



# Big Data sources applied to rural tourism

**Eduardo Cebrián Cerdá**

**Advisor: Josep Domenech i De Soria**

**Valencia, May 2024**



**UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA**

PhD Thesis

Student: Eduardo Cebrián Cerdá

Director: Josep Domenech i De Soria

Universitat Politècnica de València

*May 2<sup>nd</sup>, 2024*

## *Acknowledgements / Agradecimientos / Agraïments*

En el camino de preparación de esta Tesis Doctoral han habido muchas personas que merecen mi agradecimiento y, como mínimo, una mención en este documento por su contribución a mi formación como persona y como científico.

La primera es mi madre, Fina Cerdá Costa. Tú me introdujiste al mundo de la ciencia, me inculcaste curiosidad, y el leer 'Maravillas de la Ciencia' o 'El Porqué de las Cosas' juntos tiene mucho que ver con el camino profesional que he elegido de adulto.

Mi padre, Eduardo José Cebrián Correcher. Has sido y eres para mí un ejemplo de cómo debe comportarse y como debe trabajar un profesional íntegro y con valores. Muchas gracias por dejarme jugar con los 'Legos' y explorar mi curiosidad científica. Al final el niño sí que inventa cosas (o lo intenta).

Gracias a mi hermana Aida por convivir conmigo cuando no era tan fácil y por soportarme mucho más de lo que lo debieras haber hecho. Aunque nos hagamos mayores, siempre me vas a tener.

Gracias a mi abuela Josefa Costa González por haber creído siempre en mí, todo tiempo que hemos compartido ha sido un regalo. Siempre te llevaré conmigo.

No quiero olvidarme de mis tíos, Merche Villar y Tomás Caballero, porque fuisteis dos referentes intelectuales para mí desde muy pequeño, y porque os valoro mucho más de lo que podéis imaginar. La familia no es de sangre, es de corazón.

Más allá del ambiente familiar, y conforme he ido creciendo, nueva gente ha aparecido en mi vida y contribuyendo a que esta sea mejor.

Sheila, tú has sido mi compañera de camino en los momentos más ilusionantes pero también en los más tensos, y siempre me has mostrado un apoyo incondicional. Esta Tesis es tan tuya como mía. Gracias por sujetar mi mano durante todo el camino y por entenderme y tratarme tan bien.

Rafa Cortina y Héctor Mirete, habéis sido mis dos mejores amigos durante los últimos 14 años y los dos hermanos mayores que nunca tuve. *'M es de Mirete, la A es porque faltaba una vocal, EK son las dos primeras letras de mi nombre sin la 'H', y la 'K' quedaba mejor que la 'C'.*

En la Universidad conocí a un grupo de amigos a los que guardo en una estima que va mucho más allá de la distancia que nos separa. Dr. Pau Insa-Sánchez, Adrián Abad, Emma Revert y Diego Loras, vosotros fuisteis los primeros que me apoyastéis en este sueño, a través de las conversaciones entre clases y a través de todas las tardes que hemos pasado juntos. Muchas gracias por haber compartido conmigo todo este tiempo.

Por otro lado, no quiero tampoco olvidar a otros grandes amigos que aunque no lleven tantos años conmigo, sí que querría que quedasen reconocidos por su apoyo y su amistad durante este periodo de mi vida. José Francisco Espinosa, Héctor Dust, Isabel Nidáguila, Antonio Alcalde, Dra. Dafne Calvo y Diego García.

Sin embargo, no solo la familia y los amigos dan forma a quienes somos o cómo nos comportamos. También hay una gran parte de la formación como persona que depende de la educación que recibimos en los distintos centros educativos. Como docentes universitarios, nuestra profesión va más allá de la ciencia, consiste en transmitir un conjunto de valores positivos a las nuevas generaciones. Por eso, llegado a este punto, quiero agradecer a algunos fantásticos docentes que me han marcado a nivel personal y a los que dedico los capítulos principales de esta tesis.

Francisco Ferrer Falcón, tus clases de tecnología son lo que menos recuerdo de ti, no porque fuesen malas, si no porque los valores humanos que nos transmitías no estaban plasmados en ningún libro de texto. Recuerdo haber confiado en ti en mis momentos más duros cuando creía que nadie más me iba a comprender y jamás me defraudaste. Gracias a ti, me empecé a sentir un poquito más fuerte.

Victoria Gutiérrez Díez, no solo fuiste una excelente docente, si no que además, me hiciste entender y disfrutar de las matemáticas. Aún hoy en día empleo muchas de las técnicas y de las metodologías que tú nos enseñabas para preparar mis propias clases. Si consigo parecerme en algo a ti, sé que será señal de que estoy en el camino correcto.

Dr. Alfonso Díez Minguela, gracias a verte dar clase, supe qué era lo que quería hacer cuando acabase mis estudios. Tu pasión y tu sencillez al explicar me sirvieron de guía para saber qué tipo de docente y qué tipo de científico quería ser.

