Navigating Landscapes: Approaches to Data Collection and Analysis in Tourism

Karol Jan Borowiecki
University of Southern Denmark

Maja Uhre Pedersen
University of Southern Denmark

Sara Beth Mitchell
University of Southern Denmark

Shahedul Alam Khan
University of Southern Denmark

February 5, 2024

Abstract:

In an era where data is likened to the lifeblood of innovation, its role in the cultural and natural heritage sectors emerges as both pivotal and transformative. Beyond being a mere repository of information, data in tourism is akin to a compass, guiding strategies and illuminating paths previously unexplored. This chapter delves into the rich tapestry of cultural and natural heritage, illustrating how the strategic utilization of data is not just beneficial but essential. The focus here is on collection of visitor data, and we approach a number of practical concerns, including how to design and implement a survey at a heritage site. In doing this, we present challenges related to the design of a survey, logistics, response biases, frequency, and seasonality. These considerations are complemented with suggestions on how to overcome data limitations by using data scraping algorithms or novel data sources. Finally, we provide examples for data-based storytelling by highlighting a couple of examples to illustrate what can (or cannot) be reliably deducted from data, and how to interpret it.

Keywords: Cultural Tourism, Nature-based Tourism, Data Collection, Survey Design, Data Analysis, Data-Driven Decision Making

1 Introduction

As we stand at the threshold of a new epoch in cultural tourism, the significance of data analysis in this domain cannot be overstated (Borowiecki et al., 2023a). The advent of data collection and analysis in tourism marks a paradigm shift, transitioning from intuition-based decision-making to insights driven by data.[1] In this chapter, we endeavour to unfold the layers of complexity in cultural and nature-based tourism, addressing the potential challenges and unveiling the latent opportunities within.[2] At the heart of our discourse is the conviction that data, in its myriad forms, is the key to unlocking the secrets of tourist behaviours, cultural influences, and their interplay with the host communities. Data is also essential to understand the impact of any innovative action within the tourism sector. Here, we embark on a quest to navigate the intricacies of data collection, its purpose, its potential, and its profound impact on the realm of tourism. The aim of the chapter is to illustrate the importance of data and how data collection and the analysis can become an integrated part of project planning and management.

Before starting to collect data, it is important to determine the purpose of the data collection. Cultural and nature-based tourism can be considered as an interplay of history, heritage, nature and humans that makes it complex to track the nuances and complexities defining the field. Data collection can be the foundation of our understanding of tourist behaviours, cultural influences, and the influence of tourism on host communities. Furthermore, data is also an important part of evaluating innovative projects within the tourism sector.

To enable enhancement of the tourism experience and informed decision making, we can use data to discover the insights into inclinations, and patterns. Data on tourism not only develops knowledge for researchers and professionals but also creates value to the very communities at the centre of tourism. The outcome is versatile, ranging from strategies for sustainable tourism and enrichment of tourist experiences to the protection of heritage. Data in this field must capture the essence of the culture and the travel experience. Deciding the methods of data collection is crucial but can be a challenging decision, considering the time and resource constraints. Cultural indicators, visitor demographics, economic metrics and more can do the job, however, it is the objectives of the research that will define the scope. The approach to data collection should be a blend of art and science, considering the cultural sensitivity, ethical soundness, and adaptability to circumstances of the destination. Like any research, the analytical techniques should maintain alignment between the

[1] For an early empirical study of cultural tourism, refer to Borowiecki and Castiglione (2014), who investigate the association between participation in cultural activities and tourism flows in Italian provinces.

[2] For a definition of the term cultural tourism see e.g. Du Cros and McKercher (2020) and nature-based tourism Kuenzi and McNeely (2008).

research objectives and collected data. Analysis should be able to transform data into knowledge that will illuminate our preliminary research questions. Tourism is a dynamic domain, where progress measurement goes beyond the quantitative milestones. Sustainable tourism can be an ideal mechanism for gauging progress by tracking the preservation of heritage and optimising the impact on the local community. One of the major challenges that remains is the identification of tactics to use the gathered knowledge in the continuous development of the pilots.

The chapter unfolds with a focus on data integration in project planning and management (Section 2), followed by an in-depth look at data collection and handling biases (Section 3). We then explore data analysis through diverse sources like Eurostat and Tripadvisor (Section 4), concluding with insights on the impact of these methods in cultural tourism (Section 5).

2 Data management

Data can be used during the different phases of implementation of the cultural tourism projects and should therefore be considered an integrated part of the project and project planning. In the initial phases data can be used to explore the opportunities of action, while the collection of data during and after implementation can be used to assess the impact of the pilot action. In all cases data can come from different sources, based on the needs and intentions of the project.

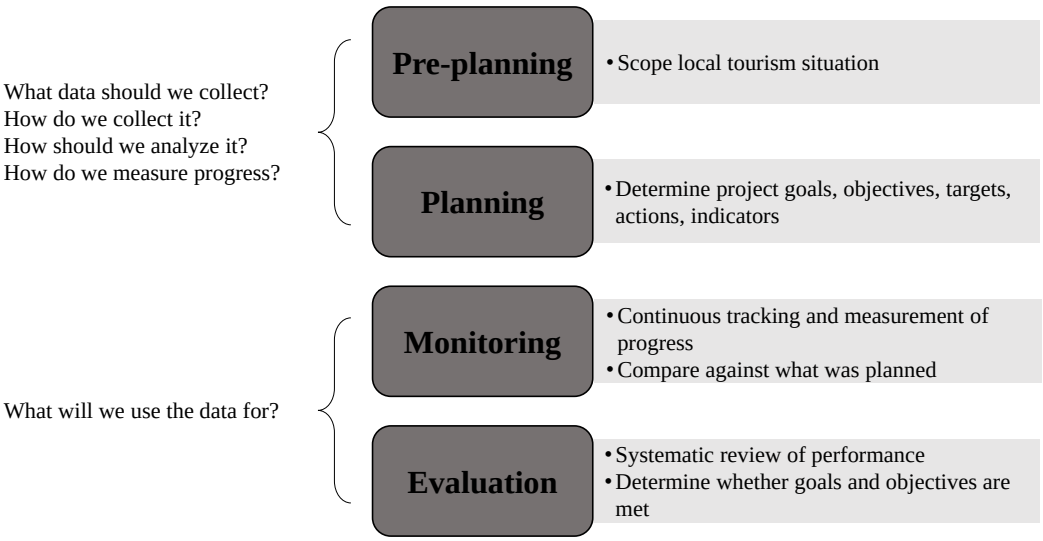
The importance of collecting data comes from the power of measuring results at all stages of a project (Gudda, 2011). For example, if the results of a project are not measured, it is impossible to tell success from failure. If the success is not measured it cannot be rewarded and we cannot learn from it. The use of data, is an important part of measuring the results.

There is a large literature regarding both management and program planning- and evaluation. The field of tourism and specifically cultural tourism is no exception. Several handbooks have been written to aid the implementation of successful interventions within tourism (e.g., Smith and Richards, 2013). However, in such handbooks, there is little to none coverage on the collection of data.

