

Bollenbach, Jessica; Rebholz, Dominik; Keller, Robert

**Article — Published Version**

## The road not taken: Representing expert knowledge for route similarities in sustainable tourism using machine learning

Electronic Markets

**Provided in Cooperation with:**

Springer Nature

*Suggested Citation:* Bollenbach, Jessica; Rebholz, Dominik; Keller, Robert (2025) : The road not taken: Representing expert knowledge for route similarities in sustainable tourism using machine learning, Electronic Markets, ISSN 1422-8890, Springer, Berlin, Heidelberg, Vol. 35, Iss. 1, <https://doi.org/10.1007/s12525-025-00816-5>

This Version is available at:

<https://hdl.handle.net/10419/330655>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



<https://creativecommons.org/licenses/by/4.0/>



# The road not taken: Representing expert knowledge for route similarities in sustainable tourism using machine learning

Jessica Bollenbach<sup>1,2,3</sup> · Dominik Rebholz<sup>1,2,3</sup> · Robert Keller<sup>4</sup>

Received: 29 October 2024 / Accepted: 1 July 2025  
© The Author(s) 2025

## Abstract

As recreational tourism in rural areas rises in popularity, overtourism, and crowding pose growing challenges, impacting both society and the environment. To support sustainable smart tourism, an information system for visitor management offers a valuable approach. A significant challenge in this context is the identification of suitable alternatives to congested areas. This paper proposes a method to calculate route similarities with distance-based algorithms and machine learning models using descriptive data to redirect visitors to less-crowded paths. A case study in a nature park validates the approach, using real-world hiking data from an online outdoor platform. Expert surveys on route similarity are used to train the models and evaluate the results. Machine learning significantly outperforms traditional similarity algorithms, achieving up to 117% higher  $R^2$  values (0.448 vs. 0.206), 26% lower MSE values (0.530 vs. 0.719), and 40% higher Spearman correlations (0.699 vs. 0.498). The random forest regression model yields the best results. This research provides a foundation for future efforts to enhance sustainable tourism by offering a data-driven approach to identifying alternative routes that align with visitor preferences.

**Keywords** Visitor management · Route similarity · Machine learning

**JEL Classification** C38 · L83 · Z32

## Introduction

*“Two roads diverged in a wood, and I—  
I took the one less traveled by,  
And that has made all the difference.”*  
~ The Road Not Taken, Robert Frost (1916)

The tourism industry has experienced rapid growth in past decades, from 25 million international arrivals in 1950 to over 1.4 billion in 2019 (UN Tourism, 2022). Although the COVID-19 pandemic immensely disrupted the industry, international tourism recovered 63% of pre-pandemic levels in 2022 and even reached 98% in the first nine months of 2024 (UN Tourism, 2023, 2024). With the rapid recovery of tourism, the benefits of economic growth and employment opportunities are returning, however, so is the problem of crowding, which poses a major concern for many popular tourist destinations (Palacios-Florencio et al., 2021; Tiwari & Chowdhary, 2021).

Generally, the arrival of too many tourists at popular destinations leads to unsustainable tourism (Capocchi et al., 2019; Mihalic, 2020). UN Tourism defines the capacity of a tourist destination according to sustainable tourism as “the maximum number of people that may visit a tourist destination at the same time, without causing destruction of the physical, economic and sociocultural environment and an unacceptable decrease in the quality of visitors’ satisfaction” (UN Tourism, 2018, p. 3). Thus, crowding and overtourism are widely discussed topics in urban (Koens et al., 2018; UN Tourism, 2018)

---

Responsible Editor: Jinwon Kim

---

✉ Dominik Rebholz  
dominik.rebholz@fit.fraunhofer.de

<sup>1</sup> Branch Business & Information Systems Engineering of the Fraunhofer FIT, Alter Postweg 101, 86159 Augsburg, Germany

<sup>2</sup> FIM Research Center for Information Management, Alter Postweg 101, 86159 Augsburg, Germany

<sup>3</sup> University of Bayreuth, Wittelsbacherring 10, 95444 Bayreuth, Germany

<sup>4</sup> University of Applied Sciences Kempten, INIT, Bahnhofstraße, 61, 87435 Kempten (Allgäu), Germany

and rural (Drápela et al., 2021) tourist destinations, whereby a tourist destination can be either a single point of interest (POI) or a route. Crowding causes environmental degradation, loss of natural biodiversity, and increased pollution (Peeters et al., 2018). Furthermore, it negatively impacts the local infrastructure, tourist infrastructure, and living conditions, as well as the cultural heritage of the residents (Adie et al., 2020; Drápela et al., 2021; Milano et al., 2021). While the latter primarily affects nature and local residents, crowding also deteriorates the tourist experience (Tokarchuk et al., 2022).

Most efforts to mitigate the effects of crowding are often insufficient to actually prevent crowding as a cause of these effects, as many approaches focus on a single tourist destination (Butler & Dodds, 2022). These approaches include investments in better local and tourist infrastructure, e.g., additional parking lots, bypasses, better trails, or the limitation of the number of visitors at a specific POI (Dodds & Butler, 2019; Koens et al., 2018; Oklevik et al., 2019). There are already some approaches in the information systems (IS) community to reduce crowding with smart tourism (Gretzel et al., 2015), such as data-driven smart cities (Kühne & Heidel, 2021; Lamsfus et al., 2015) and decision support or recommender systems (Bauer et al., 2021; Piliponyte et al., 2024). By predicting the tourism demand with machine learning algorithms, tourist destinations are able to better prepare in advance for days with peak demand (Bollenbach et al., 2022). However, to avoid crowding and enable sustainable tourism, preparation at the destination itself is not sufficient. Therefore, overarching measures are required to distribute visitors across multiple tourist destinations and reduce the number of people at crowded destinations without overloading another region. One approach for sustainable IS are “data-based systems that recommend sustainable product alternatives” (Lehnhoff et al., 2021, p. 219; Tomkins et al., 2018). This concept is deeply intertwined with the principles of Green IS, which seeks to transform economic systems and individual behaviors to foster environmentally responsible practices (Henkel et al., 2019; Kranz et al., 2015). In this case, a visitor guidance concept represents a potential solution. In the event of predicted crowding at a local hotspot (Bollenbach et al., 2022), a recommendation system proactively nudges tourists to alternative tourist destinations, i.e., an alternative POI or route in advance. Thus, the effectiveness of these systems strongly depends on how alternative destinations are identified. Most current approaches recommend tourist destinations based on geological data and distance, e.g., the same starting point, which does not solve the crowding problem as most tourists still use the same infrastructure (Pham et al., 2017; Wang et al., 2015; Yin et al., 2013). The alternative route may not align with the visitor’s preferences. Furthermore, many recommendation algorithms often suggest popular POI instead of less overcrowded alternatives (Yao et al., 2016; Yuan et al.,

2013, 2014). This only increases the number of tourists at trending tourist destinations.

To address the crowding problem, the recommendation system should be based on the similarity of the tourist destinations rather than solely based on geographical proximity. A study from the United States Department of Agriculture (2023) National Visitor Use Monitoring Program reveals that over half of visitors (54%) choose alternative destinations based on activity similarity, underscoring the importance of this factor. However, the same report indicates that geographical proximity also plays a critical role, as visitors prefer alternatives close to their residences (United States Department of Agriculture, 2023). Therefore, with the similarity of destinations as the basis of the recommendation system, tourists can choose from more suitable alternative destinations in an acceptable distance from their residences, and the crowding of overloaded regions can be reduced. Furthermore, identifying similar routes should be widely automated and digitized to enable large-scale use. It is important to note that more drastic measures, such as limiting the number of tourists or offering timed-entry reservations, might be more suitable at very popular tourist hotspots such as Venice (Bertocchi et al., 2020), Machu Picchu (Larson & Poudyal, 2012), and the US National Parks (Creany et al., 2024), since tourists at these iconic sites are often highly focused on visiting specific landmarks and are less likely to be persuaded by alternative recommendations.

While the similarity of POIs has been addressed in prior research (Qiu et al., 2021; Wang et al., 2019; Zhao et al., 2019), there is a notable lack of studies focusing on the similarity of entire routes. Unlike POIs, routes are complex entities composed of multiple POIs and diverse descriptive features such as distance and elevation. As a result, crowding in rural tourism settings often extends beyond individual POIs to affect entire route networks. To effectively mitigate such crowding, alternative routes must be identified outside of these congested areas, rather than simply redirecting visitors from one crowded POI to another nearby. This issue is especially pressing in rural tourism, where hiking and other recreational activities have seen an increase in popularity in recent years (Calbimonte et al., 2021). However, due to the structural and functional differences between POIs and routes, existing POI similarity methods cannot be readily transferred to route analysis. Despite the growing importance of route-based crowding management, systematic methods for assessing route similarity remain scarce. Thus, to the best of our knowledge, we are among the first to develop a method to determine the similarity of routes based on their descriptive features and propose the following research questions:

**RQ1:** *How accurately do various distance-based similarity and machine learning algorithms calculate route similarities using labeled and unlabeled data?*

**RQ2:** *What are relevant features and how do they influence the prediction of similar routes?*

To answer the research questions, we propose a method to evaluate the similarity of routes by applying and comparing two distance-based similarity and five machine learning algorithms. This approach is then applied and tested in a case study about the similarity of hiking routes in a nature park in the South of Germany with descriptive company data provided by the online outdoor platform “Outdooractive.” In the case study, we integrate human expert knowledge by training and evaluating the models using survey results on the similarities of hiking routes conducted by nature park rangers. Furthermore, since many machine learning algorithms act like a “black box” (Bauer et al., 2021; Pfeuffer et al., 2023), we use Shapley Additive exPlanation (SHAP) values to shed light on the most significant descriptive features of the hiking routes and how they influence the prediction model.

By answering the research question, we advance the literature in four key ways. First, we shift focus from POI and itinerary similarity (e.g., Yu et al. (2022); Qiu et al. (2021)) to the under-explored domain of route similarity, using hiking routes in rural areas as a case. While prior work has incorporated POI popularity, distance, and preferences, route similarity as a structured item remains largely unaddressed, despite its critical relevance for managing visitor flows and promoting sustainable tourism in increasingly overburdened rural areas. Second, we propose a systematic methodology combining distance-based and machine learning approaches to quantify route similarity based on descriptive attributes, independent of user histories. This extends item-to-item similarity concepts (Stamatelatos et al., 2021) to a novel domain and data structure. Third, we evaluate algorithmic performance against expert similarity ratings using explainable AI (SHAP), identifying which features shape perceived similarity. Finally, our research yields actionable insights into route similarity that can serve as input for recommender systems in the context of IS. Ultimately, our work contributes to the development of sustainable, data-driven visitor guidance systems, laying the groundwork for intelligent IS solutions that promote more sustainable tourism practices.

The present paper is structured as follows: First, the relevant background in visitor management and similarity is explored. Subsequently, the applied methods, as well as the case study, are presented. The results are described, evaluated, and discussed. Finally, the paper is summarized, mentioning the limitations as well as future work.

## Background and related work

### Crowding and visitor management

Overtourism is an umbrella term for issues such as crowding, environmental degradation, and social and cultural change (Benner, 2020). Although the research area of overtourism has become increasingly popular in recent

years, scientists have already been studying the underlying adverse impact on tourist destinations as early as the 1960s (Koens et al., 2018). Most studies on overtourism focus on urban areas, most prominently Venice (Bertocchi et al., 2020; Seraphin et al., 2018) or Barcelona (Alonso-Almeida et al. 2019). Nevertheless, overtourism also strongly affects rural areas (Drápela et al., 2021; Peeters et al., 2018). The demand for domestic tourism and new recreational activities, such as day and half-day trips to nature-based areas, has increased significantly (Falk et al., 2022; Fillisch et al., 2020; Megerle, 2021). Contrary to closed areas such as buildings, the carrying capacity in an open space is difficult to define. Therefore, we use the wording “crowding” instead of “overcrowding.” The rising demand for outdoor activities impacts not only the tourist destination but also all surroundings like infrastructure, residents, and tourist experience. Therefore, since rural tourist destinations often lack the infrastructure required to support the rising numbers of tourists, crowding leads to several problems in the local infrastructure, e.g., parked cars in inappropriate locations or street congestions, and the tourist infrastructure, e.g., crowded parking lots or damaged trails (Drápela et al., 2021; Purwandani & Pakan, 2022). Furthermore, the environmental impact of overtourism is well established in current research. Natural sites with too many visitors experience garbage and waste dumping (Peeters et al., 2018), damage from erosion, as well as endangerment of native species (Drápela, 2023; Drápela et al., 2021; Megerle, 2021). Hiking, in particular, has become increasingly popular in recent years, increasing the chance of crowding on hiking routes and posing a substantial negative impact on tourist destinations, causing environmental degradation and loss of biodiversity (Bichler & Peters, 2021; Calbimonte et al., 2021; Chhetri et al., 2004; Megerle, 2021; Nordbø & Prebensen, 2015; Smith et al., 2023). Moreover, house prices in popular tourist areas increase, leading to gentrification and decreased living conditions of residents in the area (Drápela et al., 2021). In some areas, this results in tourism-phobia and anti-tourism sentiment among locals (Insch, 2020; Zerva et al., 2019). However, some scholars, such as Prof. Butler, critique the discourse on overtourism as being frequently exaggerated and ideologically driven (Gillbanks, 2023). While acknowledging issues of overcrowding, he argues that mass tourism is frequently used as a scapegoat for broader urban and environmental challenges. He emphasizes the need for a more pragmatic, evidence-based approach to tourism management, particularly in rural areas, where sudden visitor influxes strain infrastructure.