Dra. Guadalupe Serrano Domingo, gracias por tratarme tan bien y por todas las conversaciones después de las clases, por seguir despertando mi curiosidad y por ayudarme aún cuando no era tu trabajo hacerlo. Es un orgullo haber sido tu alumno.

Per últim, no vull oblidar-me del director de la meua Tesi, el Dr. Josep Doménech i de Soria. És un honor treballar amb tu cada dia, perquè eres un exemple per a tots els que aspirem a ser docents universitaris. Gràcies per transmetre-me com treballar amb diligència, integritat i amb uns estàndards de qualitat que pense que son el camí a seguir per a tots els científics. Gràcies per haver confiat en mi i en el meu treball, tant científic com organitzatiu. Espere que el tancament d'aquesta tesi no supose el final del camí, sinò el principi.

## *Abstract*

Technological advances in recent years have enabled the emergence of new data sources and with it, the storage of large amounts of data or 'Big Data' has become increasingly important. More and more scientific studies are using these Big Data sources to try to improve understanding in various scientific fields. In tourism economics, many of these sources have already been used to predict the behavior of real variables. In tourism, the usefulness of these new data sources lies in the fact that they can help to understand the behavior of tourists, from their spatial-temporal patterns to which attractions and activities are the most popular in the destination, and therefore, they can help in the decision making of economic agents.

Therefore, this thesis tries to better understand which Big Data sources are the most useful when dealing with tourism variables and also to propose methodological improvements so that these sources can be applied to the field of rural tourism, and more specifically, to the prediction of tourists.

In this thesis several advances in this aspect are presented: First, a classification of data sources that every tourist generates during his tourist process and that compose his Digital Footprint. Then, with respect to this classification, Google Trends is chosen as the most appropriate source to help predict tourist demand, but accuracy problems are found, which are demonstrated and exemplified. Further on, it is demonstrated how this accuracy error is generated through the GT sampling process and solutions are proposed to alleviate this error, namely by obtaining more extractions and using their mean. Finally, this method is tested for the prediction of monthly overnight stays in rural tourism accommodations in Spain.

In summary, the contribution that this thesis aims to make is to provide a better understanding of Big Data sources and help to generate good practices in the use of them so that they can be applied to the prediction of real variables in rural tourism, in a way that streamlines and improves the decision making of economic agents.

## *Resumen*

Los avances tecnológicos de los últimos años han permitido la aparición de nuevas fuentes de datos y con ello, el almacenamiento de grandes cantidades de datos o 'Big Data' se ha cobrado cada vez mayor importancia. Cada vez más y más estudios científicos utilizan estas fuentes de 'Big Data' para tratar de mejorar el entendimiento en diversos campos científicos. En la economía del turismo ya se han utilizado muchas de estas fuentes para predecir el comportamiento de variables reales. En turismo, la utilidad de estas nuevas fuentes de datos reside en que pueden ayudar a entender el comportamiento de los turistas, desde sus patrones espaciotemporales hasta qué atracciones y actividades son las más populares en el destino, y por tanto, pueden ayudar en la toma de decisiones de los agentes económicos.

Por tanto, esta tesis intenta entender mejor cuáles son las fuentes de Big Data que resultan más útiles a la hora de lidiar con variables turísticas y además proponer mejoras metodológicas para que dichas fuentes se puedan aplicar al campo del turismo rural, y más concretamente, a la predicción de turistas.

En esta tesis se presentan varios avances en este aspecto: Primero, una clasificación de fuentes de datos que genera todo turista durante su proceso turístico y que componen su huella digital. Después, respecto a esta clasificación, se escoge Google Trends como la fuente más adecuada para ayudar a predecir la demanda turística, pero se encuentran problemas de precisión, que son demostrados y ejemplificados. Más adelante, se demuestra cómo se genera este error de precisión a través del proceso de muestreo de GT y se proponen soluciones para aliviar este error, a saber, obteniendo más extracciones y utilizando su media. Finalmente, este método se pone a prueba para la predicción de pernoctaciones mensuales en alojamientos de turismo rural en España.

En resumen, la contribución que esta tesis pretende hacer es aportar una mayor comprensión de las fuentes de Big Data y ayudar a generar buenas prácticas en el uso de las mismas para que se puedan aplicar a la predicción de variables reales en el turismo rural, de forma que agilice y mejore la toma de decisiones de los agentes económicos.

## *Resum*

Els avanços tecnològics dels últims anys han permès l'aparició de noves fonts de dades i amb això, l'emmagatzematge de grans quantitats de dades o 'Big Data' s'ha cobrat cada vegada major importància. Cada vegada més i més estudis científics utilitzen aquestes fonts de 'Big Data' per a tractar de millorar l'enteniment en diversos camps científics. En l'economia del turisme ja s'han utilitzat moltes d'aquestes fonts per a predir el comportament de variables reals. En turisme, la utilitat d'aquestes noves fonts de dades resideix en què poden ajudar a entendre el comportament dels turistes, des dels seus patrons espaciotemporals fins a quines atraccions i activitats són les més populars en el destí, i per tant, poden ajudar en la presa de decisions dels agents econòmics.