The data planning should be regarded as an integrated part of the project planning. It covers the part of decisions regarding what data to be collected, what methods and tools to use in data collection, and how the data should be analyzed subsequently. It is important to consider the framework of the project to ensure that the outcomes are aligned with the objectives of the project. Overall, project planning in cultural tourism can be considered as a four-stage process, as illustrated in Figure 1, each of which we will now discuss more in detail. Common challenges such as forms

of data, method of collection, analytical tools, and measuring progress of the pilot, will usually be covered in the first two stages of the data planning. On the other hand, the monitoring and evaluation phases form the purpose of the collection of data (UN, 2009).

Figure 1: Phases in project planning



Notes: This figure illustrates the different phases of project planning. *Source:* Own elaboration based on UNECE (2017); IOM (2021)

The pre-planning makes use of existing, to analyze the present condition of the tourism pilot and the surroundings. In this stage we should ask questions like:

- What is the tourism situation at the pilot site/ the wider region like now?
- How does the offering at the pilot site fit into the tourism offering of the wider region?

The second stage is the planning, which includes the determination of goals and objectives for the project.The planning stage can be divided into four distinct stages: 1) formulate goals and objectives, 2) develop strategies, 3) identify polices, programs, projects, and activities, and 4) develop a monitoring and evaluation strategy. The first step is to identify the intended achievements and the time frame for these achievements. Once these have been identified we can develop the strategies and the action required to reach the goals. Finally, the planning stage should also consider how the action should be monitored and how the project progress should be evaluated. To complete the planning we should ask questions like:

- What do you want to achieve for tourism in the future and when?

- How will you get from the situation you are at now to where you want to be in the future?
- What specific actions will you take to implement the strategies?
- How will you measure the progress?

In the monitoring stage we continuously monitor and measure the progress within different areas of the project. Here the progress is compared to the planned metrics such as the maintenance of the time schedule and budget and the attainment of goals. The monitoring also includes identification of causes behind delays and unexpected results together with the adaption of the plans due to changes in circumstances. In this stage we should ask questions like:

- Are the activities leading to the expected outputs?
- Are activities being implemented on schedule and within the budget?
- What is causing delays or unexpected results?
- Is anything happened that should lead management to modify implementation plan?
- How do stakeholders feel about the pilot?

Finally, the evaluation stage consists of a systematical review of the performance and effectiveness of the goal attainment of the pilot project. This evaluation should be conducted periodically and should also include an evaluation of the resource usage such as time-, financial- and human resources. In this phase the following questions should be asked:

- Have the aims/objectives been achieved?
- How well were resources used?
- Are benefits associated with pilot project likely to last?
- Are pilot project aims/objectives/actions responding to the needs of the local community?
- How well do the project actions fit the needs of the wider region?

In all phases of project planning, it is important to have clear goals. At the initial stages it should be determined what achievements the project is aiming for, and a clear time frame for both the short-, medium- and long term. This will give stakeholders involved a clear message of all

objectives. To set clear objectives the SMART criteria is a useful tool. SMART is a set of criteria which guides towards setting goals to achieve better results in management and was first proposed in Doran (1981). Even though initially proposed as a tool for management, it has been further developed and is widely cited within the program planning/evaluation literature (see eg. Bjerke and Renger, 2017; Chen, 2014; Gudda, 2011). The objectives must be *specific*, where the actions, roles, responsibilities, and accountabilities must be clearly mentioned. They must be *measurable* by developing appropriate metrics to observe, analyse and verify the outcomes of the efforts. The objectives must be *attainable* considering the given time and resources, and *relevant* considering whether they are useful in obtaining the expected outcomes and results. Finally, the objectives must be *time-bounded*, i.e. they must assign a time budget to achieve the objectives. Once the objectives have been determined, the actions should be specified together with the targets. This is the stage where data can be linked to the project. In determining the targets we should start by considering the baseline situation, i.e. the status quo, without the implementation of the new project. To determine the baseline situation, historical data covering the last few years, for example 3-5 years, can be used to see the current levels and trends in the sector. The baseline can be used to reflect upon the project objectives, actions and targets. After having determined the baseline a few realistic and feasible potential scenarios after implementation can be compared to this. The potential scenarios should include a best-case and a worst-case scenario together with the most realistic. This part of the planning helps detecting the best and worst possible outcomes of the project.

There are some potential pitfalls and barriers to effective project planning. A common challenge is to set too many goals, which will lead to a overwhelming work burden. It is also important to avoid goals that are not well defined or too broad. A poorly defined goal can create confusion and a lack of focus, hampering the effectiveness of the project. Finally, another obstacle is determining unachievable aims considering the given time frame, leading to a potential failure.

A good way to overcome the above challenges is to use the *Problem tree analysis* where the problem is mapped out to problem, causes and effect. Potentially, the pilot project has the aim of reversing the final effect, e.g. to reverse depopulation in the area. The main problems behind the cause will often require wider policy changes which are beyond the scope of the project. Instead of looking directly at the problems, they can be mapped out into causes towards which the pilot action can be directed.

3 Data collection

Data can be collected during the different phases explained in Section 2 and can come from different sources, primary and secondary. Since the collection of primary data can be both costly and time consuming, it is good practice to investigate what data already exist from other studies. In this chapter we consider three types of collected data each with its own specific advantages and disadvantages. In the remaining part of this section, the main focus will be on the collection of primary data specifically related to the tourism sector, i.e. visitor surveys. However, part of the discussion can also apply to other data sources which we will present in section 4.

3.1 Data in the pre-planning phase

As a part of the pre-planning process, it is important to develop a profile of the local tourism industry and a profile of the visitors. The key elements of the tourism industry profile are illustrated in Table 1 panel a together with the list of actions to take. The local tourism industry profile is used to identify the scope of tourism and the related problems to be solved. The first two elements consider existing factors which have an impact on the project and helps identifying what new possibilities can be created. The third element consider possible concerns of visitors such as lack of public transportation, disability access, language barriers, theft, hazards etc. These are all areas of improvement of the existing infrastructure and possibilities.

The elements of the visitor profile can be seen in Table 1 panel b, together with a list of content/questions to ask about the visitors. The visitor profile helps identifying the market of local tourists and their needs. The first part of the profile covers basic information about demographics of the visitors and the characteristics of the stay. The second part relates to the visited site. Here it is especially important to know how they have heard about the place. In case of projects creating new attractions/sites it might not be possible to have site level information about the visitors. In this case a visitor profile should be created for the wider area/region and it should be considered what segments of tourists to attract. In some cases, the visitor profile will show that the new site should attract new segments, while in other cases it it possible to draw from the existing tourist segment(s) of the wider area.