Introducing a visitor management system is crucial to mitigate the negative effects of crowding. Shackley (2000) defines the goal of visitor management as balancing the

desires of visitors with the impact of visitors. There are two approaches to implementing visitor management, referred to as “hard” and “soft” approaches (Kuo, 2002). While “hard” visitor management is mandatory, “soft” visitor management is optional and may result in the self-regulating behavior of the tourists (Mason, 2005). Kuo (2002) separates the “hard” approach in three subcategories: (1) physical (e.g., constructions to protect nature sites, viewing platforms) (Butler & Dodds, 2022); (2) regulatory (e.g., rules and regulations to limit the number of visitors) (Bertocchi et al., 2020); (3) economic (e.g., dynamic pricing or increased entry fees to distribute visitors to less popular days) (Enseñat-Soberanis et al., 2019; Orams, 1996). On the other hand, “soft” approaches appeal more to the visitor’s goodwill by informing them about alternative tourist destinations or offering directional information like signs or a code of conduct (Kuo, 2002). In addition, UN Tourism proposes eleven strategies, including both “hard” and “soft” measures, to implement visitor management in cities to avoid crowding (UN Tourism, 2018). The strategies can be transferred well to rural areas, whereby the explicit measures have to be adapted to the circumstances. The first strategy describes the dispersal of tourists to alternative locations as a soft approach and, thus, is the most relevant for this paper. While promoting locations solely based on a lower visitor density is possible, visitors are more likely to accept alternative locations suitable to their individual requirements and preferences (Boháč & Drápela, 2022; Calbimonte et al., 2021). Therefore, in this paper, we determine the similarity of routes to enable data-driven recommendations in rural areas. We aim to achieve a higher acceptance among tourists for alternative routes and, thus, reduce crowding.

## Similarity

To identify the similarity of routes, we first evaluate research on similarity in general. Ifenthaler (2012, p. 2148) defines similarity as “a corresponding feature in which two objects or variables are alike.” In mathematics and information science, similarity is usually a numerical value between “0” (no similarity) and “1” (perfect similarity), whereby the opposite of similarity is the distance or dissimilarity. Hence, similarity is a broad concept that is required in various kinds of use cases, e.g., similarity between values (e.g., between “0” and “1”), string similarity (“eloquently” and “ineloquent”), semantic similarity (“car” and “automobile”), etc. (Lin, 1998). Building upon this concept, similarity analysis is an established method in various fields, such as biomedical science (Lord et al., 2003; Pesquita et al., 2009), chemistry (Maggiore et al., 2014; Maldonado et al., 2006), retail (Chen et al., 2005; Choi et al., 2010; Hwangbo et al., 2018), e-commerce (Li et al., 2013; Maleki Shoja & Tabrizi, 2019), social media (Anderson et al., 2012; Becker et al., 2010), etc. The

applied method within the similarity analysis depends on the data and differs between use cases. Since our focus lies on route similarity, we briefly analyze the most relevant papers in the tourism domain that include similarity analysis.

In tourism, similarity analysis is mostly common for urban regions by determining the similarity of POIs within a tourism destination. POIs can be compared by analyzing the description or online reviews of POIs (Stamatelatos et al., 2021), for example, using generative topic modeling (Adams & Raubal, 2014) or neural-network-based semantic models, such as Word2Vec or recurrent neural networks (Qiu et al., 2021; Wang et al., 2019; Zhao et al., 2019). Other approaches to identify similarities between POIs include using a representation of the POI as points in a vector space by collecting detailed data (e.g., categorical labels). This data is used to compute the pairwise similarity using the standard cosine similarity (Preotăciuc-Pietro et al., 2013). Furthermore, Cao and Thomas (2021) use anonymized cookie IDs from search logs to discover which destinations were searched by the same user. If various users search for two destinations on the web, they conclude that those destinations are similar.

Turning to route similarity, multiple approaches analyze user-to-user similarities (Stamatelatos et al., 2021), such as driver GPS data (Oliveira e Silva et al., 2022) or tourist movement patterns using GPS data, to capture similarity patterns between different persons (Chen et al., 2020; Cui et al., 2018; Zheng et al., 2019). Tsai et al. (2024), in contrast, cluster tourists based on the similarity of their visited POI sequences. However, these approaches primarily investigate individual trajectories and movement behavior, rather than assessing the similarity of predefined routes. A related line of research shifts the focus to the characteristics of the routes themselves. For instance, Calbimonte et al. (2020) categorize the difficulty of hiking routes based on descriptive attributes, e.g., the name, description, coordinates, and user ratings. This classification can be used to compare routes with similar difficulty and to identify suitable routes for restricted users, e.g., due to disabilities (Calbimonte et al., 2021). While this approach provides important insights for specific user groups, it misses the overarching comparability of routes. Vías and Ocaña (2014) follow a different approach to determine the suitability of hiking routes: using a geographic IS, they analyze various features such as the “hikeability” (e.g., type of surface, slope), natural interests (e.g., vegetation, POIs), and land management (e.g., public spaces, protected natural areas). However, while Vías and Ocaña (2014) calculate the similarity of hiking routes, they only do so compared to an ideal route. Issa et al. (2016) go a step further and calculate the similarity between different running routes based on numerical and categorical features such as distance, elevation, off- or on-road segments, location, etc., with the Euclidean distance. Although they calculate



the overarching similarity of running routes, they only use Euclidean distance, and their primary focus is evaluating a recommendation system instead of identifying the best similarity analysis method for routes. Thus, the accuracy of the route-to-route similarity is not evaluated by including further methods such as Gower distance or machine learning approaches.

Furthermore, to facilitate a comprehensive exploration, comparing the similarity of routes has been done in practical analytics without influence in research. Candocia (2020) uses GPS data of running routes to compare their similarity. Thus, if sections of the routes are the same, the similarity is higher. Needham (2018) follows a similar approach to calculate the similarity of two runs using coordinates and dynamic time warping. However, they do not provide a detailed evaluation of the results. In addition, outdoor companies such as Komoot, Outdooractive, AllTrails, or Bergfex offer applications including navigation and route planning for tours, often recommending similar or alternative hiking routes (Bergfex, 2023; Komoot, 2021; Outdooractive, 2023). Based on an expert interview, those similarity and recommendation algorithms are still relatively simple and determine similar routes based on the route's location, category, and popularity. These simple recommendation algorithms are very restricted because they mostly recommend routes based on the distance of the selected route with the same starting point. However, the mere fact that two routes originate from the same point does not guarantee a high degree of similarity. Additionally, if the same starting point is used, the same infrastructure (e.g., parking) is employed for both routes and thus will still be affected by crowding or local restrictions.

To give a brief overview of the related work, we organize the literature into four thematic clusters: POI similarity, route assessment, route recommendation, and route similarity. Studies in the first category predominantly focus on identifying similarities between individual points of interest (POIs) using text-based descriptions, user behavior, or semantic embeddings (e.g., Adams and Raubal (2014); Stamatelatos et al. (2021)). These approaches typically compare discrete attractions rather than entire travel paths. The second category, route assessment, emphasizes evaluating hiking or travel routes based on objective characteristics such as terrain, accessibility, signage, or perceived difficulty (e.g., Calbimonte et al. (2020); Vías and Ocaña (2014)). This strand does not focus on similarity per se but rather on classifying or improving individual routes. Route recommendation models, in contrast, leverage historical movement data or collaborative filtering to suggest personalized itineraries, often based on shared patterns among tourists (e.g., Cui et al. (2018); Mou et al. (2022)). While these studies consider user preferences, they rarely examine route-to-route similarity explicitly. Finally, route similarity research engages most

directly with the notion of comparing entire travel paths, typically using GPS traces or spatiotemporal sequences (e.g., Chen et al. (2020); Zheng et al. (2019)). However, these studies often analyze user trajectories rather than predefined routes, or they rely on limited similarity metrics such as Euclidean distance (Issa et al., 2016), without systematically comparing methodological alternatives. A more detailed overview is depicted in the appendix (Table A.2).

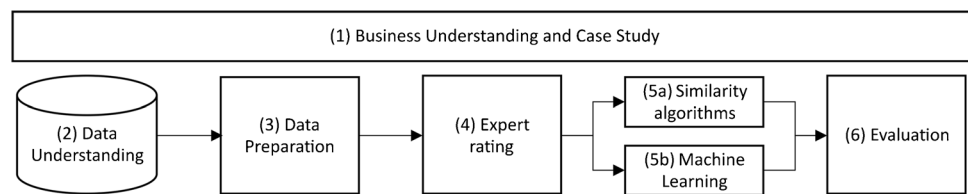
Our study departs from previous work by offering a systematic evaluation of different methods to quantify the similarity between predefined hiking routes in rural areas. Unlike approaches focused on route recommendation or trajectory clustering, our emphasis lies not on user personalization, but on the inherent comparability of the routes themselves, irrespective of who traverses them. In doing so, this paper contributes to the IS literature in three distinct ways. First, it provides an extensive analysis of route similarity based on descriptive route attributes, employing both conventional similarity metrics and advanced machine learning techniques. Second, an expert rating for a human similarity assessment is conducted, which allows an accurate evaluation of the calculated results. Third, it applies explainable AI to identify the most influential features for perceived similarity, thereby offering deeper insights into the structure of route comparability. We thus provide a comprehensive methodology for assessing the similarity of routes and create the basis for data-driven recommendation algorithms to avoid crowding that can be transferred to other use cases besides hiking. Accordingly, we make a valuable contribution to the IS community by providing tools that enhance understanding of similarity concepts, which serve as algorithmic input for future IS and recommendation engines.

## Methods and case study

The methodology of this paper is based on the Data Science Trajectory (DST) map by Martinez-Plumed et al. (2021). The DST map allows a more flexible approach to the cross-industry standard process for data mining (CRISP-DM) method, which includes six goal-oriented phases from business understanding to deployment in a fixed order (Wirth & Hipp, 2000). Figure 1 depicts the applied methodological approach for this paper.

An overview of the underlying business case is provided (1), focusing on a nature park located in southern Germany, near the border with Austria. The hiking routes from this region available on the Outdooractive database are identified during the data understanding phase (2) and prepared for the further process (3). Following, fifty sample routes are randomly selected from the prepared dataset for the rangers from the nature park to rate the similarity with a rating tool (4). To calculate the similarities of these routes, we use

**Fig. 1** Overview of the research methodology, portraying the six phases from business understanding to evaluation



distance-based similarity algorithms (5a) as well as machine learning models to improve the results (5b). Both results are evaluated on their conformity, correlation, and performance in comparison to the expert ratings (6).

### Business understanding and case study

Nowadays, many outdoor trips are planned individually by using platforms (e.g., Komoot, AllTrails) or AI-powered itineraries (e.g., ChatGPT, BookedAI, mindtrip.ai) on the internet, which often do not consider local regulations and protected areas. Outdooractive is an example of these online outdoor platforms. As of June 2023, their database consists of over 960,000 routes and about 760,000 POIs (Corporate Outdooractive, 2023). It includes routes for hiking, cycling, mountain biking, mountain hiking, climbing, snowshoeing, cross-country skiing, etc., around the globe. These can be accessed on the Outdooractive website or the Android and iOS apps. Community members can upload and rate routes without having in-depth knowledge of the region, resulting in routes that may include restricted or protected areas. Unsustainable routes that were an insider tip can easily gain popularity on the internet, leading to conflicts between people and nature. Although the undesirable routes can be manually deleted, their creation cannot be controlled, resulting in an enormous additional effort. To overcome this problem, Outdooractive enables representatives of protected areas to promote more sustainable routes as verified partners to raise awareness. Furthermore, the protected areas can also be added to the Outdooractive database to reduce unwanted visitor flows into ecologically sensitive zones (Wimmer & Rödig, 2019). To prevent the utilization of unofficial paths, the rangers of the nature park, who also answered our expert survey, started an initiative called “Digital Ranger” to analyze the most successful measures to guide tourists to sustainable routes (Schwietering et al., 2023). One way to mitigate the effects of crowding on unsustainable routes would be to recommend similar routes on sustainable routes on outdoor platforms such as Outdooractive.

### Data understanding

All relevant hiking route data was identified based on the geographical location within the nature park and downloaded using the Outdooractive API. The complete data

consists of 73,976 routes and 143 features. The column information can be grouped into the following two classes: Descriptive data describes the route, e.g., the *length*, *title*, *GPS data*, *short text description*. Metadata is necessary for depicting the route on the Outdooractive platform, e.g., the *route’s creator*, *publishing date*, *images*, the *color* in which the route is depicted on the map. Many features include a large amount of missing data, requiring further data preparation. Potential reasons include incomplete entries on the database by the route creators, category-specific columns (e.g., *rope length* for climbing routes), and data only suitable for specific routes (e.g., *labels.top* is only applicable to the most popular routes which earned the label “top”).

### Data preparation

The removal of large amounts of missing values is common, but also widely discussed in research (Liu et al., 2020; Penone et al., 2014). The main reasoning against the removal is the decrease in statistical power due to a smaller sample size. This rationale does not apply to this use case since the goal of this process is to create two sample datasets of 50 routes with excellent data quality. One sample dataset includes only a minimum number of features, while the extended dataset also includes tags of the routes. The following steps are undertaken for the data selection and cleaning process.

Outdooractive has an internal measurement to review the quality of the route data called ranking, ranging from 0 (low data quality) to 100 (high data quality). The ranking considers the level of route detail, e.g., whether all information is complete, route descriptions are filled out, and user interactions like comments and ratings exist. After evaluating the potential cut-off point, all routes with a ranking lower than 50 are removed. A higher cut-off point would have resulted in better data quality, but also a lower number of remaining routes. Furthermore, since a high amount of missing data can have negative effects on the similarity algorithm as well as the machine learning model (Zhang et al., 2003), all features with over 40% missing values are removed. Low percentages of missing data can be imputed, but the accuracy of the imputation methods decreases with over 30–60% missing values (Penone et al., 2014; Xue et al., 2022).