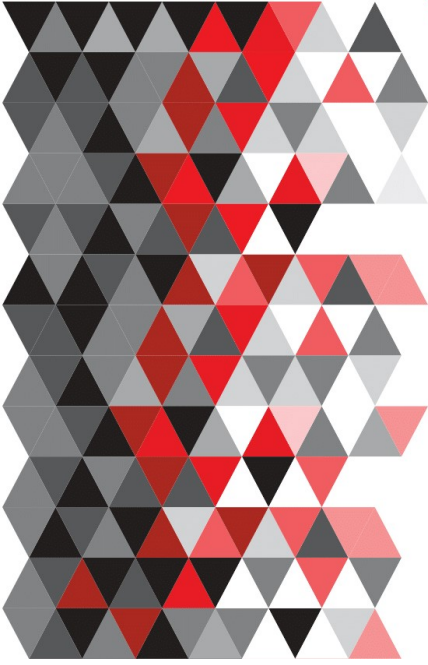
Per tant, aquesta tesi intenta entendre millor quines són les fonts de Big Data que resulten més útils a l'hora de bregar amb variables turístiques i a més proposar millores metodològiques perquè aquestes fonts es puguin aplicar al camp del turisme rural, i més concretament, a la predicció de turistes.

En aquesta tesi es presenten diversos avanços en aquest aspecte: Primer, una classificació de fonts de dades que genera tot turista durant el seu procés turístic i que componen la seua empremta digital. Després, respecte a aquesta classificació, es tria Google Trends com la font més adequada per a ajudar a predir la demanda turística, però es troben problemes de precisió, que són demostrats i exemplificats. Més endavant, es demostra com es genera aquest error de precisió a través del procés de mostreig de GT i es proposen solucions per a alleujar aquest error, a saber, obtenint més extraccions i utilitzant la seua mitjana. Finalment, aquest mètode es posa a prova per a la predicció de pernотacions mensuals en allotjaments de turisme rural a Espanya.

En resum, la contribució que aquesta tesi pretén fer, és aportar una major comprensió de les fonts de Big Data i ajudar a generar bones pràctiques en l'ús de les mateixes perquè es puguin aplicar a la predicció de variables reals en el turisme rural, de manera que agilitze i millore la presa de decisions dels agents econòmics.







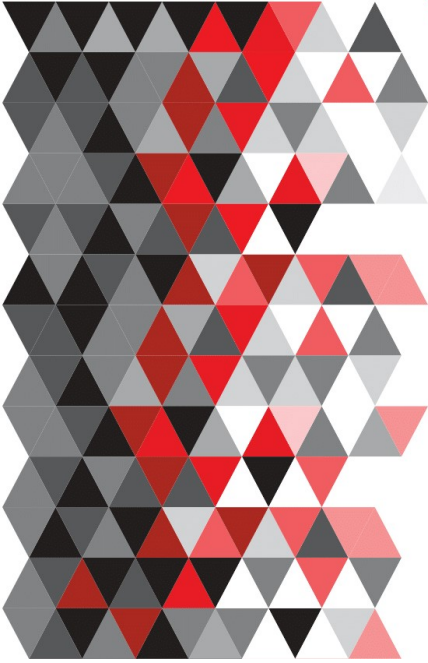
# Table of Contents

<b>List of Figures</b> .....	<b>5</b>
<b>List of Tables</b> .....	<b>7</b>
<b>1 Introduction</b> .....	<b>9</b>
1.1 <b>Internet, Big Data, Digital Footprint and Economic Indicators</b>	<b>9</b>
1.1.1 From Big Data to Digital Footprint .....	9
1.1.2 Tourism and its forecasting .....	12
1.1.3 Rural Tourism .....	13
1.1.4 Big Data and its applications to Tourism .....	14
1.2 <b>Purpose, Objectives and Hypotheses</b>	<b>17</b>
1.3 <b>Structure</b>	<b>17</b>
<b>2 Digital footprint for tourism research</b> .....	<b>21</b>
2.1 <b>Introduction</b>	<b>22</b>
2.2 <b>A Purchase-Consumption System for Tourism</b>	<b>23</b>
2.3 <b>Digital Footprint Sources in Travel and Tourism</b>	<b>25</b>
2.3.1 Internet Sources .....	25
2.3.2 Non-Internet Sources .....	29