3.2 Data considerations

Before presenting the collected data, there are some important steps to take. First, the data should be cleaned up, to ensure accurate and reliable results during the analysis. Secondly, the

Table 1: Content of tourism and visitor profiles

Panel (a) - Local tourism industry profile	
Tourism resources and assets	List of attractions
	List of Facilities
Institutional elements	List of stakeholders
	Overview of local tourism sector
	Local government capacity
	Infrastructure
Tourist concerns	List of possible concerns of tourists
Panel (b) - Visitor profile	
Visitor demographics	Age
	Gender
	Nationality
	Who they travel with (number, type)
Characteristics of stay	Length of stay
	Overnight stay
	Travel as part of a package tour
	Which attractions do they visit
	What is the estimated spending
Site discovery	How did they hear about the site
	What resources were used to explore/learn before visiting
Site-level information	Profile visitors from wider area or region
	Attract new tourist segment(s)
	Draw part of existing tourist segment(s)

Notes: This table illustrates the content of a profile of the local tourism industry in Panel a and of a visitor profile in panel b. *Source:* Own illustration based on UNECE (2017).

missing values should be treated properly. Finally, different potential biases in the data should be considered and discussed.

To clean up the data we look for errors such as missing and extreme values. It is good practice to double check the data and create a summary table with min, max and average values to check for

abnormalities. If the data set is large, it can be an option to do nothing, especially if the margin of error is small. However, for small data sets treating the errors is important. The errors can only be corrected if the accurate answer can be confirmed or if the intention is obvious. Otherwise, the solution could be to delete the incorrect observations. If data is deleted it is important to assure that the mistakes are not systematic and the exclusion criteria should be transparent.

Treating missing values can be a tricky task. First of all, it is important to get an overview of how many missing values exist and for which variables. It is also good practice to get an understanding of why the values are missing. For example, in a survey it is always suggested to include an option such as "I don't know", "I don't want to answer", "not applicable", etc. Once the missing values have been identified one possibility is to determine if the missing values can be imputed based on other available data. If imputation is not a possibility, another solution is to exclude the missing values or variables with many missing values from the analysis. In all circumstances, it is important to determine if the data are missing at random or systematically. If data is systematically missing, the implications or potential bias should be considered.

Finally, different biases can distort the results and conclusions from the actual scenario. It is not always possible to treat biases in the data, but a section should be dedicated to a discussion of the potential biases acknowledging their existence and, in case they are dealt with, how this is done. Biases can exist in all types of data collection methods, but when the data is collected from surveys there are two main groups of biases which can occur: respondent biases and researcher biases (IOM, 2021).

Respondent biases are biases related to the answers provided in the survey questionnaire. We will go through some of the common potential biases and how they can be treated. Often, respondent biases can be reduced by constructing an adequate questionnaire which reduces the risk of a biased answer. In other cases, the bias should be treated during the initial phases of selecting the representative sample.

Selection Bias: Selection bias can arise when certain groups of respondents may systematically agree or disagree to participate. It can be a concern when people volunteer to participate in a study as a respondent, they may answer differently than the people who did not volunteer.

Non-response bias: This refers to respondents who refuse to or are not able to respond to the study. In such a case the collected data will not properly represent the perception of the target population.

Attrition bias: If a study requires more than one round of answers from the same respondents there is a risk of attrition bias. The data collected may fail to represent the population, if re-

spondents drop out of the study mid-way through and force the project personnel to adjust the sample.

Acquiescence bias: When respondents have the tendency to respond positively towards every question in the survey, it creates biases. In this regard, questions can be revised in a form to get the actual reply from the respondents.

Social desirability bias: In this bias, the respondents tend to give what they think is the socially acceptable answer. To deal with this bias, the question asked should be indirect, so that they do not have the pressure of social acceptance while answering (see e.g. Fisher, 1993).

Anchoring bias: Regarding this bias, the respondent's answer is influenced by a reference point. In answering questions, respondents may rely on the information given in the earlier stages of the survey. This information can work as an anchor and influence them to give a biased answer. The best way to deal with this bias is to avoid leading or suggesting language, provide diverse perspectives, and randomise the order of questions while preparing the survey questionnaire.

Recall bias: In some cases, the respondents may have difficulty remembering certain information. In such issues, we may refer to them some key facts that will help them to recall the relevant information.

Of particular relevance in when conducting visitor surveys, is the social desirability bias, which can have important implications for the results. In Dahlgren and Hansen (2015) they explain how the nationality of the interviewer can influence the answers of the respondents. When the interviewer is of the same nationality as the target destination, they show that respondents will assess more positively the attraction. Therefore, the quite common practice of a local or domestic interviewer who interviews tourists at a destination, is severely prone to biased results and should be taken into account when planning the survey.

The second category of biases is related to the person conducting the survey/analysis, known as researcher bias.

Question-order bias: The sequence of questions may influence the response and create bias. Selection of words and presentation of ideas may create a partial impression in the mind of the respondents and influence the subsequent answers. To reduce the impact of this bias, general information can be sought before specific information.

Leading questions/ wording bias: Wording as mentioned earlier may nudge the respondent to a particular answer. To reduce this bias, researchers should frame questions using the language familiar to the respondents and refrain from paraphrasing their responses from their own

perspective.

Confirmation bias: Researchers focus on information that reinforces or confirms their hypothesis or belief. To reduce the bias, the personnel should frequently review the imprints of respondents and question existing assumptions and hypotheses.

3.3 Presenting the results

The presentation of the data is vital in the communication of the findings of the pilot project. The presentation and discussion of the data should be based on knowledge and understanding of the data and the topic of study. Furthermore, the data should be put in the context of short- and long term trends and explore relationships, causes and effects. In the following, two common ways to present data will be explained: tables and charts.

Smaller presentation tables are usually used to supplement the information given in the text and contain key figures of the results, i.e. summary statistics. On the other hand, reference tables are longer tables which contain the exact data. These are usually only referred to, and not presented directly in the text.

The charts are used as a way to visualise the results, and a well organised graph can contain large amounts of information. There are different types of charts each with a specific purpose. Deciding which one to use, depends on the results we want to illustrate and what knowledge the reader should obtain from the graph. It is also important to consider the target reader, and adapt the charts to the level of understanding of the reader. Charts are very good when we want to compare variables, e.g. comparing the number of visitors in two different periods. They can also be used to show changes over times, e.g. a line chart showing the change in the number of visitors over the last years. A chart showing the frequency distribution can illustrate occurrences within different categories such as visitors using different types of transportation. Finally, charts are useful to show correlations between variables such as the correlation between the number of tour guides in a location and the number of visitors.

The following guidelines prepared by UNECE (2009), are useful to take into consideration when presenting the data:

- **The target group**: Tailor the writing according to the knowledge and interests of the target group.
- **The role of the graphic in the overall presentation**: Graphs can only add value to the

presentation when aligned with the intended message of the report in terms of highlighting contrasts or emphasising trends.

- **How and where the message will be presented**: The approach to presentation will vary depending on the platforms and the audience.
- **Contextual issues that may distort understanding**: It is important to consider the socio-cultural, historical, or economic issues that may distort the understanding of the audience.
- **Whether textual analysis or a data table is the better solution**: Plain textual analysis can be more effective than a complicated table in highlighting the valuable insight of a circumstance.
- **Accessibility considerations**: Everyone should be able to access and understand the data, regardless of technology and disabilities.
- **Consistency across data visualisations**: The use of colours, scales, and labelling norms should be consistent across data visualisations.
- **Size, duration, and complexity**: Consider that, when a long report is presented to the audience, requiring a huge amount of time to read, it makes the understanding more complex.
- **Possibility of misinterpretation**: When readers lack the literacy to interpret graphs, charts and complex statistics, it creates possibilities for misinterpretation.