For this research, only one-day routes are considered. There is no universal definition of how many hours a one-day hike takes. For an experienced athlete, a one-day hike can last over twelve hours, while inexperienced hikers might consider a 6-h hike a one-day one. Anagnostopoulos et al. (2017) suggest that while a ten-hour walk is possible for a group of young adults, more senior visitors are often only able to walk for half the time. Vías et al. (2018) set the maximum limit for hiking routes at nine hours. Everything over that limit is considered “extreme” by them. For this paper, a limit of ten hours is set to allow a broader range of hiking routes, including more challenging routes for experienced athletes. All routes that take more than ten hours (approximately 3.69% of all routes) are removed from the dataset. Other data preparation steps include the removal of non-numeric features (e.g., *route descriptions*), subjective ratings that were inconsistently rated between different authors (e.g., rating of the landscape/experience), and hiking-irrelevant data (e.g., *rope length*). These features were excluded to ensure consistency in the dataset. Furthermore, routes which are either impossible to be finished, e.g., with the specified duration and length, or just unlikely are removed based on an outlier analysis. Univariate outliers are identified using Tukey’s fences (Tukey, 1977). For the multivariate outliers, the Mahalanobis distance (Maesschalck et al., 2000), Isolation Forest (Liu et al., 2008), and DBSCAN (Ester et al., 1996) were tested on non-transformed and log-transformed data due to skewness of the dataset. The application of DBSCAN on the log-transformed data detects all outliers while retaining a high variance in the data and thus is used for the outlier detection. Table 1 provides the descriptive statistics of the continuous features remaining in the dataset.

During feature engineering, we create new features because raw data is often unsuitable for similarity and machine learning algorithms (Heaton, 2016). This includes the creation of a *roundtrip* feature, which is true if the first and last location in the GPS data are within a close range, as well as the identification of the most relevant tags (e.g., *scenic*, *dining*, *suitability for families*, *culture*, etc.) in the dataset. Tags are binary features that represent certain properties

of the route, for example, the tag *dining* means that there are restaurants or other places to buy food along the route. Due to a high percentage of missing values, only 20 of 34 tags are used in the dataset. Furthermore, the *months* in which the routes are suitable, originally included as separate features in the dataset, exert a strong influence on the similarity algorithms. Therefore, they are removed, and the relevant temporal information is consolidated into a single feature (Lesot et al., 2009). Over 95% of the routes are suitable for summer, while less than 50% are suitable for winter. Since primarily the winter months differ in suitability, a new feature *winter tour* is introduced, which is true if the route is suitable in summer and winter.

After cleaning, 50 routes are picked randomly from the remaining 441 routes using random sampling to get two sample datasets for further evaluation (Singh, 2003). Both contain the same routes, but in addition to the base dataset, the extended dataset also contains the tag features (e.g., *dining*, *suitable for families*). While the selected tag features provide useful additional information, certain potentially relevant details, such as quantifiable aspects of tags (e.g., exact number or quality ratings of dining places) or additional POI data (e.g., waterfalls and caves), were not available within our existing dataset and thus could not be included.

Both datasets are standardized before calculating the similarities to ensure that all features are on the same scale and, therefore, no feature with a larger scale (e.g., *length*) could have a greater impact (Lesot et al., 2009). After evaluating multiple methods, we decided to use the Standard Score to normalize the values (Kumar & Nancy, 2014). While applying Standard Score normalization to binary data is less common, it ensures a consistent scale across all features, preventing unintended dominance of certain variables in the similarity calculations.

The data is tested for multicollinearity using the Variance Inflation Factor (VIF) (Alin, 2010). The *length* and *duration* have a high VIF (8.487 and 12.491, respectively). This is due to the nature of the data being dependent on each other. Nevertheless, both features are relevant and cannot be easily combined into a single feature. Therefore, both

**Table 1** Descriptive statistics of the continuous route features in the base dataset ( $N=50$ ). Values include mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum for route length, duration, ascent, descent, and altitude measures

	Length (in meters)	Duration (in meters)	Ascent (in meters)	Descent (in meters)	Min Altitude (in meters)	Max Altitude (in meters)
Mean	12,301.8	257.64	609.8	650.18	864.32	1399.82
Std	5508.9	128.94	393.32	399.65	215.12	353.46
Min	2491.9	45	33	31	548	597
25%	7730.8	157.75	357.25	401.75	730	1095
50%	11,255.9	229.5	524.5	622	777.5	1485.5
75%	16,310.3	360	827	868	945	1727
Max	26,417.2	514	1516	1649	1422	1834



features are kept separate with the knowledge of them being collinear. This does not necessarily influence the results of the models directly but could reduce the significance of the feature importance for these two features. In addition to the VIF analysis, we further investigated potential multicollinearity among the binary features using the Phi coefficient (Ekström, 2011). The Phi coefficient, which measures the strength of association between two binary variables, revealed that the majority of pairs of binary variables show no significant correlations. However, the *fauna* and *flora* features, with a correlation of 0.677, and the *exposed* and *ridge* features (correlation of 0.66), indicate some degree of collinearity. While moderate correlations may reduce the feature importance, these features do not compromise the model results and remain relevant to the analysis and are therefore kept in the dataset.

### Expert survey

An expert survey is conducted in collaboration with Outdooractive and the nature park on a sample dataset of 50 routes to produce a labeled dataset of route similarity in order to train the machine learning models and evaluate the results. This process leverages the deep field knowledge and practical experience of nature park rangers, who are tasked with rating the similarity of two routes via a specifically designed online tool, as depicted in Fig. 2 (left). The tool depicts both routes on a map with additional information such as the route name, a short description, information about the length, duration, altitude, tags, and a link to the route on the Outdooractive website for more information. We intentionally refrained from providing explicit criteria to the experts for evaluating route similarity. This deliberate decision was made to avoid nudging and to allow the experts to rely on their implicit, experience-based knowledge. By not imposing strict guidelines, we aimed to capture their

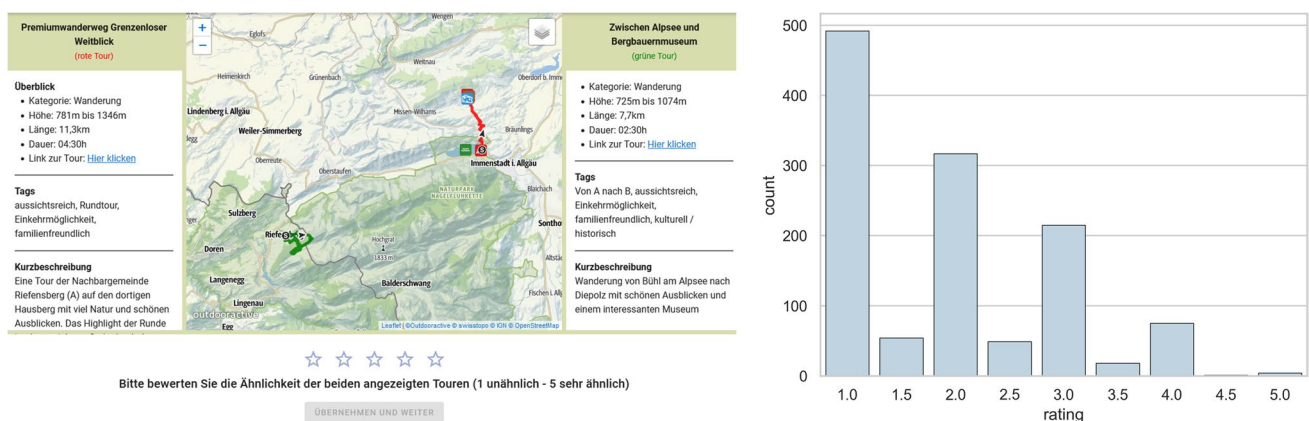
natural and subjective perceptions of route similarity derived from their extensive regional expertise.

The rating is on a scale from 1 (dissimilar) to 5 (very alike), whereby  $\binom{n}{2} = 1225$  unique route combinations are rated. After each rating, the next route pair is randomly picked from all the routes with the least number of ratings. Thus, it can be assured that the number of ratings for the route combinations is equally distributed. For routes that are rated more than once, the mean value of all ratings represents the final rating.

The survey was conducted over five weeks in the summer of 2022. Altogether, there are 1551 ratings, meaning the rangers rated each route pair at least once, with 326 route pairs rated twice. The results are displayed in Fig. 2 (right), illustrating that most route pairs are rated as dissimilar (1). For the double-rated routes, most of the routes received the same value or only differed by one level, supporting the robustness of the evaluation.

### Similarity and machine learning algorithms

To automate the digitization of expert knowledge on route similarity, we follow Shmueli and Koppius (2011) and apply two distance-based similarity algorithms and enhance the results by integrating predictive analytics using five machine learning algorithms. As an initial approach without prior assumptions regarding feature importance, we use the two unweighted algorithms: Euclidean distance and Gower distance (Gower, 1971). Thus, each feature has the same influence on the result. To improve the route similarity results and match reality more closely, we add weights to the features and train them on the expert survey ratings using machine learning algorithms. By adding weights to the distance function, the effect of certain features increases



**Fig. 2** Left: Screenshot of the interface of the expert survey tool for pairwise route similarity assessment; Right: Distribution of similarity ratings (1 = very dissimilar, 5 = very similar)

or decreases and better reflects human interpretation of similarity (Lesot et al., 2009; Shirkhorshidi et al., 2015).

The Euclidean distance, as the most commonly used distance-based algorithm (Lesot et al., 2009), provides baseline results. The algorithm is sensitive to outliers (Shirkhorshidi et al., 2015), but since outliers are already removed from the dataset, they do not influence the algorithm's results. Furthermore, the algorithm is unweighted and treats each feature as continuous data. In addition, we evaluate the performance of the Gower similarity, proposed by Gower (1971), as this similarity measure can handle datasets with mixed data types (numerical and categorical) (dos Santos & Zárate, 2015; Kuhn & Johnson, 2019). Since the sample dataset contains both continuous and binary features, we expect different results in comparison to the Euclidean distance. The predictions of the similarity algorithms are scaled to match the rating scale used by the experts by applying a Linear regression to the data with the expert rating as the dependent variable. The data is then multiplied with the model coefficient and added to the intercept.

To prepare the data for the training of the machine learning algorithms, we follow the transformation approach of Zerzucha and Walczak (2012) and Ahdid et al. (2016): The distance of a feature (e.g., *length*) across two routes is treated as the new feature. Each row in the transformed dataset represents a route pair combination. We apply the Euclidean distance; therefore, this feature describes the difference between both routes, since for one-dimensional data  $\sqrt{(x_2 - x_1)^2} = x_2 - x_1$ . This means that instead of indicating, for example, the *length* of a route, the feature is now the difference of the *lengths* between two routes. These distance features are then employed as inputs for commonly used machine learning regression algorithms (Ahdid et al., 2016). This transformation process is depicted in Fig. 3. The variables  $x$  and  $y$  represent two features (e.g., *length*, *duration*), and  $i$  and  $j$  two routes for which the pairwise similarity is calculated. All comparisons between the same routes (e.g.,  $i, i$  or  $j, j$ ) are removed from the dataset since they are exactly the same and the distance is always zero.

The dataset is split into training and test data for an accurate evaluation. To avoid overfitting, most research suggests

that 70–80% should be used for training and the remaining 30–20% for testing; thus, we use a 70/30 split in this study (Gholamy et al., 2018). The hyperparameters of the machine learning algorithms are tuned using grid search (Bergstra & Bengio, 2012; Liashchynskiy & Liashchynskiy 2019).

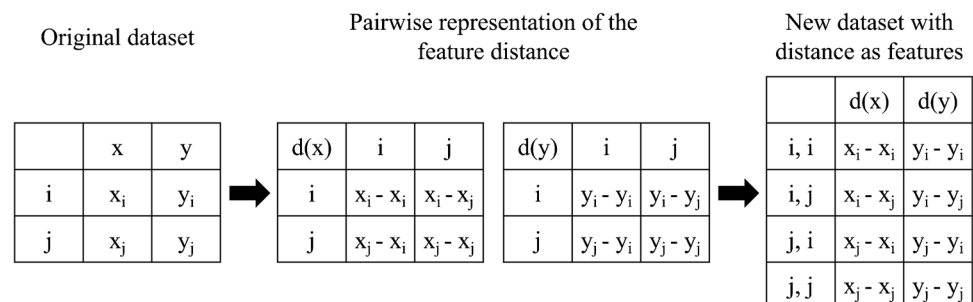
To evaluate the performance of simple and more complex machine learning algorithms, we selected a range of models. Given our dataset's unique structure, consisting exclusively of distance-based features, the algorithm selection must account for linearity, interpretability, handling of mixed feature types, and robustness to dataset size and dimensionality.

*Linear regression* is an established base method and describes the linear relationship between the independent and dependent variables (Weisberg, 2005). The method is transparent, computationally efficient, and robust for smaller datasets with straightforward linear relationships, making it suitable as a foundational benchmark. However, linear regression is limited because it assumes a linear relationship, which may not adequately capture complex, nonlinear patterns in how route similarity is judged by experts.

Support vector machines are one of the most popular supervised machine learning algorithms. Although primarily used for classification, they can also be used for regression tasks, which is then called *support vector regression* (SVR) (Martin, 2002). SVR addresses the limitations of linear regression by using kernel functions to map data into higher-dimensional feature spaces, enabling it to capture more complex, nonlinear relationships. We employ a radial basis function kernel for SVR due to its established generalization capabilities (Javed et al., 2009). Both SVR and the linear regression need standardized data so that they can appropriately weigh binary features in comparison to metric features (Shen et al., 2016; Tran et al., 2022).

