **Staycations**: The concept of staycations has gained popularity, especially during the COVID-19 pandemic. Travelers opt to explore their own regions or nearby destinations, often to support local businesses and minimize travel risks. By analyzing the yearly evolution of clusters, one can find how local people visit their lands.
- **Slow Travel**: Slow travel encourages travelers to take a more relaxed and immersive approach to exploring a destination. It involves spending more time in one place, getting to know the local culture, and minimizing rushed itineraries. By analyzing the yearly evolution of clusters, one can find how the leisure time size changes over time.

## 6.4 Evolution over time

By regularly updating the dataset and reapplying the segmentation process, it becomes possible to observe how tourism segments shift over time. This temporal analysis provides insights into trends, emerging patterns, and changes in traveler behavior. For instance, one can detect if certain segments are growing or declining, and whether new segments are emerging. This can be especially valuable for businesses and destinations aiming to adapt their strategies to evolving market dynamics.

Events, whether they are political, related to urban development, or global in scale, can significantly influence tourism behavior. The proposed method allows for the examination of how such events impact tourism segments. For example:

- **Political Changes**: Changes in government policies, regulations, or geopolitical events can lead to shifts in tourism segments. By analyzing data before and after such changes, one can reveal how political events alter traveler preferences, destinations, or travel patterns.
- **Urban Development**: Urban transformations, including the construction of new attractions, infrastructure improvements, or changes in city planning, can have a

profound effect on tourism. The approach can capture the evolving dynamics of tourism segments in response to these urban developments.

- **Global Events**: The COVID-19 pandemic disrupted the tourism industry on an unprecedented scale. The method can be used to study how the pandemic impacted different tourism segments. This can involve analyzing data from both pre- and post-pandemic periods to understand how traveler preferences and behavior have changed.

By conducting such analyses, businesses, policymakers, and tourism industry stakeholders can make informed decisions, adapt their strategies, and anticipate the needs and preferences of travelers in the wake of significant events. This not only provides a means to react effectively to changing circumstances but also to proactively shape the future of tourism in response to evolving trends and challenges.

### 6.5 Mixed Data Clustering challenges

Analyzing tourism data, especially when dealing with mixed data and big data, presents several challenges. These challenges often include managing diverse data types, extracting meaningful insights, and ensuring scalability. The presented metric, the TPM, is designed to tackle these challenges effectively.

Mixed data in tourism analytics refers to the coexistence of both categorical and numerical data, such as traveler demographics and location-based information. Traditional clustering algorithms struggle to handle mixed data efficiently because they often require data to be homogenized or transformed into a single data type, which can lead to information loss.

TPM overcomes this challenge by providing a metric that can accommodate mixed data seamlessly. It calculates the distance between two stays, considering both their context and content, which may consist of categorical and numerical attributes. This allows TPM to retain the richness of mixed data, ensuring that no information is sacrificed during the analysis.

In addition to addressing data challenges, TPM provides interpretable results. It calculates distances between stays based on context and content, which can be translated into meaningful insights about tourism behavior. This interpretability is essential for stakeholders in the tourism industry who aim to understand their audience and tailor their services accordingly.

## 7 Conclusion

Throughout this article, the needs for a tourism segmentation approach have been discussed. The TPM method is designed to efficiently segment tourists based on a data-driven approach, proposing an innovative way to understand the tourism industry. In summary, a data processing stage has been combined with a clustering process, introducing a new metric measure called TPM. The clustering algorithm to complement it is the well-known hierarchical clustering, AGNES. The validity of the approach is demonstrated through experiments calibrated with those of a naive approach. Furthermore, the approach successfully provides segments that can be matched with tourism

management reports. Thus, the method can be considered both reliable and effective. The code and all results obtained by the approach and the naive approach are available on GitHub[5].

The tourism industry can utilize the method to complement tourism management studies. Given that the method requires the analysis of digital traces, it negates the complicated and expensive framework inherent in tourism management studies. It is envisioned that tourism experts will enhance this study to better analyze the segments established by the method.

In addition to the suggestions made, efforts will be made to ameliorate the method in future studies. Notably, obtaining more information about tourist static and dynamic data, such as revenues of tourists, presence of other travelers during the stay, etc., would be a valuable addition. Lastly, the approach will be evaluated in another geographical area and during the COVID epidemic years to compare and challenge the obtained results.