<b>2.4</b>	<b>Digital Footprint Sources by Stage in the PCS Model</b>	<b>30</b>
2.4.1	Stage 1: Pre-trip . . . . .	30
2.4.2	Stage 2: During-trip . . . . .	32
2.4.3	Stage 3: Post-trip . . . . .	35
<b>2.5</b>	<b>Discussion</b>	<b>36</b>
<b>2.6</b>	<b>Conclusions</b>	<b>42</b>
<b>3</b>	<b>Is Google Trends a Quality Data Source? . . . . .</b>	<b>45</b>
<b>3.1</b>	<b>Introduction</b>	<b>46</b>
<b>3.2</b>	<b>Google Trends</b>	<b>46</b>
<b>3.3</b>	<b>Quality of Data Sources</b>	<b>47</b>
<b>3.4</b>	<b>Empirical Evidence</b>	<b>48</b>
<b>3.5</b>	<b>Conclusions</b>	<b>51</b>
<b>4</b>	<b>Addressing Google Trends inconsistencies . . . . .</b>	<b>53</b>
<b>4.1</b>	<b>Introduction</b>	<b>54</b>
<b>4.2</b>	<b>Related Work</b>	<b>55</b>
<b>4.3</b>	<b>Google Trends sampling</b>	<b>58</b>
<b>4.4</b>	<b>Simulating GT sampling</b>	<b>60</b>
4.4.1	Modeling . . . . .	60
4.4.2	Scenarios . . . . .	61
4.4.3	Simulation results . . . . .	61
<b>4.5</b>	<b>Alleviating Google Trends inconsistencies</b>	<b>66</b>
4.5.1	A Measure for Popularity . . . . .	66
4.5.2	A Measure for Inconsistency . . . . .	67
4.5.3	Reducing inconsistencies by averaging extractions . . . . .	67
4.5.4	Empirical validation . . . . .	69
<b>4.6</b>	<b>Conclusions</b>	<b>70</b>

<b>5</b>	<b>Can Google Trends predict rural tourism? The case of Spain</b> .....	<b>75</b>
5.1	Introduction	75
5.2	Literature Review	76
5.3	Methodology	78
5.3.1	Data .....	78
5.3.2	Models .....	79
5.3.3	Estimation results .....	80
5.4	Forecasting Results	82
5.5	Conclusions	84
<b>6</b>	<b>Conclusions</b> .....	<b>87</b>
6.1	Main contributions	87
6.2	Implications	88
6.3	Limitations	89
6.4	Future Work	90
	<b>Bibliography</b> .....	<b>93</b>





# List of Figures

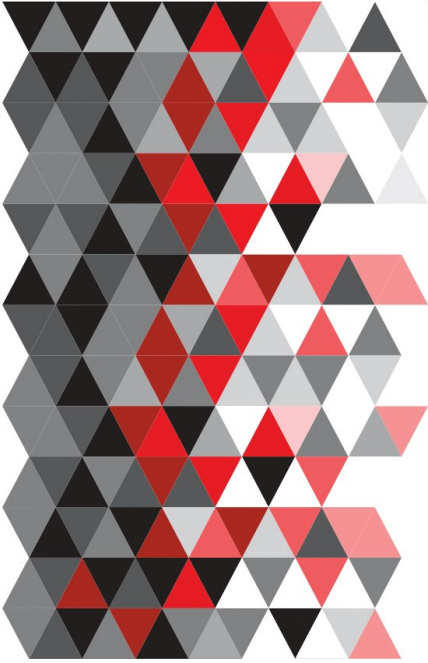
- 1.1 Structure of the thesis, hypotheses and objectives. . . . . 19
  
- 2.1 The Purchase-Consumption System model for travel and tourism, adapted from Woodside and King (2001), which includes three stages: pre-trip, during-trip, and post-trip, with potential variables affecting traveler’s choices. . . . . 24
- 2.2 Evolution of the use of Digital Footprint sources for analyzing tourist behavior, comparing the use of internet to non-internet sources. . . . . 39
- 2.3 Evolution of the use of Digital Footprint sources for analyzing tourist behavior, comparing Social Media sources to non-Social Media sources. . . . . 40
- 2.4 Evolution of the use of Digital Footprint sources for analyzing tourist behavior by stages in the Purchase-Consumption System model. . . . . 41
  
- 3.1 GT reports of searches for four Austrian cities collected at three different dates. 50
  
- 4.1 Process GT follows to compute an SVI time series. The upper part represents the general process followed by the data. The lower part gives a simple example of how the sampling affects the computation of the SVI for a term. The example assumes that the GT report for the term  $a$  from  $t_1$  to  $t_3$  is requested. . . . . 59
- 4.2 Actual SVI for a Seasonal-dominant pattern (a) and a Trend-dominant pattern (b) 62

4.3 Simulation of GT data generation process for a term with seasonal-dominant pattern. Each individual extraction is shown in light blue. The darker line represents the average of all extractions.  $r$  and  $r_s$  are the Pearson and Spearman Correlation coefficients respectively between the dark line and the actual values. . . . . 64

4.4 Simulation of GT data generation process for a term with trend-dominant pattern. Each individual extraction is shown in light blue. The darker line represents the average of all extractions.  $r$  and  $r_s$  are the Pearson and Spearman Correlation coefficients respectively between the dark line and the actual values. . . . . 65

4.5 Relationship between MAPE and the number of averaged extractions (log-log scale) under six different scenarios: two search patterns (seasonal and trend-dominant) combined with three popularity levels (High, red lines; Medium, blue lines; Low, orange lines) . 67