*Decision trees* partition the data and fit a simple prediction model for each partition. Therefore, the model is highly interpretable, clearly indicating which features are most important and how these features interact with each other, while also effectively modeling nonlinear relationships (Tso & Yau, 2007; Xu et al., 2005). A higher depth of the tree might allow a better fitting of the model, which, on the other hand, might result in overfitting (Hawkins, 2004). We address this by systematically controlling tree depth.

**Fig. 3** Transformation of the original dataset into a pairwise format for machine learning. Feature distances are computed for each pair of instances and used as input features in a new dataset



*Random forest regression* is an ensemble method employing the bagging (bootstrap aggregating) method. This means that the model is created using the average results of various decision trees to improve the prediction. This approach significantly reduces overfitting risks, proving particularly beneficial given our limited dataset size and the discrete nature of our binary distance-based features (Biau & Scornet, 2016). Random forests often outperform neural networks, especially on less complex datasets (Fernández-Delgado et al., 2014).

The *XGBoost* (extreme gradient boosting) algorithm is also an ensemble method constructed from decision trees. It is a boosting algorithm, sequentially constructing decision trees, each addressing residual errors from preceding models. This sequential training strategy captures complex nonlinear relationships while ensuring high predictive performance (Chen & Guestrin, 2016). This makes this algorithm particularly suitable for our dataset's structure and size.

In addition to the models, we also evaluated the potential of neural networks and Gaussian process regression, which have shown great usefulness in modeling complicated nonlinear patterns across various research fields (Jin & Xu, 2024a; Jin et al., 2025). Neural networks are known for their capacity to model highly nonlinear relationships through deep layered architectures. Their ability to automatically learn feature representations makes them particularly powerful in capturing subtle patterns within complex datasets (Jin and Xu 2024c; LeCun et al., 2015). However, neural networks typically require large datasets to avoid overfitting and to achieve reliable generalization, which poses a challenge in our case given the limited sample size (D'souza et al., 2020). Moreover, their limited interpretability (Benitez et al., 1997; Liang et al., 2021) undermines our analytical focus on understanding the specific contributions of each distance-based feature to the model's predictions. Gaussian process regression represents another advanced approach we considered. As a nonparametric, probabilistic model, it is well-suited for regression tasks involving small datasets and offers informative uncertainty estimates for each prediction (Jin & Xu, 2024b; Rasmussen & Williams, 2006). However, standard Gaussian processes are mainly designed for continuous input variables, which makes it challenging to handle our mix of binary and metric features without custom kernel design (Kuok et al., 2025). While neural networks and Gaussian Process Regression offer valuable perspectives, the selected methods provide a more appropriate balance of interpretability, performance, and robustness for the dataset at hand. We consider the integration of more complex models an interesting avenue for future work when larger and more diverse datasets become available.

## Evaluation

To compare the results of the similarity algorithms and machine learning models with the expert ratings, we use

three primary metrics: Spearman correlation, mean squared error (MSE), and the coefficient of determination ( $R^2$ ). Furthermore, the best-performing model is analyzed in detail using SHAP (SHapley Additive exPlanations) values.

The Spearman correlation is a statistical metric for measuring association and is especially suitable for ordinal data. Spearman's rank correlation coefficient ( $\rho$ ) is a measure of rank correlation, which measures the monotonic relationship between two variables. This is done by replacing the observed values with rank scores ( $r_x, r_y$ ) and then calculating the Pearson correlation using the rank scores (Raadt et al., 2021).

The mean squared error is a well-established and widely accepted measure for evaluating the results of a regression model that penalizes large errors more heavily than smaller errors. MSE calculates how close the regression line fits the actual point, whereby a value of 0 is the best possible result and the worst value approaches  $+\infty$ . This measure should only be used to compare different machine learning models based on the same dataset (Chicco et al., 2021).

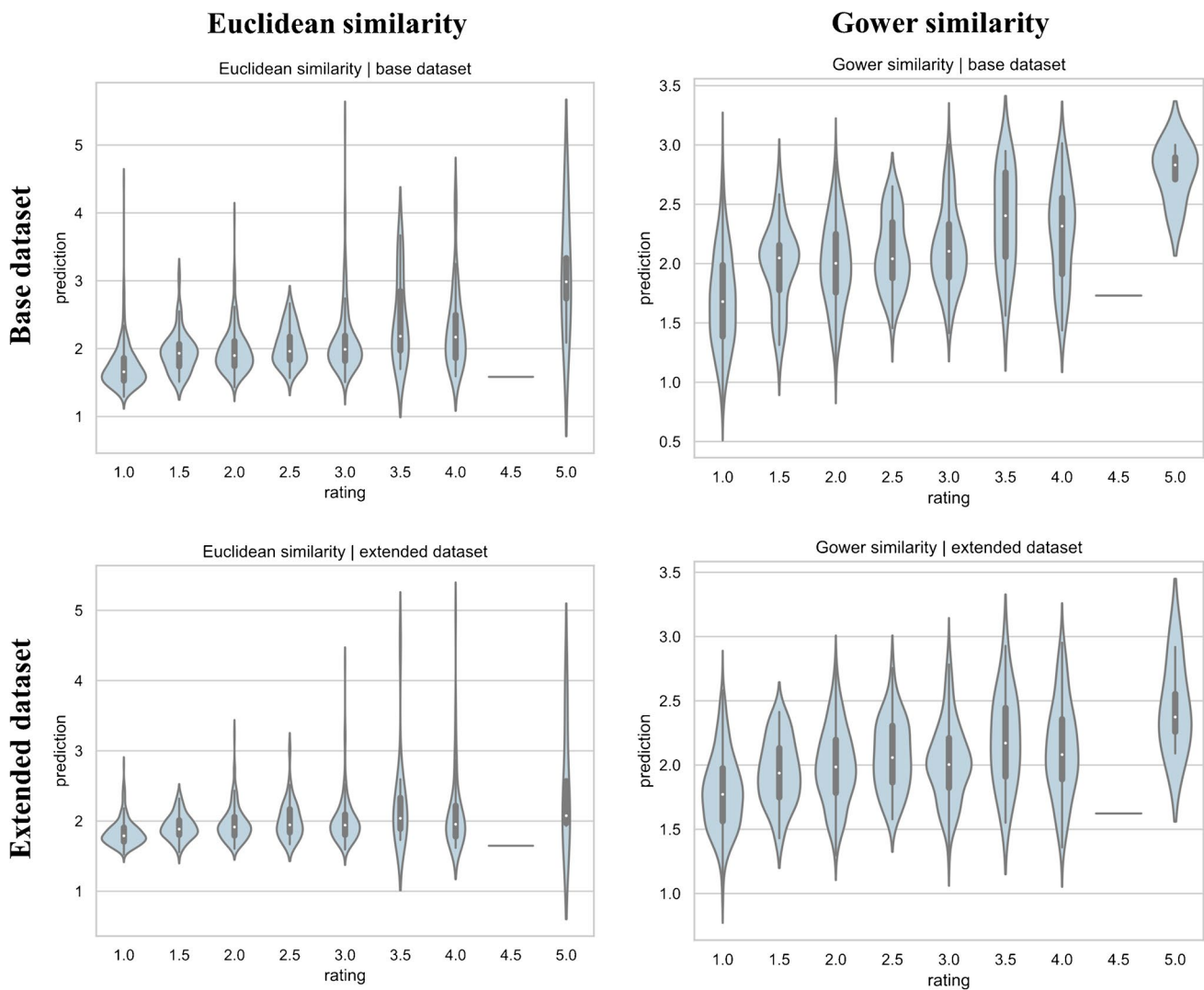
$R^2$  is a measure that allows the comparison of models across different datasets. Similar to MSE, it also describes how well a regression model fits the data. The possible range for  $R^2$  is from  $-\infty$  to 1, with 1 being the best possible value. The closer a value is to 1, the better the regression model (Chicco et al., 2021). Chin (1998) categorizes  $R^2$  at 0.19 as weak, 0.33 as moderate, and 0.67 as substantial.

SHAP is an approach introduced by Lundberg and Lee (2017) to interpret the predictions of a model. It assigns an importance value for each prediction to the features, e.g., how they influence the prediction. This paper uses this approach to accurately identify the importance of certain features and how they influence the model.

## Results

### Similarity algorithms

To determine the similarities between the hiking routes, we first calculate the Euclidean and Gower similarity for each route pair in the base dataset and in the extended dataset on unlabeled data. Figure 4 displays the calculated similarity compared with the rating from the expert survey. The figure illustrates a notable distinction between Euclidean and Gower similarity. The range of the Euclidean similarity is wider than the range of the Gower similarity. For instance, the Euclidean similarity in the base dataset has values ranging from 1.29 to 5.31. In contrast, the Gower similarity for the same dataset only varies from 0.74 to 3.11. We can observe similar results for the extended dataset, with the values of the Euclidean similarity ranging from 1.52 to 4.95 and the Gower similarity ranging from 0.95 to 2.95.



**Fig. 4** Violin plots showing the prediction distributions of the Euclidean and Gower similarity algorithms across the expert-rated similarity levels (1–5). Results are presented for both the base and extended datasets

Examining the correlation between prediction and rating reveals a positive correlation. This is supported by the Spearman rank correlation, presented in Table 2. All combinations exhibit a positive Spearman value ranging from 0.321 for the extended dataset with Euclidean similarity to 0.498 for the base dataset with Euclidean similarity. When comparing the base dataset with the extended dataset, both

similarity algorithms outperform across all metrics on the base dataset. Although the Gower similarity algorithm achieves a slightly higher  $R^2$  and lower MSE score than the Euclidean algorithm, the Spearman rank correlation is higher for the Euclidean similarity.

Thus, the distance-based similarity algorithms enable a general prediction of similarity. Nevertheless, the algorithms

**Table 2** Performance metrics of the similarity algorithms across the base and extended datasets. Reported are mean squared error (MSE),  $R^2$ , and Spearman's rank correlation between predicted and expert-rated similarity. Best results per metric are highlighted

	Euclidean base	Euc. extended	Gower base	Gower extended
MSE	0.724	0.820	0.719	0.797
$R^2$	0.20	0.095	0.206	0.120
Spearman	0.498	0.321	0.455	0.351



can only predict tendencies rather than exact ratings, as indicated by the relatively low  $R^2$  results.

## Machine learning models

We split the dataset into training data (857 route pairs) and test data (368 route pairs). To ensure comparability, the same route pairs were consistently assigned to either the training or test dataset for both the base and extended data. Subsequently, we describe only the best-performing models for each algorithm.

The evaluations of the different machine learning models, illustrated in Table 3, indicate that random forest regression on the extended dataset achieves the lowest MSE (0.53) and highest  $R^2$  (0.448) results. The highest Spearman value is reached with SVR (0.699). The linear regression, a more basic algorithm, performs slightly worse than SVR, XGBoost, and random forest regression; however, it outperforms the decision tree regression. Furthermore, while XGBoost performs slightly worse than random forest regression and SVR on the extended dataset, its model trained on the base dataset surpasses the other algorithms trained on the base dataset.

Comparing the models' performance between the base and extended dataset, the models trained on the extended dataset either perform better than or equal to the models trained solely on the base dataset. The mean value of the  $R^2$  results for the base dataset is 0.398 (MSE 0.577), while for the extended dataset, it is 0.414 (MSE 0.561). Regarding the Spearman correlation, the algorithms trained on the base dataset achieve a mean value of 0.652, compared to 0.669 for the extended dataset.

Additionally, we observe that certain algorithms, such as random forest regression and linear regression, utilize all 10 features in the base dataset and all 30 features in the extended dataset for prediction. In contrast, other algorithms only use a fraction of the available features, e.g., the decision tree regression uses six out of ten and six out of thirty features for the base and extended dataset, respectively. The optimal performance of the decision tree algorithm is achieved with a maximum tree depth of four. Increasing the depth of the decision tree does not enhance these results; in fact, it may lead to overfitting as the algorithm incorporates

more features. Conversely, a shallower tree may not capture the necessary information.

The utilization of SHAP values allows a closer examination of the various features. The SHAP summary plot provides feature importance via top-down ranking. Furthermore, each dot on the graph represents one distance value for a route combination. The horizontal position of the dot indicates whether the feature value has a negative (left side) or a positive (right side) impact on the model output. An accumulation of dots represents many similar rows with a similar impact on the model output. The color of each dot indicates if the feature value is low or high. Specifically, blue dots denote a low feature distance; ergo, a similar route combination for this feature. Conversely, red dots indicate a large feature distance, such as comparing the *length* feature between a short and a long route. Furthermore, the correlation is represented by the combination of color and horizontal position. A high value (red color) with a negative impact (left side) signifies a negative correlation between the feature and the target value, and vice versa. In Fig. 5, we present the SHAP summary plot for the random forest regression, the best-performing model, trained and evaluated on the extended dataset. All features with "tag" in their names exclusively belong to the extended dataset, while all other features are part of both datasets. We focus on the analysis of the twenty most important features.