In conclusion, when presenting data it is important to have a clear message and to keep the information simple without providing unnecessary information. It is also important to make sure that missing values and abbreviations are explained properly in the text. Finally, the target audience should be kept in mind.

3.4 Visitor surveys

We will here cover some of the important considerations the pilot should include when conducting visitor surveys to obtain data. The survey can be conducted in different ways, most commonly via an online survey or a physical survey (e.g. paper questionnaire). The visitor survey should collect data on demographics of the visitors (what kind of people visit, where are they from,- how long are they staying, etc.). This survey can also include questions about the visitor experience (how would they rate their trip, how likely are they to recommend the site to others, etc.). In some cases, the

pilot project involves digital platforms (for examples, see Borowiecki et al., 2016). Here an effort should be made to collect data on digital engagement and (if feasible) conduct a survey of online visitors. It is possible to collect basic information about digital engagement/website visitors using tools such as Google analytics. Appendix A provides a sample visitor survey with an introduction and questions to be asked.

To have reliable and useful results, visitor surveys should be conducted on a day-to-day basis for all visitors at the site. For more remote sites where this is not possible, an effort should be made to conduct a visitor survey twice a year (once during low season and again during high season). During each of these survey periods, an effort should be made to conduct the survey at least once on a weekday (Monday-Thursday) and at least once on a weekend (Saturday or Sunday). Furthermore, it is important that the survey is distributed randomly to all visitors to assure that the results are representative.

Finally, it is important to consider the content of the survey, to assure that it fits to the potential respondents. For example, people walking a trail may not have the patience to conduct a long survey while online users who actively use the digital tool may have better conditions to take a longer survey. It is also important, that the questions are impartial, clear and precise to reduce potential biases and missing values.

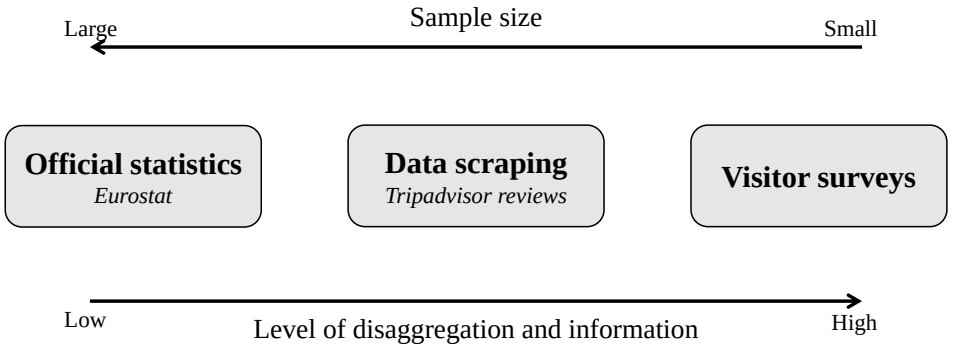
4 Data analysis

Throughout this chapter we have emphasised the importance of data in all steps of a project, from pre-planning to implementation and evaluation. A special focus has been on data collected from primary sources such as visitor surveys. However, the type of data to be collected, is specific to the available budget, context and needs.

The aim of this section is to provide an illustrative example of an analysis of tourism trends in a selected location, using different sources of data, each with its own advantages and limitations. The analysis complements the previous sections, by providing different sources of data, and briefly explaining the advantages and limitations in using each of them. Furthermore, it is also an illustration of how a simple analysis can be conducted, and how the data can be used and presented to show tourism trends. This section also illustrates how alternative and innovative data sources can be implemented. Finally, it improves the understanding of the different data sources, by illustrating what type of information, and at what level of detail, each source can offer. As mentioned in Section 3, data can be collected from both primary and secondary sources, and exist at different

levels of aggregation and with different levels of information included. We will present data from the following three sources: Eurostat, Tripadvisor and INCULTUM pilot visitor surveys. Each type of data has its own advantages and disadvantages which should be considered before deciding what data to collect. The list of sources is not exhaustive, but is illustrative of different levels of detail and aggregation. In Figure 2 we present an overview of the three levels of data presented in this section. Each of the three selected sources will be explained in the following subsections.

Figure 2: Selected data sources



Notes: This figure illustrates different data sources together with their sample size and level of aggregation and information. *Source:* Own illustration.

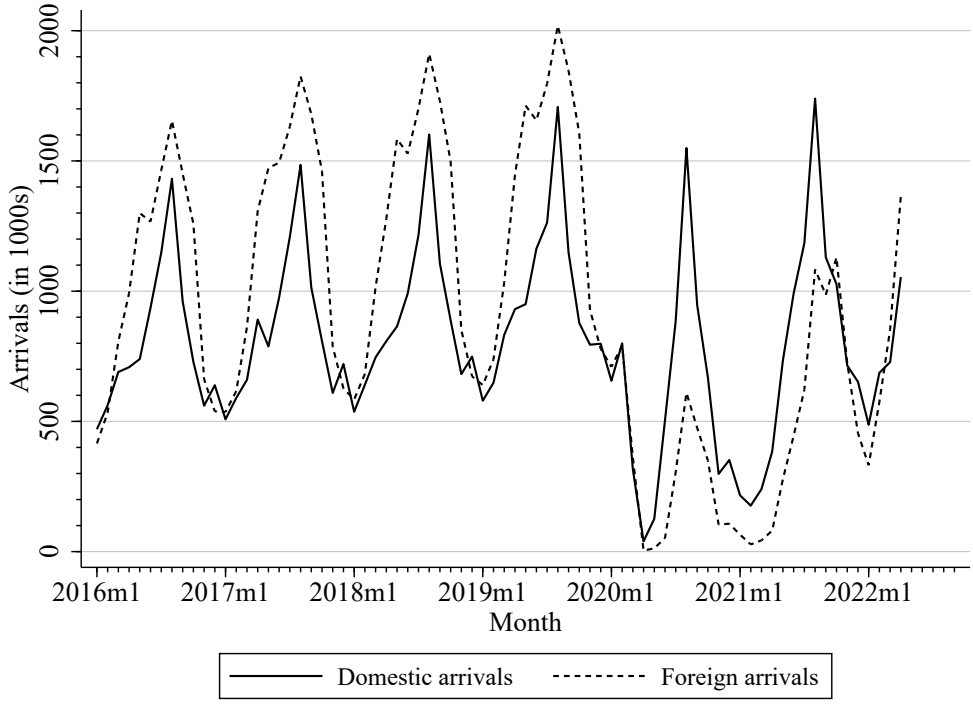
As an illustrative example, we concentrate on the Portuguese INCULTUM pilot site, and present the data from the three aforementioned sources to show details about tourism and tourists in this location.