4.6 Comparison of empirical Mean Absolute Percentage Error (MAPE) with theoretical values from Equation 4.8 for selected Google Trends terms. Red lines indicate empirical MAPE for a single averaging of GT series, while shaded regions represent 95% confidence intervals determined through bootstrapping. Each subfigure corresponds to a specific term and search location, together with its associated mean standard deviation ( $\bar{s}$ ). . . . . 71



## List of Tables

2.1	Matrix representing the association between the sources of Digital Footprint and the stage in the PCS model where they have been utilized, as observed in the analyzed literature. The symbol 'X' indicates that some papers were found for a given combination. . . .	37
3.1	Google Trends parameters in the experimental setting. Four searches, one per each search term, were explored. . . . .	49
3.2	Pearson Correlation coefficient of the GT data on February 4, 2021 with GT data returned on different dates. . . . .	51
3.3	Spearman Correlation coefficient of the GT data on February 4, 2021 with GT data returned on different dates. . . . .	51
3.4	Difference in forecast arrivals by retrieving GT data on different days and for different forecasting horizons ( $h$ ). . . . .	52
4.1	Sets of parameters used in the simulation. . . . .	62
4.2	Summary of Google Trends terms extracted from the literature with their respective periods, geography, mean standard deviation ( $\bar{\sigma}$ ). . . . .	69
4.3	Summary of Google Trends terms extracted from the literature with their mean standard deviation ( $\bar{\sigma}$ ) and the necessary amount of extractions to obtain a 1% MAPE according to Equation 4.9. . . . .	69
4.4	Models tested for the relationship between term popularity and MAPE of the SVI 72	
4.5	Mean of the standard deviation ( $\bar{\sigma}$ ) of the simulation scenarios. . . . .	73

---

4.6	Mean of the standard deviation ( $\bar{s}$ ) of 30 different extractions for a selection of GT queries used in the literature. . . . .	74
5.1	Search term and category combinations and their correlation to the dependent variable. . . . .	79
5.2	Necessary extractions to obtain a 1% MAPE for each search term. . . . .	80
5.3	In-sample fit model estimations . . . . .	81
5.4	Out of sample forecasting, forecast horizons 1, 2, 3, 6 and 12. . . . .	83





# 1. Introduction

## 1.1 Internet, Big Data, Digital Footprint and Economic Indicators

In the last decades, the popularity of the Internet has massively increased across all spheres of society, which has caused all economic agents to produce huge amounts of data. Moreover, the advances in the fields of analytics and computation from the last few years have allowed for the storage, processing, treatment and analysis of these vast quantities of data, which are commonly referred to as 'Big Data'. The combination of these two trends now provides the opportunity to trace users' online activity across all data repositories, also known as their 'Digital Footprint' in a very detailed manner. This information can be really helpful in understanding the social and economic behavior of these users, which in turn, can potentially be transformed into economic indicators which can provide fast and accurate results for planning and decision making in both the public and the private sector. Although research in these areas has advanced recently, questions remain about what is the best way to apply these data to obtain reliable and accurate results, and therefore, research which deals with methodologies for the better use of 'Big Data' becomes nowadays as relevant as it has ever been.

### 1.1.1 From Big Data to Digital Footprint

Even though the term 'Big Data' was first coined in 1997 (Cox and Ellsworth, 1997), it does not have a clear homogeneous definition. According to De Mauro et al. (2016) the definitions in the literature can be classified in to one of four types.

The first type of definitions are those which deal with the characteristics of Big Data.

In this sense, Laney (2001) focuses on the volume, velocity and variety of the data as the three main features of Big Data and which have since been adopted in the literature (Ghasemaghaei and Calic, 2020; Bydon et al., 2020; Chang and Grady, 2019; Vogel et al., 2019). As the field has evolved, authors have added different features to this first definition such as veracity (Schroeck et al., 2012) or value (Dijcks, 2013). In fact, some authors such as (Kitchin and McArdle, 2016) have gone on so far as to study what are the ontological characteristics of Big Data. The second type refers to definitions which deal with the technological requirements necessary for the handling of Big Data, which might be associated with computing power or the necessary architecture to deal with the processing and handling of Big Data (Bydon et al., 2020; Vogel et al., 2019). Other definitions suggest that data can only be considered Big Data as long as certain thresholds are met. In this vein, Dumbill (2013) argues that we may consider data to be Big Data once an alternative way of processing data becomes necessary due to the inability of conventional databases to handle such data (Chang and Grady, 2019; Favaretto et al., 2020). Finally, the fourth type of definitions in De Mauro et al. (2016) are those which deal with the Social Impact of Big Data. Here, the definitions of Boyd and Crawford (2013) or Mayer-Schönberger and Cukier (2013) might be included. The former define Big Data as a phenomenon with a cultural, technological and scholarly dimension, while the latter explains the impact of 'Big Data' through the changes it produces in the way in which data is analyzed.