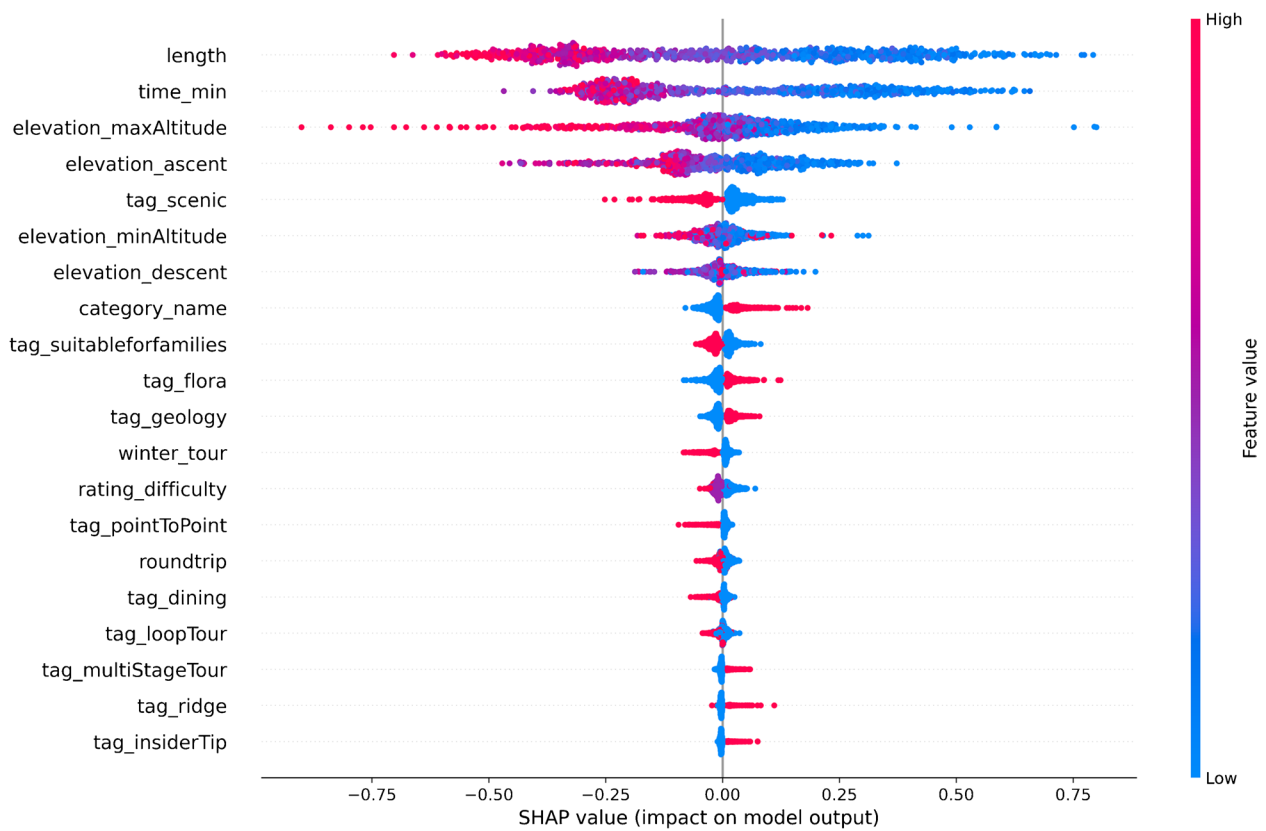
Similar to the feature importance derived from the models, SHAP values also identify *length* as the most important feature, followed by the *duration*, *maximum elevation*, and *total ascent*. Due to the multicollinearity between *length* and *duration*, the actual importance of these features might vary. The most important additional feature from the extended dataset is *scenic*, followed by the *suitability of a route for families*. Interestingly, despite the 15 least important features being tags (five depicted, ten not depicted), models trained on the extended dataset outperform those on the base dataset.

Examining the distribution of features impact, we observe that the *length* values are mostly evenly distributed, although slightly skewed to the right. The *time* feature exhibits a high concentration of values with a negative impact of approximately  $-0.3$ . In contrast, values with a positive impact do not display a similar accumulation. The *maximum altitude*

**Table 3** Performance comparison of the five machine learning models on the base and extended datasets. Metrics include mean squared error (MSE),  $R^2$ , and Spearman's rank correlation with expert ratings. Best results per metric and dataset are highlighted

Dataset	Linear regression		SVR		Decision tree		Random forest		XGBoost	
	Base	Ext	Base	Ext	Base	Ext	Base	Ext	Base	Ext
MSE	0.589	0.582	0.567	0.533	0.623	0.624	0.561	0.530	0.546	0.539
$R^2$	0.385	0.393	0.409	0.444	0.350	0.348	0.415	0.448	0.430	0.437
Spearman	0.655	0.670	0.662	0.699	0.617	0.617	0.657	0.679	0.668	0.678





**Fig. 5** SHAP summary plot for the random forest regression model, showing the impact of each feature on the model output. Each point represents a SHAP value for an individual prediction, with color indicating the feature value (red = high; blue = low)

contains numerous values with minimal impact and several extreme outliers that notably influence the model output. For the remaining features, most of their values are distributed around the center and, therefore, do not greatly impact the model, except for some outliers.

In terms of correlation, we generally expect a negative correlation between the features and the target data, given that the features represent the distance between two routes. This holds true for most features, although with some exceptions: the values of the *elevation descent* feature lack clear separation, aligning with the linear regression coefficient results where *elevation descent* has a positive coefficient. The feature *category name* and the tags *flora*, *geology*, *multi-stage tour*, and *insider tip* all display a distinct separation between low and high values. However, contrary to our expectation, the predicted similarity increases with the rising distance between the routes for these features. The tag *ridge* follows a similar pattern, though the distinction is not as straightforward as in the features mentioned above, as some high values have a low negative impact.

Concluding, the analysis of SHAP values allows for an interpretation of how various features influence prediction performance. However, it is important to note that SHAP values describe the model's interpretation of features and

may not capture general relations, given that the random forest regression model does not achieve perfect results. This becomes evident as some features influence the model in the expected way, while others do the opposite.

### Comparison between distance-based similarity and machine learning algorithms

For the comparison, Table 4 displays the results of the best-performing models – those with the lowest MSE, highest  $R^2$ , and highest Spearman—separated for the base and the extended dataset.

The machine learning algorithms demonstrate superior performance compared to the similarity algorithms, particularly when trained on the extended dataset. There is a 4.2% increase in the  $R^2$  score from the base to the extended dataset. Among the machine learning algorithms, random forest regression performs the best, with SVR achieving lower  $R^2$  and MSE scores but better results for the Spearman correlation. Despite the differences in performance among machine learning algorithms, the results are closely clustered. For the base dataset, linear regression achieves an  $R^2$  score of 0.385, and XGBoost 0.43, reflecting an 11.7% increase. Conversely, for the distance-based similarity algorithms,

**Table 4** Comparison of the best-performing distance-based similarity algorithm and machine learning model on the base and extended datasets. Metrics include mean squared error (MSE),  $R^2$ , and Spearman's rank correlation with expert similarity ratings

	Similarity algorithm		Machine learning	
	Base	Extended	Base	Extended
<b>MSE</b>	0.719	0.797	0.546	0.53
<b><math>R^2</math></b>	0.206	0.12	0.43	0.448
<b>Spearman</b>	0.498	0.351	0.668	0.699

the base dataset outperforms the extended dataset, with a 72% higher  $R^2$  score for the smaller dataset. Nevertheless, even with this improvement, the similarity algorithm results are less than half as effective as the machine learning results when comparing the  $R^2$  score.

## Discussion

Regarding the research questions, our analysis demonstrates the feasibility of calculating the similarity of hiking routes using both similarity algorithms and machine learning models. Evaluating and discussing the models' performance involves considering expert ratings, route dataset, and the models themselves, including their features.

The experts likely incorporate descriptive features such as *length*, *duration*, and *altitude* in their rating process, as revealed by the correlation with basic distance-based similarity algorithms. However, there are discrepancies between the algorithm's results and those of the experts (Fig. 4), likely due to the subjective nature of the ratings. The experts were asked to rate the routes based on their personal experiences, which may have influenced their evaluations. Therefore, even with local experts conducting the ratings, factors such as duration or points of interest along the route may vary individually. Additionally, a majority of routes receive dissimilar ratings, with only a small percentage being rated as very alike (Fig. 2). Obtaining more data on similar routes could likely enhance the overall performance of machine learning models. Notably, in this research, predictions for higher-rated similarities tend to be underestimated.

Besides the expert ratings, the raw tour data represents the second important data source. To ensure data quality, this research implements a rigorous data cleaning process. However, it is essential to note that the routes are primarily uploaded by individuals to the Outdooractive platform and are not formally reviewed by the platform. Consequently, the accuracy of the provided route information cannot be entirely guaranteed. To address this concern, we mitigate the issue by selecting a subset of 50 different routes, which are then carefully evaluated for plausibility. When extending

our approach to other regions and contexts, we would recommend utilizing a valid and reviewed source of data to reduce the effort required for data cleaning.

Turning to the features, we identified the most important features and their impact on predictions. However, the SHAP value analysis (Fig. 5) only refers to features present in the model, leaving room for potential unaccounted features. To assess the potential impact of additional features, we compared model and algorithm performance with both base and extended dataset. On the one hand, the results of the distance-based similarity algorithms on unweighted features indicate that more features do not necessarily improve results (Table 2). The lack of weights leads to a decreased influence of crucial features in relation to all features. Although the extended dataset offers more information, this information is not successfully used to convey knowledge. On the other hand, machine learning algorithms, capable of assigning weights or determining feature influence, effectively use additional information from the extended dataset without diminishing the importance of crucial features (Table 3). Overall, machine learning models outperform distance-based similarity algorithms, with the random forest regression and SVR achieving the best results. Despite slight differences, even basic models like linear regression produce satisfactory results ( $R^2$  scores ranging from 0.348 for the base to 0.448 for the extended dataset). Given the similarity in outcomes, deploying a fast and computationally efficient model like the linear regression, including all available features, may be practical in real-world applications at a slight expense of accuracy.

This study extends beyond mere application towards a theoretical contribution, offering significant value to the smart tourism community. By analyzing and comparing various algorithms for distance-based similarity alongside machine learning approaches, we not only aim to pave the way for more sustainable practices in recreational tourism but also enrich the toolkit available to researchers and practitioners for recommendation systems in diverse domains. Considering the first research question, we evaluated how accurately the multiple approaches calculate the similarity and which model performance surpasses the others. In relation to the second research question, we offer insights into the role of dataset size and conduct a comprehensive analysis of feature relevance using SHAP values. Additionally, our approach, which is evaluated on a real-life data basis by using not publicly available company data, can enhance existing work analyzing routes: for example, while Calbimonte et al. (2020) focuses on the suitability of a route for a hiker, a combined approach with our work allows the personalized recommendation of suitable alternative routes which are less overcrowded. Moreover, most papers analyze the similarities in POI (Adams & Raubal, 2014; Cao & Thomas, 2021;

Preoțiu-Pietro et al., 2013; Qiu et al., 2021; Zhao et al., 2019), whereas our approach allows better recommendations for cases where the journey itself is the destination.

Contributing to practice, these insights enable the implementation of a recommendation system fit for the interest of users and to prevent crowding. Contrary to many other recommendation systems, this approach can suggest routes outside of a crowded region based on inherent route similarity, supporting a more even distribution of visitors. Building upon existing initial efforts, nature parks can directly cooperate with outdoor platforms to redirect visitor flows. As our approach is developed and tested using real-world company data, it demonstrates not only theoretical relevance but also practical applicability in a specific case. Moreover, combining route similarity with individual user preferences could further enhance the relevance and value of recommended alternatives for tourists.

While our case study focuses on one region, the methodology can be easily extended to other regions with similar characteristics, i.e., for activities such as hiking in mountainous terrains. In general, the findings of this work focus on tourist regions in the westernized world and on the characteristics of middle European landscapes but can be adapted to other countries as well. Given that the features and their importance remain the same for hiking activities, the trained machine learning model can directly be applied to new route data. For regions with distinct characteristics, such as flatter terrain, adjustments to the models and a new training dataset may be necessary, emphasizing factors like *length* or *duration* over *elevation*. Additionally, the same concept can be applied to other outdoor recreational activities on foot, such as jogging and Nordic walking. Furthermore, the same approach could be applied to other sports following a route, such as biking, cross-country skiing, canoeing, or climbing. To conclude, our approach not only provides the necessary information to propose recommendations aimed at protecting the environment, increasing the acceptance of tourism, and promoting sustainable tourism but also offers a theoretical contribution to the IS community by enhancing the understanding and application of complex similarity concepts in diverse settings.

## Conclusion, limitations, and future research

In this paper, we present a novel method for determining route similarity based on descriptive features and analyzed the performance of various similarity algorithms and machine learning models. Our primary goal is to establish a foundation for developing a recommendation system to facilitate visitor management and prevent crowding. Such

a recommendation system requires route suggestions based on similar descriptive features, rather than relying solely on geographical proximity. To validate our approach, we conducted a case study involving the analysis and preparation of hiking route data from the Outdooractive database. We curated a sample dataset comprising fifty hiking routes within a nature park, whose similarities of route combinations have been evaluated by experts. This rating served as a benchmark to assess the performance of the distance-based similarity algorithms and to train and evaluate the machine learning models. Furthermore, for the best-performing machine learning model, we identified the most important features and how they influence the similarity prediction with a SHAP-value analysis.

The results of this research reveal that the calculated similarities of the distance-based algorithms exhibit a moderate correlation with the expert ratings. Interestingly, an increase in the number of features does not necessarily lead to improved results. The unweighted nature of the algorithms implies equal importance for all features, causing more influential features to diminish in significance with the addition of others. When utilizing labeled data, the training of machine learning algorithms proves to be advantageous over basic algorithms. Unlike distance-based similarity algorithms, incorporating additional features enhances the model's informational depth by allowing it to assign weights to the feature importance. Among the models assessed, the random forest regression demonstrates superior performance, closely followed by the support vector regression, XGBoost, and linear regression. Furthermore, the *length* of the route is identified as the most crucial feature for the similarity of hiking routes.