# References


[1] Hu, F., Li, Z., Yang, C., Jiang, Y.: A graph-based approach to detecting tourist movement patterns using social media data. Cartography and Geographic Information Science **46**(4), 368–382 (2019)

[2] Zhang, K., Chen, D., Li, C.: How are tourists different?-reading geo-tagged photos through a deep learning model. Journal of Quality Assurance in Hospitality & Tourism **21**(2), 234–243 (2020)

[3] Alén, E., Losada, N., Domínguez, T.: The impact of ageing on the tourism industry: An approach to the senior tourist profile. Social Indicators Research **127**(1), 303–322 (2016)

[4] Rafael, C., Almeida, A.: Socio-demographic tourist profile and destination image in online environment. Journal of Advanced Management Science **5**(5) (2017)

[5] Chareyron, G., Da-Rugna, J., Raimbault, T.: Big data: A new challenge for tourism. In: 2014 IEEE International Conference on Big Data (Big Data), pp. 5–7 (2014). IEEE

[6] Amaro, S., Duarte, P., Henriques, C.: Travelers' use of social media: A clustering approach. Annals of Tourism Research **59**, 1–15 (2016)

[7] Godoy, D., Amandi, A.: User profiling in personal information agents: a survey. The Knowledge Engineering Review **20**(4), 329–361 (2005)

[8] Eke, C.I., Norman, A.A., Shuib, L., Nweke, H.F.: A survey of user profiling: State-of-the-art, challenges, and solutions. IEEE Access **7**, 144907–144924 (2019)


[5] https://github.com/SmartGridandCity/TourismProfileMeasure

[9] Araniti, G., De Meo, P., Iera, A., Ursino, D.: Adaptively controlling the qos of multimedia wireless applications through" user profiling" techniques. IEEE Journal on Selected Areas in Communications **21**(10), 1546–1556 (2003)

[10] Nilashi, M., Ibrahim, O., Ithnin, N., Sarmin, N.H.: A multi-criteria collaborative filtering recommender system for the tourism domain using expectation maximization (em) and pca–anfis. Electronic Commerce Research and Applications **14**(6), 542–562 (2015)

[11] Cufoglu, A.: User profiling-a short review. International Journal of Computer Applications **108**(3) (2014)

[12] Gavalas, D., Kenteris, M.: A web-based pervasive recommendation system for mobile tourist guides. Personal and Ubiquitous Computing **15**(7), 759–770 (2011)

[13] Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., Li, X.: Efficient user profiling based intelligent travel recommender system for individual and group of users. Mobile Networks and Applications **24**(3), 1018–1033 (2019)

[14] Refanidis, I., Emmanouilidis, C., Sakellariou, I., Alexiadis, A., Koutsiamanis, R.-A., Agnantis, K., Tasidou, A., Kokkoras, F., Efraimidis, P.S.: myvisitplanner gr: Personalized itinerary planning system for tourism. In: Hellenic Conference on Artificial Intelligence, pp. 615–629 (2014). Springer

[15] Abbasi-Moud, Z., Vahdat-Nejad, H., Sadri, J.: Tourism recommendation system based on semantic clustering and sentiment analysis. Expert Systems with Applications **167**, 114324 (2021)

[16] Cao, L., Luo, J., Gallagher, A., Jin, X., Han, J., Huang, T.S.: Aworldwide tourism recommendation system based on geotaggedweb photos. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2274–2277 (2010). IEEE

[17] Massimo, D., Ricci, F.: Clustering users' pois visit trajectories for next-poi recommendation. In: Information and Communication Technologies in Tourism 2019, pp. 3–14. Springer, ??? (2019)

[18] Rodríguez, J., Semanjski, I., Gautama, S., Weghe, N., Ochoa, D.: Unsupervised hierarchical clustering approach for tourism market segmentation based on crowdsourced mobile phone data. Sensors **18**(9), 2972 (2018)

[19] Wind, Y.J., Bell, D.R.: Market Segmentation. Routledge, ??? (2008)

[20] Calantone, R.J., Johar, J.S.: Seasonal segmentation of the tourism market using a benefit segmentation framework. Journal of Travel Research **23**(2), 14–24 (1984)

[21] Dolnicar, S.: Market segmentation for e-tourism. Handbook of e-Tourism, 1–15

(2020)

[22] Deseure-Charron, F., Djebali, S., Guérard, G.: Clustering method for touristic photographic spots recommendation. In: Advanced Data Mining and Applications: 18th International Conference, ADMA 2022, Brisbane, QLD, Australia, November 28–30, 2022, Proceedings, Part II, pp. 223–237 (2022). Springer

[23] Xu, D., Tian, Y.: A comprehensive survey of clustering algorithms. Annals of Data Science **2**(2), 165–193 (2015)

[24] Backer, E., Jain, A.K.: A clustering performance measure based on fuzzy set decomposition. IEEE Transactions on Pattern Analysis and Machine Intelligence (1), 66–75 (1981)

[25] Blanco-Moreno, S., González-Fernández, A.M., Muñoz-Gallego, P.A.: Big data in tourism marketing: past research and future opportunities. Spanish Journal of Marketing-ESIC (ahead-of-print) (2023)

[26] Dolnicar, S.: A review of data-driven market segmentation in tourism. Journal of Travel & Tourism Marketing **12**(1), 1–22 (2002)