Despite of its lack of a clear definition, Big Data and or the methods of 'Big Data' are used today by most economic agents because data can be accessed much faster and in bigger volumes than ever before (Blazquez and Domenech, 2018), which means that its applications are relevant in plenty of economic sectors such as Finance (Hasan et al., 2020), Engineering (Deepa et al., 2022) or Tourism (Xiang and Fesenmaier, 2017). This is specially true in an online setting where the owners of webpages store every click and every possible trace of activity from each of its visitors, which allows them to obtain the necessary data to make more informed decisions in the future (Askitas and Zimmermann, 2015; Sestino et al., 2020; Lutfi et al., 2023). However, this phenomenon is not exclusive to direct Internet-based activity, as in modern society most of the daily transactions such as paying for groceries or using the public transport card are digitalised and therefore also leave an important trace of data. The collection of these type of data occurs through the processes which can be classified as part of the Internet of Things (IoT).

IoT is based on the idea that *'the connection of physical things to the Internet makes it possible to access remote sensor data and to control the physical world from a distance'*

(Kopetz and Steiner, 2022). IoT makes use of 'smart objects' which are physical systems connected to the Internet. In this vein, RFID technology as well as Bluetooth or Mobile Roaming, among others, fit this definition. IoT is also useful in scientific research as it allows to follow the activities and movement patterns of individuals, and finds applications in areas such as tourism (Qian et al., 2021), industrial processing (Khanna and Kaur, 2020), healthcare (Singh et al., 2020) and even agriculture (Kassim, 2020).

Big Data repositories generate huge quantities of valuable information about the daily activities of individuals, and the combination of all these information from the different data repositories creates a 'Digital Footprint' of each individual. Digital Footprint can be defined as *'a high dimensional and constantly growing space characterized by digital transactions, augmented by surveillance, and influenced by associations and patterns through space and time'* (Weaver and Gahegan, 2007). This means that the Digital Footprint of each individual currently encompasses a big portion their daily lives and for that reason, its study can improve the understanding of economic and social behavior and therefore improve the decision making of both public and private economic agents.

So much so, that the European Statistical System (ESS) has put forward a series of projects dedicated to the study of Digital Footprint in different ways:

- ESSnet Big Data I is a project dedicated to integrate certain sources of Big Data into the process of producing official statistics. The data collected in the project is added to official statistics to create new estimates in statistics which can help study of the Digital Footprint of individuals. In this case, the data is collected from five different Big Data sources. The first source of data is the webscraping of job portals in order to better understand the job market. Secondly, data is also collected from webscraping of enterprises webpages. Thirdly, data from smart electricity meters is also collected with the purpose of generating statistics about energy consumption or housing. Then, ship positioning data is also retrieved in order to generate statistics about traffic as well as pollution. Finally, mobile phone data is also retrieved to improve statistical estimates.
- ESSnet Big Data II is a follow-up project from ESSnet Big Data I which is focused on the production of official statistics in topics such as tourism or transportation through the use of mobile network data (Oancea et al., 2019). The main objective of the project is to generate a framework which can help produce these statistics. This is done through the development of a simulation model that allows to compute certain statistics such as the location of a mobile device or the movement pattern of

population.

- The Trusted Smart Statistics - Web Intelligence Network project serves as a continuation from ESSnet Big Data I and ESSnet Big Data II and its focus is to establish Web Intelligente Network (WIN) as a tool to facilitate the integration of web data into official statistics at the ESS level. Through this implementation, official statistics can be produced much faster and in a standardized manner across all National Statistics Institutes.

One of the main applications of Big Data is the research in tourism, which is often employed to deepen the understanding of the Digital Footprint of individuals and the potential insight that can be gained from its tracking. Therefore, the prevalence of these projects show the importance of developing methods to understand not only tourism as a whole, but also the major role that the different sources of Big Data play in the development of such methods.

### 1.1.2 Tourism and its forecasting

In order to conceive a clear definition of tourism, the concepts of traveller and visitor need to be addressed first. The UNWTO (2010) defines a traveller as '*someone who moves between different geographic locations for any purpose and any duration*'. Then, visitors are travellers whose main destination is outside of their usual environment, who travel for less than a year, and for reasons other than employment in the destination. If a trip meets the aforementioned criteria, then the individual can be considered a visitor. Finally, Tourism can finally be defined as '*the activities of visitors*' (UNWTO, 2010).

Before the COVID-19 pandemic, tourism represented, on average, 4.4% of GDP and 6.9% of employment of OECD countries (OECD, 2022), and though it is still recovering today, UNWTO (2023c) reports that the pre-pandemic levels of World Tourism had almost been recovered by the end of 2023. Therefore, understanding tourist behavior becomes hugely relevant, given the potential impact that it can have in an economy. For that reason, academics have attempted to understand what are the key factors which make a destination more desirable over the years, so that management and policy makers can make better decisions in the planning of tourism. These factors might range from the geographical distance between origin and destination (Tavares and Leitão, 2016), to economic factors such as the relative prices between origin and destination (Muryani et al., 2020) to the cultural resources of the destination (Salinas Fernández et al., 2020; Noolan, 2023), or even the air transport infrastructures or the ICT readiness (Salinas Fernández et al., 2020).