This paper introduces a methodologically innovative approach to assessing route similarity that advances the IS discipline by leveraging descriptive features over geographical proximity, offering a new perspective on data analysis and application. The insights gained from comparing machine learning models with traditional algorithms enhance our understanding of decision-making processes in IS, highlighting the benefits of machine learning models trained on labeled data.

However, this paper is also subject to some limitations that should be addressed in future research. While route similarity was rated by experts, there may still be individual influencing factors that vary across different groups. Future studies could therefore investigate how similarity ratings differ depending on user background, preferences, or demographic characteristics, allowing for a more nuanced understanding of perceived similarity. Another limitation lies in the granularity of our dataset, particularly with regard to tag features and POI-related information. For instance, while the binary tag for *dining* indicates the presence of food outlets, it does not reflect quantifiable aspects such as

the number or quality of these facilities. Moreover, while detailed POI data is not available, this omission may be of limited consequence given that all routes are located within the same nature park and share similar, appealing natural features. Nevertheless, incorporating more detailed data, such as the number of dining options, the mean or median altitude, the maximum incline, additional POI data, and other amenities along the trail, could enhance similarity predictions, as these aspects may be more relevant to some visitors than purely quantitative metrics like route *length* or *duration*. These features may be especially relevant for visitors with specific needs or preferences, and thus contribute to more personalized recommendations. Furthermore, while we implemented various data quality checks, the data is created by individuals and might contain incomplete or inconsistently tagged information. Another limitation is the focus of the study on a single region and a specific type of outdoor activity. Exploring the applicability of the proposed methodology in diverse regions, different route types (e.g., urban walks, bike trails), or various sports activities could provide valuable insights.

Looking ahead, future research should explore how expanding the dataset can further improve the robustness of similarity predictions. Incorporating user preferences, including those of specific demographic segments, will be essential for enhancing the relevance and acceptance of recommendations. In addition, investigating the role of AI-based travel assistants may offer new avenues for dynamic and context-aware route suggestions. Moreover, considering the spatial distance between users' residences and suggested routes would increase the applicability of recommendations in real-world planning. Finally, the deployment and evaluation of additional machine learning models may yield performance gains and support further methodological advancement. Beyond improving the models for route similarity, the findings of this work should be used to develop a recommendation system for outdoor websites to enable sustainable growth in tourism in the long term.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s12525-025-00816-5>.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Data availability** The route data used in the case study is not publicly available, but the routes can be individually viewed on the Outdooractive platform (available at <https://www.outdooractive.com/>). The code that supports the findings of this study is available from the corresponding author upon reasonable request.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes

were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Adams, B., & Raubal, M. (2014). Identifying salient topics for personalized place similarity. In S. Winter & C. Rizos (Eds.), *Proceedings of Research at Locate'14, CEUR Workshop Proceedings* (Vol. 1142, pp. 1–12). Canberra. <https://ceur-ws.org/Vol-1142/paper1.pdf>. Accessed 19 Sep 2024.
- Adie, B. A., Falk, M., & Savioli, M. (2020). Overtourism as a perceived threat to cultural heritage in Europe. *Current Issues in Tourism*, 23(14), 1737–1741. <https://doi.org/10.1080/13683500.2019.1687661>
- Ahdid, R., Taifi, K., Safi, S., & Manaut, B. (2016). Euclidean & geodesic distance between a facial feature points in two-dimensional face recognition system. In *ACIT'2016*. <https://www.acit2k.org/acit/images/stories/year2014/month1/proceeding/37.pdf>. Accessed 15 Feb 2024.
- Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3), 370–374. <https://doi.org/10.1002/wics.84>
- Alonso-Almeida, M. D., Borrajo-Millán, F., & Yi, L. (2019). Are social media data pushing overtourism? The case of Barcelona and Chinese tourists. *Sustainability*, 11(12), 3356. <https://doi.org/10.3390/su11123356>
- Anagnostopoulos, A., Atassi, R., Becchetti, L., Fazzzone, A., & Silvestri, F. (2017). Tour recommendation for groups. *Data Mining and Knowledge Discovery*, 31(5), 1157–1188. <https://doi.org/10.1007/s10618-016-0477-7>
- Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2012). Effects of user similarity in social media. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM. <https://doi.org/10.1145/2124295.2124378>
- Bauer, K., Hinz, O., van der Aalst, W., & Weinhardt, C. (2021). Expl(AI)n it to me – Explainable AI and information systems research. *Business & Information Systems Engineering*, 63(2), 79–82. <https://doi.org/10.1007/s12599-021-00683-2>
- Becker, H., Naaman, M., & Gravano, L. (2010). Learning similarity metrics for event identification in social media. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM. <https://doi.org/10.1145/1718487.1718524>
- Benitez, J. M., Castro, J. L., & Requena, I. (1997). Are artificial neural networks black boxes? *IEEE Transactions on Neural Networks*, 8(5), 1156–1164. <https://doi.org/10.1109/72.623216>
- Benner, M. (2020). The decline of tourist destinations: An evolutionary perspective on overtourism. *Sustainability*, 12(9), Article 3653. <https://doi.org/10.3390/su12093653>
- Bergfex. (2023). *BERGFEX- Deutschland - Wandern & mountainbiken*. <https://www.bergfex.de/sommer/deutschland/touren/>. Accessed 24 Aug 2023.
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281–305. <https://doi.org/10.5555/2188385.2188395>
- Bertocchi, D., Camatti, N., Giove, S., & van der Borg, J. (2020). Venice and overtourism: Simulating sustainable development scenarios through a tourism carrying capacity model. *Sustainability*, 12(2), Article 512. <https://doi.org/10.3390/su12020512>



- Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2), 197–227. <https://doi.org/10.1007/s11749-016-0481-7>
- Bichler, B. F., & Peters, M. (2021). Soft adventure motivation: An exploratory study of hiking tourism. *Tourism Review*, 76(2), 473–488. <https://doi.org/10.1108/TR-10-2019-0403>
- Boháč, A., & Drápela, E. (2022). Overtourism hotspots: Both a threat and opportunity for rural tourism. *European Countryside*, 14(1), 157–179. <https://doi.org/10.2478/euco-2022-0009>
- Bollenbach, J., Neubig, S., Hein, A., Keller, R., & Krcmar, H. (2022). Using machine learning to predict POI occupancy to reduce overcrowding. In D. Demmler, D. Krupka, & H. Federath (Eds.), *Lecture Notes in Informatics (LNI) 2022* (P-326, pp. 393–408). Köllen Druck+Verlag GmbH. [https://doi.org/10.18420/inf2022\\_34](https://doi.org/10.18420/inf2022_34)
- Butler, R. W., & Dodds, R. (2022). Overcoming overtourism: A review of failure. *Tourism Review*, 77(1), 35–53. <https://doi.org/10.1108/TR-04-2021-0215>
- Calbimonte, J.-P., Martin, S., Calvaresi, D., Zappelaz, N., & Cotting, A. (2020). Semantic data models for hiking trail difficulty assessment. In J. Neidhardt & W. Wörndl (Eds.), *Information and communication technologies in tourism 2020: Proceedings of the International Conference in Surrey, United Kingdom, January 08–10, 2020 / Julia Neidhardt, Wolfgang Wörndl, editors* (pp. 295–306). Springer. [https://doi.org/10.1007/978-3-030-36737-4\\_24](https://doi.org/10.1007/978-3-030-36737-4_24)
- Calbimonte, J.-P., Martin, S., Calvaresi, D., & Cotting, A. (2021). A platform for difficulty assessment and recommendation of hiking trails. In W. Wörndl, C. Koo, & J. L. Stienmetz (Eds.), *Information and Communication Technologies in Tourism 2021: Proceedings of the ENTER 2021 eTourism Conference, January 19–22, 2021 / Wolfgang Wörndl, Chulmo Koo, Jason L. Stienmetz, editors* (pp. 109–122). Springer. [https://doi.org/10.1007/978-3-030-65785-7\\_9](https://doi.org/10.1007/978-3-030-65785-7_9)
- Candocia, M. (2020). *Calculating similarity of running routes*. <https://maxcandocia.com/article/2020/Sep/13/run-path-similarity/>. Accessed 24 May 2022.
- Cao, H., & Thomas, E. (2021). *Destination similarity based on implicit user interest*. <https://doi.org/10.48550/arXiv.2102.06687>
- Capocchi, A., Vallone, C., Pierotti, M., & Amaduzzi, A. (2019). Overtourism: A literature review to assess implications and future perspectives. *Sustainability*, 11(12), Article 3303. <https://doi.org/10.3390/su11123303>
- Carvalhinho, L., Rosa, P., & Gomes, F. (2015). Hiking trails evaluation in the natural park of Serras de Aire e Candeeiros, Portugal. *European Journal of Tourism, Hospitality and Recreation*, 6(2), 139–156. [https://www.researchgate.net/publication/282012572\\_HIKING\\_TRAILS\\_EVALUATION\\_IN\\_THE\\_NATURAL\\_PARK\\_OF\\_SERRAS\\_DE\\_AIRE\\_E\\_CANDEEIROS\\_PORTUGAL](https://www.researchgate.net/publication/282012572_HIKING_TRAILS_EVALUATION_IN_THE_NATURAL_PARK_OF_SERRAS_DE_AIRE_E_CANDEEIROS_PORTUGAL). Accessed 30 June 2022.
- Chen, M.-C., Chiu, A.-L., & Chang, H.-H. (2005). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, 28(4), 773–781. <https://doi.org/10.1016/j.eswa.2004.12.033>
- Chen, J., Huang, L., Wang, C., & Zheng, N. (2020). Discovering travel spatiotemporal pattern based on sequential events similarity. *Complexity*, 2020, 1–10. <https://doi.org/10.1155/2020/6632956>
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. <https://doi.org/10.1145/2939672.2939785>
- Chhetri, P., Arrowsmith, C., & Jackson, M. (2004). Determining hiking experiences in nature-based tourist destinations. *Tourism Management*, 25(1), 31–43. [https://doi.org/10.1016/s0261-5177\(03\)00057-8](https://doi.org/10.1016/s0261-5177(03)00057-8)
- Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7, e623. <https://doi.org/10.7717/peerj-cs.623>
- Chin, W. W. (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.), *Quantitative methodology series. Modern methods for business research* (pp. 295–336). Psychology Press. <https://psycnet.apa.org/record/1998-07269-010>. Accessed 20 Sept 2022.
- Choi, J., Hui, S. K., & Bell, D. R. (2010). Spatiotemporal analysis of imitation behavior across new buyers at an online grocery retailer. *Journal of Marketing Research*, 47(1), 75–89. <https://doi.org/10.1509/jmkr.47.1.75>
- Corporate Outdooractive. (2023). *Platform*. Corporate Outdooractive. <https://corporate.outdooractive.com/en/platform/>. Accessed 20 June 2022.
- Creany, N., Monz, C. A., & Esser, S. M. (2024). Understanding visitor attitudes towards the timed-entry reservation system in Rocky Mountain National Park: Contemporary managed access as a social-ecological system. *Journal of Outdoor Recreation and Tourism*, 45, Article 100736. <https://doi.org/10.1016/j.jort.2024.100736>
- Cui, G., Luo, J., & Wang, X. (2018). Personalized travel route recommendation using collaborative filtering based on GPS trajectories. *International Journal of Digital Earth*, 11(3), 284–307. <https://doi.org/10.1080/17538947.2017.1326535>
- D'souza, R. N., Huang, P.-Y., & Yeh, F.-C. (2020). Structural analysis and optimization of convolutional neural networks with a small sample size. *Scientific Reports*, 10(1), 834. <https://doi.org/10.1038/s41598-020-57866-2>
- de Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1), 1–18. [https://doi.org/10.1016/S0169-7439\(99\)00047-7](https://doi.org/10.1016/S0169-7439(99)00047-7)
- de Oliveira e Silva, R. A., Cui, G., Rahimi, S. M., & Wang, X. (2022). Personalized route recommendation through historical travel behavior analysis. *GeoInformatica*, 26(3), 505–540. <https://doi.org/10.1007/s10707-021-00453-y>
- de Raadt, A., Warrens, M. J., Bosker, R. J., & Kiers, H. A. L. (2021). A comparison of reliability coefficients for ordinal rating scales. *Journal of Classification*, 38(3), 519–543. <https://doi.org/10.1007/s00357-021-09386-5>
- Dodds, R., & Butler, R. (2019). The phenomena of overtourism: A review. *International Journal of Tourism Cities*, 5(4), 519–528. <https://doi.org/10.1108/IJTC-06-2019-0090>
- dos Santos, T. R., & Zárate, L. E. (2015). Categorical data clustering: What similarity measure to recommend? *Expert Systems with Applications*, 42(3), 1247–1260. <https://doi.org/10.1016/j.eswa.2014.09.012>
- Drápela, E. (2023). Geoheritage and overtourism: A case study from sandstone rock cities in the Czech Republic. *Geological Society, London, Special Publications*, 530(1), 257–275. <https://doi.org/10.1144/SP530-2022-102>
- Drápela, E., Zágorský, K., Boháč, A., & Böhm, H. (2021). Rural overtourism: A typology of negative effects. In *PUBLIC RECREATION and LANDSCAPE PROTECTION – with sense hand in hand!*. [https://www.researchgate.net/profile/artur-bohac/publication/354781623\\_rural\\_overtourism\\_a\\_typology\\_of\\_negative\\_effects](https://www.researchgate.net/profile/artur-bohac/publication/354781623_rural_overtourism_a_typology_of_negative_effects). Accessed 12 Aug 2024.
- Ekström, J. (2011). The phi-coefficient, the tetrachoric correlation coefficient, and the Pearson-Yule debate. *UCLA: Department of Statistics, UCLA*. <https://escholarship.org/uc/item/7qp4604r>. Accessed 10 Aug 2024.
- Enseñat-Soberanis, F., Frausto-Martínez, O., & Gándara-Vázquez, M. (2019). A visitor flow management process for touristified archaeological sites. *Journal of Heritage Tourism*, 14(4), 340–357. <https://doi.org/10.1080/1743873X.2018.1529179>