4.1 Official statistics

At the highest level of aggregation we have data from official statistics such as Eurostat. Official statistics have the advantage of being a reliable source and comes with a large sample size and are comparable across countries and over time. However, data is usually highly aggregated both in time and space, meaning that it can be hard to detect effects for smaller units such as the region, city or attraction level. Usually, official statistics also have a limited amount of information at the individual level, meaning that they cannot be used to establish effects regarding the visitors of a specific location. We present data related to tourism provided by the statistical office of the European Union, Eurostat. Eurostat provides the number of arrivals at tourist accommodations by month and country, and separately for domestic and foreign visitors (Eurostat, 2023). An arrival at a tourist accommodation establishment is defined as a person (tourist) who arrives at a tourist accommodation establishment and checks in. There are made no restrictions on age, meaning that adults as well as children are part of the statistic. Same-day visitors that spend only few hours (no

overnight stay) are excluded from this statistic. In Figure 3 we show the number of domestic and foreign arrivals in Portugal over time for the period 2016 to 2022. From Figure 3 there is a clear pattern of seasonality for both domestic and foreign tourists.

Figure 3: Eurostat domestic and foreign arrivals over time



Notes: This figure shows the the number of domestic and foreign Eurostat arrivals over time in Portugal. *Source:* Official statistics from Eurostat (2023).

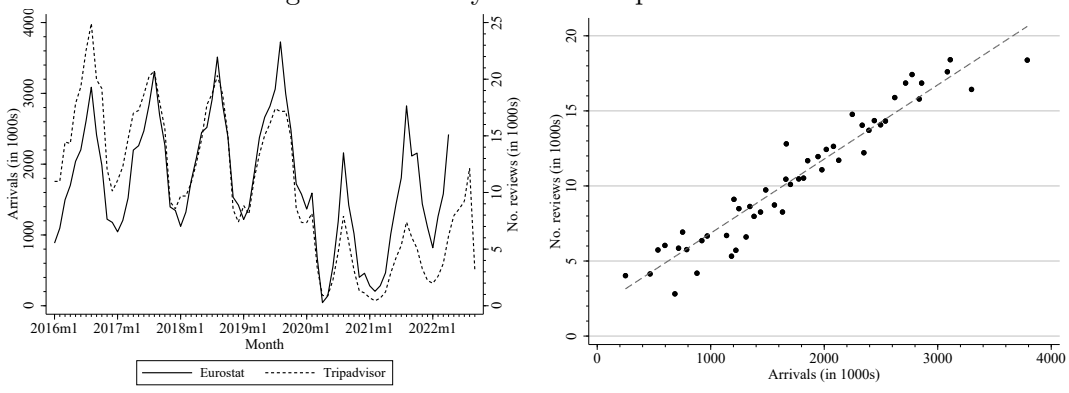
4.2 Data scraping from Tripadvisor

At the second level we have alternative data sources, such as data scraping which can be used to obtain fairly large samples of data and, at the same time, contain more information about the visitors and attractions than the official statistics. In Borowiecki et al. (2023b), we present this new method relying on reviews collected from the travel portal Tripadvisor for all attractions in various countries. In Borowiecki et al. (2023c) we re-apply the method to collect reviews for all attractions in countries in which an INCULTUM pilot site is located. The data contains a list of all reviews in English and native language published by users, together with information about the user posting the review, and all the attractions in each country. This gives us a more detailed set of information, including both the visitors and the attractions. The information included in the data set are: date of the review, name and location of attraction, location of user, type of visit (e.g. with family or friends), type of attraction (e.g. museum), rating of the attraction, and travel distance. Before inferring results from an alternative source like Tripadvisor reviews, it is important to validate the

data. The validation assures that any results obtained from the data are reliable. Borowiecki et al. (2023b) validate, both through a visual inspection and a formal analysis, the pursued approach by comparing the novel Tripadvisor data to the official statistics from Eurostat. In what follows we present methods for the visual inspection used in Borowiecki et al. (2023b).

A simple way to validate the data, is by comparing the time trends. In Panel a of Figure 4 we show the number of Tripadvisor reviews over time, together with the number of arrivals from Eurostat. The change in the number of reviews over time follow the change in the Eurostat arrivals quite well, indicating that the reviews are a good approximation of tourism flows in Portugal. A second way to inspect the validity, is to look at the binned scatter plot in Panel b of Figure 4 where we illustrate the correlation between the two variables, arrivals and reviews. The closer the points are to a straight line, the higher is the correlation. From Panel b of Figure 4, it is clear that they correlate quite well. Given the high correlation, we can establish that the Tripadvisor data is a valid way to analyse tourism flows. Further and more formal validity tests can also be carried out by estimating the correlation shown in Panel b of Figure 4 in a regression setting and looking at the significance of the estimate (see Borowiecki et al., 2023b,c).

Figure 4: Validity tests of Tripadvisor data



(a) Change in arrivals and reviews over time

(b) Binned scatter plot

Notes: This figure is a visual inspection of the validity of our Tripadvisor data using data from Portugal. Panel a shows the change in Tripadvisor reviews over time together with Eurostat arrivals. Panel b shows binned scatter plots between number of Eurostat arrivals and Tripadvisor reviews. *Source:* Arrivals from Eurostat (2023) and Tripadvisor reviews from Borowiecki et al. (2023b,c) (see Section 4.2 for details).

Once the data is validated we can use it to look at tourism flows in our selected INCULTUM location in Portugal. Given that we have information about the individual attractions and users, we can look at a smaller unit of observation than the country. This analysis can be used to create a profile of the local tourism trends close to the pilot site and also a broad profile of the visitors in terms of origin and attraction choices. In Figure 5 we show different results aggregating our data at the NUTS3 level, and concentrating on the NUTS3 regions close to the location of the