In fact, Gidebo (2021) find that the factors which make a destination more desirable are different from the perspective of the country of origin than they are from the perspective of the destinations. In this sense, in the country of origin the relevant factors for the tourist are those related to demand such as cost of travel, price or marketing. When looking in to the destination, tourists focus on factors related to supply such as accessibility or tourism infrastructure.

However, there is another component that can be really helpful in tourism planning, and that is the prediction of tourism flows. Attempting to correctly predict and forecast tourist flows has become increasingly popular due to the rise in popularity of Big Data. In fact, the aforementioned ESS projects have dedicated workpackages to this very topic. Particularly, the ESSnet Big Data I project contains a workpacakge dedicated to the application of a combination of Big Data sources and existing official statistics to the area of tourism among others. Moreover, it is also detailed in ESSnet Big Data II how the development of a framework for mobile network data could potentially be applied to tourism statistics. Finally, the Trusted Smart Statistics - Web Intelligence Network project also includes the promotion of seminars where participants are taught how to apply WIN technology to process tourism data.

In essence, tourism is one of the most influential economic sectors in most economies, and therefore understanding and forecasting touristic behavior is essential in tourism planning.

### 1.1.3 Rural Tourism

UNWTO (2019) define rural tourism as '*a type of tourism activity in which the visitor's experience is related to a wide range of products generally linked to nature-based activities, agriculture, rural lifestyle / culture, angling and sightseeing*'. Moreover, those activities have to take place in destinations where the population density is low, where landscape is mostly related to agriculture and or forestry and where there is a traditional lifestyle. Moreover, Rosalina et al. (2021) argue that analyzing other definitions of rural tourism reveals four main features which are: location, sustainable development, community-based features and experiences.

Rural tourism has emerged in the last years, and specially after the COVID-19 pandemic, as a popular and alternative way of travelling. This is also due to the rise of the Internet and online booking platforms, which has facilitated the access and appeal of rural tourism. In fact, by 2021 rural areas comprised 43.8% of accommodation beds and contributed to 37%

of overnight stays in the European Union (EPRS, 2023).

Rural tourism can help the development of rural economies by providing new sources of income (Guaita Martínez et al., 2019; Wijjayanti et al., 2020). The impact that rural tourism can have in an economy includes the stimulation of economic growth or the improvement of living standards in local communities among others (Wilson et al., 2001; Liu et al., 2023). Often times, this growth is fostered by the collaboration between local actors, which also ensures long-term sustainability of touristic activities (Kumar et al., 2022). However, the impact of rural tourism goes beyond pure economic benefits. It also encompasses the sociocultural and well-being of rural areas. The sociocultural benefits include the promotion of a community identity or of cultural heritage among others (Lane and Kastenholz, 2015). Moreover, rural tourism can serve as a mean to improve conservation efforts but also to promote sustainable habits (Rosalina et al., 2021; Jepson and Sharpley, 2015) and is therefore a huge catalyst in maintaining sustainable rural development (Tang and Xu, 2023). Therefore, it should come as no surprise that over 50% of the member states which participated in UNWTO's last survey about rural tourism consider rural tourism to be a direct priority of their country (UNWTO, 2023b).

Because of the relevancy of rural tourism, scientific literature has covered at length what can be done to unlock the potential of rural destinations by means of understanding how to foster rural tourism. In this sense, Kumar et al. (2022) find that the development of rural tourism in India has five main drivers: the development of infrastructure, the environmental conscience, the support of the local government and the community, the availability of funds from the public sector and the participation from the private sector. Other authors such as Jepson and Sharpley (2015) find that visitors to Lake District (UK) feel a sense of belonging to the area, and that the participation of tourists in specific forms of rural tourism provokes deeper emotional feelings in customers. Therefore, the authors propose promoting a 'sense of place' as a strategy to improve customer loyalty. Similarly, Rid et al. (2014) are able to identify four types of tourists in The Gambia according to their different motivations: heritage and nature seekers; multi-experience seekers; beach and multi-experience seekers and sun and beach seekers. This identification helps in creating different paths for development so that rural tourism initiatives can be targeted correctly.

#### 1.1.4 Big Data and its applications to Tourism

Technological advances in recent years have enabled the emergence of new data sources and with this, the storage of large amounts of data or Big Data has become increasingly

important (Blazquez and Domenech, 2018). More and more scientific studies are using these Big Data sources to try to improve understanding in various scientific fields such as healthcare (Khanra et al., 2020; Li et al., 2021a), finance (Goldstein et al., 2021; Bellini et al., 2020), economics (Awan et al., 2021b; Del Giudice et al., 2021) or marketing (Buhalis and Volchek, 2021; Brewis et al., 2023).

One of the main challenges highlighted by this new trend in the literature is the difficulty in understanding the correct application of these data sources. This challenge is partly attributed to the novelty of the sources, which raises concerns about their quality. A primary area of uncertainty is the usefulness of these sources, specifically their capability to aid researchers in predicting real-world variables.