- Ester, M., Kriegel, H.-P., Sander, J., & Xu, X [Xiaowei]. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise*. <https://doi.org/10.5555/3001460.3001507>. Accessed 29 Aug 2022.
- Falk, M., Hagsten, E., & Lin, X. (2022). Importance of land characteristics for resilience of domestic tourism demand. *Tourism Recreation Research*. <https://doi.org/10.1080/02508281.2022.2116541>
- Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1), 3133–3181. <https://doi.org/10.5555/2627435.2697065>
- Fillisch, J., Jordan, L., Nasebandt-Schwarz, L., & Neumeyer, E. (2020). *Branchenbefragung wandertourismus - Corona 2020*. [https://www.wanderbares-deutschland.de/\\_Resources/Persistent/33eef a583d4e927b6f4804d40dd19d182d30cbe6/Branchenbefragung%20Wandertourismus%20-%20Corona%202020.pdf](https://www.wanderbares-deutschland.de/_Resources/Persistent/33eef a583d4e927b6f4804d40dd19d182d30cbe6/Branchenbefragung%20Wandertourismus%20-%20Corona%202020.pdf). Accessed 27 Sept 2022.
- Frost, R. (1916). *Mountain interval*. Holt.
- Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. ScholarWorks@UETP. [https://scholarworks.utep.edu/cs\\_techrep/1209/](https://scholarworks.utep.edu/cs_techrep/1209/). Accessed 10 Sep 2024.
- Gillbanks, D. (2023). Prof Richard Butler on tourism's challenges and academia's inadequacies - The "good. The "Good Tourism" Blog. <https://www.goodtourismblog.com/2023/12/professor-richard-butler/>. Accessed 1 Apr 2025.
- Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27(4), 857. <https://doi.org/10.2307/2528823>
- Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: Foundations and developments. *Electronic Markets*, 25(3), 179–188. <https://doi.org/10.1007/s12525-015-0196-8>
- Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1), 1–12. <https://doi.org/10.1021/ci0342472>
- Heaton, J. (2016). An empirical analysis of feature engineering for predictive modeling. In *SoutheastCon 2016*. IEEE. <https://doi.org/10.1109/secon.2016.7506650>
- Henkel, C., Seidler, A.-R., Kranz, J., & Fiedler, M. (2019). How to nudge pro-environmental behaviour: An experimental study. In *Proceedings of the 27th European Conference on Information Systems (ECIS)*. Association for Information Systems.
- Hwangbo, H., Kim, Y. S., & Cha, K. J. (2018). Recommendation system development for fashion retail e-commerce. *Electronic Commerce Research and Applications*, 28, 94–101. <https://doi.org/10.1016/j.elerap.2018.01.012>
- Ifenthaler, D. (2012). Measures of Similarity. In N. M. Seel (Ed.), *Springer reference. Encyclopedia of the sciences of learning* (pp. 2147–2150). Springer. [https://doi.org/10.1007/978-1-4419-1428-6\\_503](https://doi.org/10.1007/978-1-4419-1428-6_503)
- Insch, A. (2020). The challenges of over-tourism facing New Zealand: Risks and responses. *Journal of Destination Marketing & Management*, 15, 100378. <https://doi.org/10.1016/j.jdmm.2019.100378>
- Issa, H., Guirguis, A., Beshara, S., Agne, S., & Dengel, A. (2016). Preference based filtering and recommendations for running routes. In *Proceedings of the 12th International Conference on Web Information Systems and Technologies*. SCITEPRESS - Science and Technology Publications. <https://doi.org/10.5220/0005897801390146>
- Javed, F., Chan, G., Savkin, A. V., Middleton, P. M., Malouf, P., Steel, E., Mackie, J., & Lovell, N. H. (2009). RBF kernel based support vector regression to estimate the blood volume and heart rate responses during hemodialysis. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. <https://doi.org/10.1109/iembs.2009.5332739>
- Jin, B., & Xu, X. [Xiaojie]. (2024a). Forecasting wholesale prices of yellow corn through the Gaussian process regression. *Neural Computing and Applications*, 36(15), 8693–8710. <https://doi.org/10.1007/s00521-024-09531-2>
- Jin, B., & Xu, X. [Xiaojie]. (2024b). Predictions of steel price indices through machine learning for the regional northeast Chinese market. *Neural Computing and Applications*, 36(33), 20863–20882. <https://doi.org/10.1007/s00521-024-10270-7>
- Jin, B., Xu, X., [Xiaojie], & Zhang, Y. (2025). Thermal coal futures trading volume predictions through the neural network. *Journal of Modelling in Management*, 20(2), 585–619. <https://doi.org/10.1108/JM2-09-2023-0207>
- Jin, B., & Xu, X [Xiaojie] (2024c). Wholesale price forecasts of green grams using the neural network. *Asian Journal of Economics and Banking*. Advance online publication. <https://doi.org/10.1108/AJEB-01-2024-0007>
- Koens, K., Postma, A., & Papp, B. (2018). Is overtourism overused? Understanding the impact of tourism in a city context. *Sustainability*, 10(12), Article 4384. <https://doi.org/10.3390/su10124384>
- Komoot. (2021). *Tourenempfehlungen von komoot (Smart Tours)*. <https://support.komoot.com/hc/de/articles/360058879211>. Accessed 17 Aug 2022.
- Kranz, J., Kolbe, L. M., Koo, C., & Boudreau, M.-C. (2015). Smart energy: Where do we stand and where should we go? *Electronic Markets*, 25(1), 7–16. <https://doi.org/10.1007/s12525-015-0180-3>
- Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. <http://www.featurengineering/index.html>. Accessed 14 Sept 2022.
- Kühne, B., & Heidel, K. (2021). How could smart cities use data? – Towards a taxonomy of data-driven smart city projects. In *Innovation through information systems. Volume I : a collection of latest research on domain issues* (pp. 351–366). Springer. [https://doi.org/10.1007/978-3-030-86790-4\\_24](https://doi.org/10.1007/978-3-030-86790-4_24)
- Kumar, S., & Nancy (2014). Efficient K-mean clustering algorithm for large datasets using data mining standard score normalization. *International Journal IJRITCC*, 2(10). <https://doi.org/10.17762/ijritcc.v2i10.3368>
- Kuo, I.-L. (2002). The effectiveness of environmental interpretation at resource-sensitive tourism destinations. *International Journal of Tourism Research*, 4(2), 87–101. <https://doi.org/10.1002/jtr.362>
- Kuok, S.-C., Yao, S.-A., Yuen, K.-V., Yan, W.-J., & Girolami, M. (2025). Bayesian generative kernel Gaussian process regression. *Mechanical Systems and Signal Processing*, 227, Article 112395. <https://doi.org/10.1016/j.ymssp.2025.112395>
- Lamsfus, C., Martín, D., Alzua-Sorzabal, A., & Torres-Manzanera, E. (2015). Smart tourism destinations: An extended conception of smart cities focusing on human mobility. In I. Tussyadiah & A. Inversini (Eds.), *Information and Communication Technologies in Tourism 2015: Proceedings of the International Conference in Lugano, Switzerland, February 3 - 6, 2015* (1st ed. 2015, pp. 363–375). Springer International Publishing; Imprint: Springer. [https://doi.org/10.1007/978-3-319-14343-9\\_27](https://doi.org/10.1007/978-3-319-14343-9_27)
- Larson, L. R., & Poudyal, N. C. (2012). Developing sustainable tourism through adaptive resource management: A case study of Machu Picchu, Peru. *Journal of Sustainable Tourism*, 20(7), 917–938. <https://doi.org/10.1080/09669582.2012.667217>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Lehnhoff, S., Staudt, P., & Watson, R. T. (2021). Changing the climate in information systems research. *Business & Information*

- Systems Engineering*, 63(3), 219–222. <https://doi.org/10.1007/s12599-021-00695-y>
- Lesot, M. J., Rifqi, M., & Benhadda, H. (2009). Similarity measures for binary and numerical data: A survey. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 1(1), Article 21985, 63. <https://doi.org/10.1504/IJKESDP.2009.021985>
- Li, Y.-M., Wu, C.-T., & Lai, C.-Y. (2013). A social recommender mechanism for e-commerce: Combining similarity, trust, and relationship. *Decision Support Systems*, 55(3), 740–752. <https://doi.org/10.1016/j.dss.2013.02.009>
- Liang, Y., Li, S., Yan, C., Li, M., & Jiang, C. (2021). Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing*, 419, 168–182. <https://doi.org/10.1016/j.neucom.2020.08.011>
- Liashchynskyi, P [Petro], & Liashchynskyi, P [Pavlo] (2019). Grid search, random search, genetic algorithm: A big comparison for NAS. <https://doi.org/10.48550/arXiv.1912.06059>
- Lin, D. (1998). An information-theoretic definition of similarity. *Icml*, 98, 296–304. <https://doi.org/10.5555/645527.657297>
- Liu, C.-H., Tsai, C.-F., Sue, K.-L., & Huang, M.-W. (2020). The feature selection effect on missing value imputation of medical datasets. *Applied Sciences*, 10(7), Article 2344. <https://doi.org/10.3390/app10072344>
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*. <https://doi.org/10.1109/ICDM.2008.17>
- Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003). Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation. *Bioinformatics*, 19(10), 1275–1283. <https://doi.org/10.1093/bioinformatics/btg153>
- Lundberg, S., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. <https://doi.org/10.48550/arXiv.1705.07874>
- Maggiore, G., Vogt, M., Stumpfe, D., & Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8), 3186–3204. <https://doi.org/10.1021/jm401411z>
- Maldonado, A. G., Doucet, J. P., Petitjean, M., & Fan, B.-T. (2006). Molecular similarity and diversity in chemoinformatics: From theory to applications. *Molecular Diversity*, 10(1), 39–79. <https://doi.org/10.1007/s11030-006-8697-1>
- Maleki Shoja, B., & Tabrizi, N. (2019). Customer reviews analysis with deep neural networks for e-commerce recommender systems. *IEEE Access*, 7, 119121–119130. <https://doi.org/10.1109/access.2019.2937518>
- Martin, M. (2002). On-line support vector machine regression. In T. Elomaa, H. Mannila, & H. Toivonen (Eds.), *13th European Conference on Machine Learning* (pp. 282–294). Springer, Berlin, Heidelberg. [https://doi.org/10.1007/3-540-36755-1\\_24](https://doi.org/10.1007/3-540-36755-1_24)
- Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8), 3048–3061. <https://doi.org/10.1109/tkde.2019.2962680>
- Mason, P. (2005). Visitor management in protected areas: From ‘hard’ to ‘soft’ approaches? *Current Issues in Tourism*, 8(2–3), 181–194. <https://doi.org/10.1080/13683500508668213>
- Megerle, H. E. (2021). Calcerous tufa as invaluable geotopes endangered by (over-)tourism: A case study in the UNESCO Global Geopark Swabian Alb, Germany. *Geosciences*, 11(5), Article 198. <https://doi.org/10.3390/geosciences11050198>
- Mihalic, T. (2020). Conceptualising overtourism: A sustainability approach. *Annals of Tourism Research*, 84, Article 103025. <https://doi.org/10.1016/j.annals.2020.103025>
- Milano, C., Novelli, M., & Cheer, J. M. (Eds.). (2021). *Travel and tourism in the age of overtourism*, 1. Routledge. <https://doi.org/10.4324/9781003140610>
- Molnár, A. J. (2020). TRAILSIGNER: A conceptual model of hiking trail networks with consistent signage planning and management. In M. Tropmann-Frick, B. Thalheim, H. Jaakkola, Y. Kiyoki, & N. Yoshida (Eds.), *Frontiers in artificial intelligence and applications*. Information Modelling and Knowledge Bases XXXII: IOS Press. <https://doi.org/10.3233/FAIA200817>
- Molokáč, M., Hlaváčová, J., Tometzová, D., & Liptáková, E. (2022). The preference analysis for hikers’ choice of hiking trail. *Sustainability*, 14(11), Article 6795. <https://doi.org/10.3390/su14116795>
- Mou, N., Jiang, Q., Zhang, L., Niu, J., Zheng, Y., Wang, Y., & Yang, T. (2022). Personalized tourist route recommendation model with a trajectory understanding via neural networks. *International Journal of Digital Earth*, 15(1), 1738–1759. <https://doi.org/10.1080/17538947.2022.2130456>
- Needham, M. (2018). *Strava: Calculating the similarity of two runs*. <https://www.markhneedham.com/blog/2018/01/18/strava-calculating-similarity-two-runs/>. Accessed 17 Aug 2022.
- Nordbø, I., & Prebensen, N. K. (2015). Hiking as mental and physical experience. In I. Nordbø & N. K. Prebensen (Eds.), *Advances in Hospitality and Leisure* (Vol. 11, pp. 169–186). Emerald Group Publishing Limited. <https://doi.org/10.1108/S1745-3542201500011010>
- Oklevik, O., Gössling, S., Hall, C. M., Steen Jacobsen, J. K., Grøtte, I. P., & McCabe, S. (2019). Overtourism, optimisation, and destination performance indicators: A case study of activities in Fjord Norway. *Journal of Sustainable Tourism*, 27(12), 1804–1824. <https://doi.org/10.1080/09669582.2018.1533020>
- Orams, M. B. (1996). A conceptual model of tourist-wildlife interaction: The case for education as a management strategy. *Australian Geographer*, 27(1), 39–51. <https://doi.org/10.1080/00049189608703156>
- Outdooractive. (2023). *Routes*. <https://www.outdooractive.com/en/routes/>. Accessed 24 Aug 2023.
- Palacios-Florencio, B., Santos-Roldán, L., Berbel-Pineda, J. M., & Castillo-Canalejo, A. M. (2021). Sustainable tourism as a driving force of the tourism industry in a post-COVID-19 scenario. *Social Indicators Research*, 158(3), 991–1011. <https://doi.org/10.1007/s11205-021-02735-2>
- Peeters, P., Gössling, S., Klijs, J., Milano, C., & Postma, A. (2018). *Overtourism: Impact and possible policy responses*. European Parliament, Policy Department for Structural and Cohesion Policies. [https://www.researchgate.net/publication/330502264\\_Overtourism\\_Impact\\_and\\_possible\\_policy\\_responses](https://www.researchgate.net/publication/330502264_Overtourism_Impact_and_possible_policy_responses)
- Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., Young, B. E., Graham, C. H., & Costa, G. C. (2014). Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods in Ecology and Evolution*, 5(9), 961–970. <https://doi.org/10.1111/2041-210X.12232>
- Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7), Article e1000443. <https://doi.org/10.1371/journal.pcbi.1000443>
- Pfeuffer, N., Baum, L., Stammer, W., Abdel-Karim, B. M., Schramowski, P., Bucher, A. M., Hügel, C., Rohde, G., Kersting, K., & Hinz, O. (2023). Explanatory interactive machine learning. *Business & Information Systems Engineering*, 65(6), 677–701. <https://doi.org/10.1007/s12599-023-00806-x>
- Pham, T.-A. N., Li, X., & Cong, G. (2017). A general model for out-of-town region recommendation. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. <https://doi.org/10.1145/3038912.3052667>