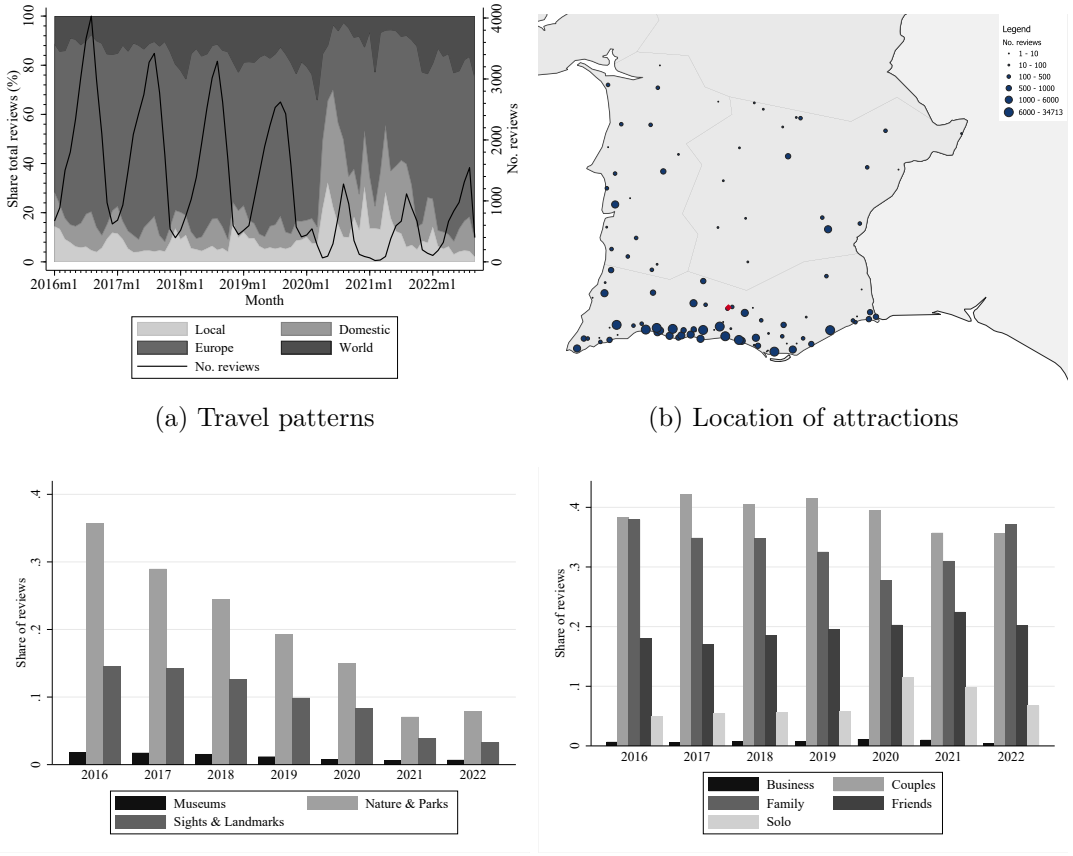
Portuguese INCULTUM pilot site. We identify all attractions located within the NUTS3 region where the pilot site is located and attractions in the bordering NUTS3 regions. In Panel a of Figure 5 we break the total number of reviews for the selected NUTS3 regions into four different travel categories: local, domestic, Europe and world. The different shades of grey, show the share of reviews for each of the four travel categories, out of the total number of reviews. It can be noticed that there is a very high share of European visitors most of which are from outside Portugal, with the exception of a shorter period following the Covid-19 pandemic in 2020. On the other hand, the share of visitors from outside Europe is quite low. In Panel b of Figure 5 it is possible to see the location of all attractions in the selected NUTS3 regions. The size of each dot is based on the number of reviews at the attraction, while the red dot indicates the approximate location of the Portuguese INCULTUM pilot site. There is a clear pattern of the location of attractions and the number of reviews, with a higher concentration close to the coast and much less attractions in the inland. In Panel c of Figure 5 we go more in detail with the type of attractions visited. Tripadvisor categories each attraction into one or more categories based on the attraction type. Given that INCULTUM has a special focus on cultural and nature-based tourism, we present results for the three attraction categories most related to these. From Panel c of Figure 5 it is possible to see the share of reviews in the three categories: museums, nature & parks, and sights & landmarks. In the Portuguese pilot area there is a high share of visitors in natural sites such as parks, followed by visits to different sights and landmarks. At the same time, there is a downward trend in all three categories, indicating that cultural tourism is trending downwards. Finally, in Panel d of Figure 5 we illustrate the annual trends in the type of visitors in the five categories: business, couples, family, friends and solo. The share of visitors, going for business is very low in all years reaching levels well below 5% in all years. The two largest categories are families and couples constituting more than 30% of all visitors each.

4.3 Data from visitor surveys

To complete the presentation of data, we look at the most detailed data, namely the data obtained from visitor surveys conducted at the pilot site. The level of information possible from visitor surveys is very high, while the sample size is usually quite small, given limited resources and time constraints. However, the additional information obtainable from a visitor survey makes this an important part of the evaluation of the pilot action.

An alternative to conducting visitor surveys is to use the surveys conducted by the European Commission. The European commission conducts surveys on travel behaviours and motivations

Figure 5: Tourism trends using Tripadvisor data



(a) Travel patterns

(b) Location of attractions

(c) Attraction categories

(d) Visitor category

Notes: This figure shows different results obtained from the Tripadvisor data. Panel a illustrates travel patterns for the four travel categories 1) Local, 2) Domestic, 3) Europe and 4) World. Panel b shows the location of attractions and number of reviews of each. The red dot indicates the approximate location of the INCULTUM pilot site in Portugal. Panel c shows the share of reviews of different attraction categories related to cultural and nature-based tourism. Panel d shows the share of reviews for different visitor types. *Source:* Tripadvisor reviews from Borowiecki et al. (2023b,c) (see Section 4.2 for details).

based on a harmonised questionnaire for a large group of countries together with other surveys of national and international travel behaviour. Such surveys are conducted at the trip level and are often of a higher quality than a stand alone survey. Furthermore, another important feature of such surveys is the stratified, representative sampling procedure and results which can be generalised. For examples of the use of such surveys see Boto-García et al. (2019) who uses survey data to analyse the length of stay in a particular location and Vergori and Arima (2020) who studies cultural tourism and seasonality.

However, in some cases, there might be very specific requirements for the survey, such as specific questions about the location, which cannot be obtained from more generic surveys. In such cases a new survey can be conducted to obtain the needed answers.

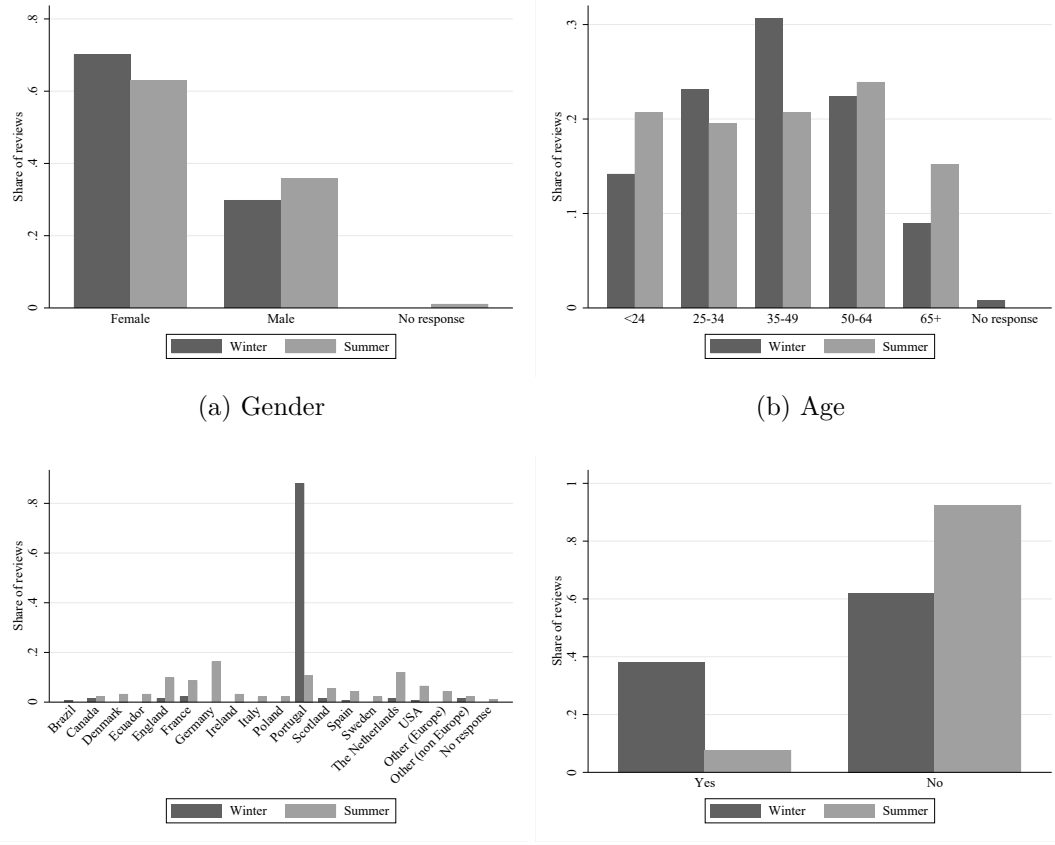
In Figure 6 we illustrate some of the results from two visitor surveys conducted close to the Por-