[27] D'Urso, P., De Giovanni, L., Disegna, M., Massari, R., Vitale, V.: A tourist segmentation based on motivation, satisfaction and prior knowledge with a socio-economic profiling: A clustering approach with mixed information. Social Indicators Research **154**(1), 335–360 (2021)

[28] Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A.: User profiles for personalized information access. The adaptive web, 54–89 (2007)

[29] Wassler, P., Nguyen, T.H.H., Schuckert, M., *et al.*: Social representations and resident attitudes: A multiple-mixed-method approach. Annals of Tourism Research **78**, 102740 (2019)

[30] Chung, M.G., Herzberger, A., Frank, K.A., Liu, J.: International tourism dynamics in a globalized world: A social network analysis approach. Journal of Travel Research **59**(3), 387–403 (2020)

[31] Leal, F., González–Vélez, H., Malheiro, B., Burguillo, J.C.: Semantic profiling and destination recommendation based on crowd-sourced tourist reviews. In: International Symposium on Distributed Computing and Artificial Intelligence, pp. 140–147 (2017). Springer

[32] Amoretti, M., Belli, L., Zanichelli, F.: Utravel: Smart mobility with a novel user profiling and recommendation approach. Pervasive and mobile computing **38**, 474–489 (2017)

[33] McKercher, B., Tolkach, D., Eka Mahadewi, N.M., Byomantara, D.G.N.: Choosing the optimal segmentation technique to understand tourist behaviour. Journal of Vacation Marketing **29**(1), 71–83 (2023)

[34] D'urso, P., Massari, R.: Fuzzy clustering of mixed data. Information Sciences **505**, 513–534 (2019)

[35] Melnykov, V., Maitra, R.: Finite mixture models and model-based clustering. Statistics Surveys **4**, 80–116 (2010)

[36] Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Transactions on neural networks **16**(3), 645–678 (2005)

[37] Singh, P.K., Othman, E., Ahmed, R., Mahmood, A., Dhahri, H., Choudhury, P.: Optimized recommendations by user profiling using apriori algorithm. Applied Soft Computing **106**, 107272 (2021)

[38] Huang, Z.: Clustering large data sets with mixed numeric and categorical values. In: Proceedings of the 1st Pacific-asia Conference on Knowledge Discovery and Data mining,(PAKDD), pp. 21–34 (1997). Citeseer

[39] Li, C., Biswas, G.: Unsupervised learning with mixed numeric and nominal data. IEEE Transactions on knowledge and data engineering **14**(4), 673–690 (2002)

[40] Cheeseman, P.C., Stutz, J.C., *et al.*: Bayesian classification (autoclass): theory and results. Advances in knowledge discovery and data mining **180**, 153–180 (1996)

[41] Hsu, C.-C.: Generalizing self-organizing map for categorical data. IEEE transactions on Neural Networks **17**(2), 294–304 (2006)

[42] Ahmad, A., Khan, S.S.: Survey of state-of-the-art mixed data clustering algorithms. Ieee Access **7**, 31883–31902 (2019)

[43] Ping, Y., Gao, C., Liu, T., Du, X., Luo, H., Jin, D., Li, Y.: User consumption intention prediction in meituan. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 3472–3482 (2021)

[44] Gössling, S.: Tourism, tourist learning and sustainability: An exploratory discussion of complexities, problems and opportunities. Journal of Sustainable Tourism **26**(2), 292–306 (2018)

[45] Niaraki, A.S., Kim, K.: Ontology based personalized route planning system using a multi-criteria decision making approach. Expert Systems with Applications **36**(2), 2250–2259 (2009)

[46] Cannas, R.: An overview of tourism seasonality: Key concepts and policies. Almatourism-Journal of Tourism, Culture and Territorial Development **3**(5),

40–58 (2012)

[47] Moreno, A., Valls, A., Isern, D., Marin, L., Borràs, J.: Sigtur/e-destination: ontology-based personalized recommendation of tourism and leisure activities. Engineering applications of artificial intelligence **26**(1), 633–651 (2013)

[48] Jia, Z., Yang, Y., Gao, W., Chen, X.: User-based collaborative filtering for tourist attraction recommendations. In: 2015 IEEE International Conference on Computational Intelligence & Communication Technology, pp. 22–25 (2015). IEEE

[49] Borràs, J., Moreno, A., Valls, A.: Intelligent tourism recommender systems: A survey. Expert systems with applications **41**(16), 7370–7389 (2014)

[50] Hakimi, S.L., Yau, S.S.: Distance matrix of a graph and its realizability. Quarterly of applied mathematics **22**(4), 305–317 (1965)

[51] Struyf, A., Hubert, M., Rousseeuw, P., *et al.*: Clustering in an object-oriented environment. Journal of Statistical Software **1**(4), 1–30 (1997)

[52] Müllner, D.: Modern hierarchical, agglomerative clustering algorithms. arXiv preprint arXiv:1109.2378 (2011)