- Piliponyte, G., Massimo, D., & Ricci, F. (2024). Simulation of recommender systems driven tourism promotion campaigns. *Information Technology & Tourism*. <https://doi.org/10.1007/s40558-024-00283-2>
- Preoțiu-Pietro, D., Cranshaw, J., & Yano, T. (2013). Exploring venue-based city-to-city similarity measures. In Y. Zheng (Ed.), *ACM Digital Library, Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing* (p. 1). ACM. <https://doi.org/10.1145/2505821.2505832>
- Purwandani, I., & Pakan, S. P. (2022). Local habitus and temporal overtourism in Yogyakarta. *Consumer Behavior in Tourism and Hospitality*, 17(4), 544–560. <https://doi.org/10.1108/CBTH-07-2021-0177>
- Qiu, P., Gao, J., & Lu, F. (2021). Identifying the relatedness between tourism attractions from online reviews with heterogeneous information network embedding. *ISPRS International Journal of Geo-Information*, 10(12), 797–817. <https://doi.org/10.3390/ijgi10120797>
- Rasmussen, C. E., & Williams, C. K. I. (2006). Gaussian processes for machine learning. *The MIT Press*. <https://doi.org/10.7551/mitpress/3206.001.0001>
- Schwietering, A., Steinbauer, M., Mangold, M., Sand, M., & Audorff, V. (2023). Digitalization of planning and navigating recreational outdoor activities. *German Journal of Exercise and Sport Research*. <https://doi.org/10.1007/s12662-023-00927-1>
- Seraphin, H., Sheeran, P., & Pilato, M. (2018). Over-tourism and the fall of Venice as a destination. *Journal of Destination Marketing & Management*, 9, 374–376. <https://doi.org/10.1016/j.jdmm.2018.01.011>
- Shackley, M. L. (2000). *Visitor management: Case studies from World Heritage Sites* (1st ed.). Butterworth-Heinemann. <https://doi.org/10.4324/9780080520681>
- Shen, X., Gong, X., Cai, Y., Guo, Y., Tu, J., Li, H., Zhang, T., Wang, J., Xue, F., & Zhu, Z.-J. (2016). Normalization and integration of large-scale metabolomics data using support vector regression. *Metabolomics*, 12(5), 1–12. <https://doi.org/10.1007/s11306-016-1026-5>
- Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0144059>
- Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, 35(3), 553. <https://doi.org/10.2307/23042796>
- Singh, S. (2003). Simple Random Sampling. In S. Singh (Ed.), *Advanced sampling theory with applications* (pp. 71–136). Springer Netherlands. [https://doi.org/10.1007/978-94-007-0789-4\\_2](https://doi.org/10.1007/978-94-007-0789-4_2)
- Smith, I., Velasquez, E., Norman, P., & Pickering, C. (2023). Effect of the COVID-19 pandemic on the popularity of protected areas for mountain biking and hiking in Australia: Insights from volunteered geographic information. *Journal of Outdoor Recreation and Tourism*, 41, Article 100588. <https://doi.org/10.1016/j.jort.2022.100588>
- Stamatelatos, G., Drosatos, G., Gyftopoulos, S., Briola, H., & Efraimidis, P. S. (2021). Point-of-interest lists and their potential in recommendation systems. *Information Technology & Tourism*, 23(2), 209–239. <https://doi.org/10.1007/s40558-021-00195-5#Sec23>
- Tiwari, P., & Chowdhary, N. (2021). Has COVID-19 brought a temporary halt to overtourism? *Turyzm/Tourism*, 31(1), 89–93. <https://doi.org/10.18778/0867-5856.31.1.20>
- Tokarchuk, O., Barr, J. C., & Cozzio, C. (2022). How much is too much? Estimating tourism carrying capacity in urban context using sentiment analysis. *Tourism Management*, 91, Article 104522. <https://doi.org/10.1016/j.tourman.2022.104522>
- Tomkins, S., Isley, S., London, B., & Getoor, L. (2018). Sustainability at scale. In S. Pera & A. S. I. G. o. C.-H. Interaction (Eds.), *ACM international conference proceedings series, Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 214–218). ACM. <https://doi.org/10.1145/3240323.3240411>
- Tran, T. N., Lam, B. M., Nguyen, A. T., & Le, Q. B. (2022). Load forecasting with support vector regression: Influence of data normalization on grid search algorithm. *International Journal of Electrical and Computer Engineering (IJECE)*, 12(4), 3410. <https://doi.org/10.11591/ijece.v12i4.pp3410-3420>
- Tsai, C.-Y., Chuang, K.-W., Jen, H.-Y., & Huang, H. (2024). A tour recommendation system considering implicit and dynamic information. *Applied Sciences*, 14(20), Article 9271. <https://doi.org/10.3390/app14209271>
- Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32(9), 1761–1768. <https://doi.org/10.1016/j.energy.2006.11.010>
- Tukey, J. W. (1977). *Exploratory data analysis*. Addison Wesley. <https://doi.org/10.1002/bimj.4710230408>
- UN Tourism. (2018). ‘Overtourism’? - Understanding and managing urban tourism growth beyond perceptions, executive summary. UN Tourism. <https://doi.org/10.18111/9789284420070>
- UN Tourism. (2022). *Tourismusaufkommen weltweit bis 2021* | Statista. <https://de.statista.com/statistik/daten/studie/37123/umfrage/weltweites-tourismusaufkommen-nach-reiseankuenften-seit-1950/>. Accessed 10 Mar 2022.
- UN Tourism. (2023). *World tourism barometer - January 2023*. UN Tourism. [https://webunwto.s3.eu-west-1.amazonaws.com/s3fs-public/2023-01/UNWTO\\_Barom23\\_01\\_January\\_EXCERPT.pdf?VersionId=\\_2bbK5GIwk5KrBGJZt5iNPAgnrWoH8NB](https://webunwto.s3.eu-west-1.amazonaws.com/s3fs-public/2023-01/UNWTO_Barom23_01_January_EXCERPT.pdf?VersionId=_2bbK5GIwk5KrBGJZt5iNPAgnrWoH8NB). Accessed 2 Feb 2023.
- UN Tourism. (2024). *Global tourism set for full recovery by end of the year with spending growing faster than arrivals*. <http://prewebunwto.s3.eu-west-1.amazonaws.com/s3fs-public/2024-12/NR%20Barometer%20EN.pdf?VersionId=mO6F27vm4cVI.0akchKW4HAuqX86Eqf2>. Accessed 4 Jan 2025.
- United States Department of Agriculture. (2023). *U.S. forest service national visitor use monitoring survey results national summary report 2022: Data collected FY 2018 through FY 2022*. Accessed 12 Mar 2025. <https://www.fs.usda.gov/sites/default/files/2022-National-Visitor-Use-Monitoring-Summary-Report.pdf>
- Vías, J., & Ocaña, C. (2014). Multicriteria evaluation by GIS to determine trail hiking suitability in a natural park. *Boletín De La Asociación De Geógrafos Españoles*, 66, 323–339. [https://www.researchgate.net/publication/279102247\\_Multicriteria\\_evaluation\\_by\\_GIS\\_to\\_determine\\_trail\\_hiking\\_suitability\\_in\\_a\\_natural\\_park](https://www.researchgate.net/publication/279102247_Multicriteria_evaluation_by_GIS_to_determine_trail_hiking_suitability_in_a_natural_park)
- Vías, J., Rolland, J., Gómez, M. L., Ocaña, C., & Luque, A. (2018). Recommendation system to determine suitable and viable hiking routes: A prototype application in Sierra de las Nieves Nature Reserve (southern Spain). *Journal of Geographical Systems*, 20(3), 275–294. <https://doi.org/10.1007/s10109-018-0271-8>
- Wang, W., Yin, H., Chen, L., Sun, Y., Sadiq, S., & Zhou, X. (2015). Geo-SAGE: A geographical sparse additive generative model for spatial item recommendation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. <https://doi.org/10.1145/2783258.2783335>
- Wang, M.-F., Lu, Y.-S., & Huang, J.-L. (2019). SPENT: A successive POI recommendation method using similarity-based POI embedding and recurrent neural network with temporal influence. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp 2019): Kyoto, Japan, 27 February–2 March* (pp. 1–8). IEEE. <https://doi.org/10.1109/BIGCO MP.2019.8679431>
- Weisberg, S. (2005). *Applied linear regression* (3., Auflage). John Wiley & Sons. <https://doi.org/10.1002/0471704091>



- Wimmer, H., & Rödiger, S. (2019). Berücksichtigung von Schutzgebieten bei der Entwicklung der Online-Plattform outdooractive.com. In R. Forst, M. Porzelt, & V. Scherfose (Eds.), *BfN-Skripten: Vol. 520. Konflikte durch Erholungsnutzung in Großschutzgebieten und deren Entschärfung durch innovatives Besuchermanagement* (pp. 91–101). <https://doi.org/10.19217/skr520>
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, I*. <https://api.semanticscholar.org/CorpusID:1211505>. Accessed 15 June 2022.
- Xu, M., Watanachaturaporn, P., Varshney, P. K., & Arora, M. K. (2005). Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment*, 97(3), 322–336. <https://doi.org/10.1016/j.rse.2005.05.008>
- Xue, Y., Tang, Y., Xu, X., & LianNeri, J. F. (2022). Multi-objective feature selection with missing data in classification. In *IEEE Transactions of emerging topics in computational intelligence*, 6(2), 355–364. <https://doi.org/10.48550/arXiv.2104.08747>
- Yao, Z., Fu, Y., Liu, B., Liu, Y., & Xiong, H. (2016). POI recommendation: A temporal matching between POI popularity and user regularity. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE. <https://doi.org/10.1109/icdm.2016.0066>
- Yin, H., Sun, Y., Cui, B., Hu, Z., & Chen, L. (2013). LCARS: A location-content-aware recommender system. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. <https://doi.org/10.1145/2487575.2487608>
- Yu, D., Yu, T., Wu, Y., & Liu, C. (2022). Personalized recommendation of collective points-of-interest with preference and context awareness. *Pattern Recognition Letters*, 153, 16–23. <https://doi.org/10.1016/j.patrec.2021.11.018>
- Yuan, Q., Cong, G., Ma, Z., Sun, A., & Thalmann, N. M. (2013). Time-aware point-of-interest recommendation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM. <https://doi.org/10.1145/2484028.2484030>
- Yuan, Q., Cong, G., & Sun, A. (2014). Graph-based point-of-interest recommendation with geographical and temporal influences. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM. <https://doi.org/10.1145/2661829.2661983>
- Zerva, K., Palou, S., Blasco, D., & Donaire, J. A. B. (2019). Tourism-philia versus tourism-phobia: Residents and destination management organization's publicly expressed tourism perceptions in Barcelona. *Tourism Geographies*, 21(2), 306–329. <https://doi.org/10.1080/14616688.2018.1522510>
- Zerzucha, P., & Walczak, B. (2012). Concept of (dis)similarity in data analysis. *TrAC Trends in Analytical Chemistry*, 38, 116–128. <https://doi.org/10.1016/j.trac.2012.05.005>
- Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, 17(5–6), 375–381. <https://doi.org/10.1080/713827180>
- Zhao, J., Peng, D., Wu, C., Chen, H., Yu, M., Zheng, W., Ma, L., Chai, H., Ye, J., & Qie, X. (2019). Incorporating semantic similarity with geographic correlation for query-POI relevance learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 1270–1277. <https://doi.org/10.1609/aaai.v33i01.33011270>
- Zheng, W., Zhou, R., Zhang, Z., Zhong, Y., Wang, S., Wei, Z., & Ji, H. (2019). Understanding the tourist mobility using GPS: How similar are the tourists? *Tourism Management*, 71, 54–66. <https://doi.org/10.1016/j.tourman.2018.09.019>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.