tuguese INCULTUM pilot site. The first survey was conducted in the period December 2022 to January 2023 and a total of 134 visitors participated in the survey. The survey was mainly targeting residents and visitors in the Algarve region and the aim was to understand the perception of the name "Campina de Faro". The second survey was conducted during the summer 2023 and was targeted mainly towards North-European beach tourists with a total of 92 respondents. The aim of the second survey was similar, with a focus on the perception and knowledge of "Campina de Faro". Given the different target populations of the two surveys, they are not entirely comparable. However, they both have a focus on the knowledge of the pilot site, and hence the results show how different populations perceive the location, which make the comparison useful. For more details about how the surveys were conducted, we refer to Chapter 3, where sampling strategy and representativeness are explained more in detail.

In Panel a of Figure 6 it is possible to see the gender distribution of the respondents while in Panel b we show the age distribution. In both surveys, more than two thirds of the respondents are women, with a slightly higher share in the survey conducted during the winter. In terms of age distribution, there are some differences between the two surveys. In the winter survey there is a higher share of visitors between 35 and 49 years of age, while in the summer survey, respondents are more evenly distributed across all age intervals. In Panel c of Figure 6 it is possible to see the nationality of the respondents. In the winter survey there is a high share of Portuguese respondents, more than 80% of the sample. This is expected, since this survey was targeted towards residents. In the summer survey the distribution of nationality is more diverse. The highest share of respondents are from Germany, followed by the Netherlands and England. Visitors from both Scandinavia and Southern Europe also have good shares. Finally, the share of visitors from Portugal is also significant in the summer survey. In Panel d of Figure 6 we show the responses to the question "Have you heard about Campina de Faro?". Clearly, a large share of respondents did not know about the place, especially in the summer survey where more than 90% answered "No" against about 60% in the winter survey. This difference is reasonable, given the different target populations, and indicate that especially foreigners are not very aware about the pilot site. This is an important point given the results from Panel a of Figure 5 showing that a large share of tourists visiting the region in which the pilot is located originate from Europe and outside Portugal.

Together, the analysis of the three selected data types, gives a comprehensive overview of tourists and tourism trends in the Algarve region in Portugal.

Figure 6: Results from on-site visitor surveys conducted in Campina de Faro (Portugal)



(a) Gender

(b) Age

(c) Nationality

(d) Heard about Campina de Faro

Notes: This figure shows results from visitor surveys conducted in Algarve, Portugal during the winter 2022-2023 and the summer 2023. In the winter survey 134 visitors participated while in the summer survey 92 visitors participated. Panel a shows the gender distribution, panel b the age distribution, panel c the nationality, and panel d the answer to the question: Have you heard about Campina de Faro? *Source:* INCULTUM pilot survey data from survey conducted in the Algarve region in Portugal.

5 Conclusion

Some regard data to be as valuable as gold. Others contest instead that data is the new oil. Whether data is shiny or black, the consensus is that it is highly valuable. Consequently, data has become a crucial foundation for business decisions and drives economic activities. In some contrast to this, the cultural heritage sector often does not exploit the full potential of data.

In this chapter, we have underscored the pivotal role of data collection and analysis in enhancing and understanding cultural tourism. As we have seen, data serves as a cornerstone in the realm of cultural tourism, not only for understanding current trends and visitor behaviours but also for planning and implementing successful cultural tourism projects.

In the initial phases of project planning, data equips stakeholders with insights to explore oppor-

tunities and set realistic goals. The alignment of data planning with project planning is crucial, ensuring that data collection and analysis are integral to each phase of a project. This holistic approach enables a comprehensive understanding of both the tourism situation at pilot sites and the broader tourism landscape.

The collection of data, whether primary or secondary, presents its own set of challenges and opportunities. Primary data, particularly from visitor surveys, offers invaluable insights into visitor demographics, behaviours, and perceptions. However, it also requires careful consideration in terms of survey design, implementation, and data cleaning processes to mitigate biases and errors.

Secondary data sources vary significantly in their scope and depth. For instance, Eurostat's official tourism statistics provide a broad overview of tourism trends and visitor profiles, offering reliable and large-scale data. However, they often lack granularity and specificity, particularly when it comes to the finer details of individual tourist experiences and behaviours.

On the other hand, novel data-science approaches, such as the analysis of reviews from a leading travel portal, open up new avenues for in-depth and granular insights. Unlike traditional statistical data, Tripadvisor reviews offer a wealth of disaggregated information. This includes detailed feedback on tourist experiences, preferences, and behaviours. More significantly, these reviews can reveal patterns in tourists' past travels, their specific interests in various aspects of cultural sites, and their subjective evaluations of their experiences. Such data can shed novel light on the nuances of visitor engagement and satisfaction, providing a more detailed and nuanced picture of cultural tourism dynamics. However, it is important to note that the collection and analysis of this type of data is neither cheap nor easy, requiring specialised skills and resources.

Data presentation, a critical step in the process, demands careful consideration to ensure clarity, relevance, and accessibility. The use of tables, charts, and other visual aids must align with the intended message and audience, facilitating effective communication of the findings.

The case studies, particularly the Portuguese INCULTUM pilot, illustrate the practical application of data collection and analysis in cultural tourism. These examples highlight the diversity of data sources and methodologies, as well as the depth of insights they can provide into cultural tourism dynamics.

In conclusion, the realm of cultural tourism is on the cusp of a transformative era, propelled by the integration of comprehensive data collection and analysis. This evolution transcends traditional decision-making and project planning, paving the way for a deeper, more nuanced understanding of cultural tourism dynamics. As we navigate this ever-evolving landscape, the strategic harnessing of data emerges not just as a tool, but as a vital catalyst in sculpting sustainable and enriching

cultural tourism experiences. Looking ahead, it is this symbiosis of data and cultural insight that promises to redefine the contours of the industry, driving innovation and fostering a more connected and culturally enriched world.

References

Bjerke, M. B. and Renger, R. (2017). Being smart about writing smart objectives. *Evaluation and Program Planning*, 61:125–127.

Borowiecki, K. J. and Castiglione, C. (2014). Cultural participation and tourism flows: an empirical investigation of italian provinces. *Tourism Economics*, 20(2):241–262.

Borowiecki, K. J., Forbes, N., and Fresa, A. (2016). *Cultural Heritage in a Changing World*. Springer Cham.

Borowiecki, K. J., Gray, C. M., and Heilbrun, J. (2023a). *The Economics of Art and Culture*. New York: Cambridge University Press.

Borowiecki, K. J., Pedersen, M. U., and Mitchell, S. B. (2023b). Using big data to measure cultural tourism in ~~europe~~ with unprecedented precision. ~~Working paper, Scandinavia Working Paper series, S-WoPEc.~~

Borowiecki, K. J., Pedersen, M. U., Mitchell, S. B., and Khan, S. A. (2023c). Final Findings Analysis Report INCULTUM. Deliverable D3.3, Visiting the Margins. INnovative CULtural ToUrisM in European peripheries.

Boto-García, D., Baños-Pino, J. F., and Álvarez, A. (2019). Determinants of tourists' length of stay: A hurdle count data approach. *Journal of Travel Research*, 58(6):977–994.

Chen, H. T. (2014). *Practical Program Evaluation: Theory-Driven Evaluation and the Integrated Evaluation Perspective, 2nd Edition*. SAGE Publications.

Dahlgren, G. H. and Hansen, H. (2015). I'd rather be nice than honest: An experimental examination of social desirability bias in tourism surveys. *Journal of Vacation Marketing*, 21(4):318–325.

Doran, G. T. (1981). There's a S.M.A.R.T. way to write management's goals and objectives. *Management Review*, 70(11):35–36.

Du Cros, H. and McKercher, B. (2020). *Cultural tourism*. Tourism Economics.

Eurostat (2023). Eurostat tourism statistics. https://ec.europa.eu/eurostat/web/tourism. Last data update: 05/01/2023.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of consumer research*, 20(2):303–315.

Gudda, P. (2011). *A Guide to Project Monitoring Evaluation*. AuthorHouse.

IOM (2021). Methodologies for data collection and analysis for monitoring and evaluation. In *IOM Monitoring and Evaluation Guidelines*. Chapter 4.

Kuenzi, C. and McNeely, J. (2008). *Nature-Based Tourism*, pages 155–178. Springer Netherlands, Dordrecht.

Smith, M. and Richards, G. (2013). *The Routledge Handbook of Cultural Tourism*. Routledge.

UN (2009). Handbook on planning, monitoring and evaluating for development results. *United Nations Development Programme*.

UNECE (2009). UNECE "Making Data Meaningful" guide series- parts 1, 2 and 3. *United Nations Economic Commission for Europe - UNECE*.

UNECE (2017). Tourism Guidebook For Local Governement Units. *Department of Tourism Philippines*.

Vergori, A. S. and Arima, S. (2020). Cultural and non-cultural tourism: Evidence from italian experience. *Tourism Management*, 78:104058.

A Appendix

Sample visitor survey

Introduction

This survey is being conducted by [the surveyor (add name)]. The survey is part of INCULTUM (2021-2024), a HORIZON2020-funded project. The main goal of this survey is to better understand visitors to [add name] and how we can improve the visitor experience. If you agree to participate, we will ask you a set of questions about you and your experiences at [add name]. The survey will take approx. 15 minutes time to be answered. Your participation is voluntary, and all information will be anonymised and kept strictly confidential in accordance with the data protection laws and guidelines.

Section A. Visitor demographics

A.1 What is your gender?

- ☐ Male
- ☐ Female
- ☐ Other
- ☐ Prefer not to respond

A.2 What is your age?

_____

A.3 What is your country of residence?

_____

A.4 If you live in [add name], please indicate the county / département / provincia where you live:

_____

A.5 What is your marital status?

- ☐ Single
- ☐ Married
- ☐ Widowed

☐ Divorced / separated

☐ Prefer not to respond

A.6 How many dependent children do you have?

_____

A.7 Which category best describes you?

☐ In full-time employment

☐ In part-time employment

☐ Student

☐ Unemployed

☐ Retired/Pensioner

☐ Housewife/househusband

☐ Other (please specify): _____

☐ Prefer not to respond

A.8 What is the highest level of education you have attained?

☐ Completed secondary school or less

☐ Bachelor's degree or equivalent

☐ Master's degree / PhD or equivalent

Section B. Details of visit

B.1 Have you visited [add name] before?

☐ Yes

☐ No

B.2 If yes, when did you last visit [add name]?

☐ Within last month

☐ Within last year

☐ Previous year

☐ More than two years ago

B.3 What is the main purpose of your visit to this area?

- ☐ Vacation / holiday
- ☐ Visiting friends / relatives
- ☐ Education / training
- ☐ Conference / large meeting
- ☐ Business / small meeting
- ☐ Event
- ☐ Other

B.4 Which best describes the group you are traveling with?

- ☐ I am traveling alone
- ☐ A couple
- ☐ A family with children
- ☐ A group of friends
- ☐ A school group
- ☐ An organised tour group (not school-related)
- ☐ Other

B.5 How did you arrive at [add name] today?

- ☐ Private car, van, or motorcycle (e.g., own, friends, family)
- ☐ Rented car, van or motorcycle
- ☐ Taxi
- ☐ Public bus or coach
- ☐ Private bus or coach
- ☐ Train
- ☐ Bicycle
- ☐ Walk

B.6 Which of the following best describes your visit to the area?

- ☐ Day trip

☐ Overnight stay

B.7 If you are staying overnight, which city are you staying in?

_____

B.8 If you are staying overnight, how many nights are you staying?

_____

B.9 Which of the following best describes the type of accommodation you are staying in?

☐ Hotel, motel, hostel

☐ Guesthouse, bed and breakfast

☐ Short-term rental (e.g., Airbnb)

☐ Caravan, camping

☐ Home of friend or relative

☐ Second home

☐ Other (please describe): _____

B.10 Have you visited any of the following sites in the area? Select all that apply.

☐ [add location 1]

☐ [add location 2]

☐ [add location 3]

☐ [add location 4]

Section C. Visitor experience

C.1 How did you find out about [add name]?

☐ Friends / relatives

☐ Tourist information centre

☐ Newspaper or magazine

☐ Search engine (do not remember which websites) Travel review site (e.g., TripAdvisor, Google Places) Facebook, blog, other social media

☐ [add website]

☐ Other (please specify): _____

C.2 What factors were important for you when choosing to visit [add name]? Select all that apply.

☐ Quality of experience

☐ Good value for money

☐ Historic interest

☐ Scenery and countryside

☐ Peace and quiet

☐ Friendliness and hospitality of locals

☐ Environmental impact

☐ Geographic proximity – I live nearby / I am staying nearby

☐ Cultural proximity – I identify with what the site represents

☐ By chance – I was just passing by / I was already visiting an area nearby A particular event (please specify): _____

☐ Other (please specify): _____

C.3 Please rate your visit to [add name] on a scale of 1 (Very poor) to 10 (Excellent).

C.4 How likely are you to recommend [add name] to someone else on a scale from 1 (Very poor) to 10 (Excellent)?