For example, Awan et al. (2021a) improve the prediction of financial stocks by including data from Twitter among others, in Machine Learning Models. Similarly, Önder et al. (2020) use data from Facebook and Google Trends to improve the forecasting of tourist demand to four Austrian Cities. Even in the field of meteorological forecasting, Zenkner and Navarro-Martinez (2023) include Big Data retrieved from two observation stations in London which are then used to produce accurate temperature forecasts in the city.

Yet, even if the results seem promising, there is no consensual way to measure the quality of this data, and therefore it becomes really challenging for researchers to be sure about the consistency and reproducibility of their results. Furthermore, even if data quality can be measured, there is no knowledge of whether if these data sources are good enough or not.

Moreover, the fact that there is so much data and from such a wide range of sources raises the question about how to solve the methodological issues which might arise from every Big Data source. Authors such as Batini et al. (2015); Wang et al. (2023) propose a framework for the evaluation of Big Data based on certain qualities such as completeness or accuracy as a solution. Others, such as Liu et al. (2016) provide a set of good practices which could help alleviate the issues faced when working with Big Data. However, not all authors deal with data quality. Instead, other authors try to solve the issues for their own specific sector (Wong and Wong, 2020; Eichenauer et al., 2022) by adapting to the characteristics of the data.

Therefore, the next logical steps are to find if these proposed solutions actually work and if the quality of Big Data is improving, for which, once again, there is no consensual solution of how to do so.

As authors argue that the quality of data is still an issue that hinders the potential of Big

Data itself (Hossen et al., 2020; Saleh et al., 2023), another challenge which needs to be tackled is the development of methodologies that allow the better use of Big Data sources as well as its expansion among the literature.

In tourism, the usefulness of these new data sources lies in the fact that they can help to understand the behavior of tourists, from their spatiotemporal patterns to which attractions and activities are the most popular in the destination, and therefore can help in public policy decision making. In this particular field many of these sources are employed successfully, as it is the case for Mobile Roaming Data (Raun et al., 2016; Qian et al., 2021), Google Trends (GT) (Dergiades et al., 2018; Bangwayo-Skeete and Skeete, 2015), Facebook (Gunter et al., 2019; Kwok et al., 2022) or even Instagram (Filieri et al., 2021; Palazzo et al., 2021).

However, this raises a new challenge, which is to understand which sources are relevant for tourism, and furthermore, for those which are relevant, what is the specific purpose behind its use. This is a relevant challenge given that the use of Big Data sources for tourism is multifaceted in its purpose: some authors successfully improve forecasting of tourist arrivals through the use of GT (Gunter et al., 2019; Rivera, 2016) or Twitter data (Carvache-Franco et al., 2022; Kim et al., 2021), while others focus on the tracking of usual touristic routes through Instagram posts (Ma et al., 2020) or GPS devices (Zheng et al., 2019), or even the choice of transport to destinations through Mobile Roaming Data (Qian et al., 2021). In addition, Big Data sources in tourism are multifaceted as well in regards to the point of the trip in which they are used: Some authors study how to influence the tourists to make a trip through different Social Media platforms such as Instagram (Aramendia-Muneta et al., 2021) or Facebook (Villamediana et al., 2019) while others focus on the activities chosen by tourists while at the destination through the study of Consumer Card data (Scuderi and Dalle Nogare, 2018). Finally, Big Data can also be used to monitor the post trip behavior and the intentions to return of a tourist through Online Reviews (Zhu et al., 2020).

Finally, given the importance of rural tourism in modern economies as well as its potential benefits (Rosalina et al., 2021; Wilson et al., 2001), it becomes hugely relevant to understand how rural tourists behave as well. However, another challenge is whether or not these Big Data sources can be applied successfully to rural tourism research as well. Because rural tourism deals with a much smaller scale of tourists at least in a destination to destination basis, Big Data sources might not be as suitable for this type of tourism.



## 1.2 Purpose, Objectives and Hypotheses

The purpose of this thesis is to apply Big Data to the rural tourism sector to improve the prediction of related variables and to help establish a set of good practices regarding the handling and processing of Big Data. Therefore, in this thesis the general objective is to obtain a broader understanding of the tourism sector, focusing on the rural side of it and also to explore how to better apply Big Data and forecasting techniques to predict variables of interest in rural tourism. Hence, the proposed hypotheses are the following:

- **Hypothesis 1:** Big Data sources are heterogeneous and each of them has different applications but also presents different methodological issues.
- **Hypothesis 2:** GT has quality issues if not treated properly.
- **Hypothesis 3:** GT can help improve the forecasting of variables related to rural tourism.

Each hypotheses is materialized through different objectives, which are tackled in Chapters 2 through 5 in this thesis. The objectives are:

- **Objective 1:** Study the relevance of Big Data sources in the tourism sector.
- **Objective 2:** Study quality aspects of GT.
- **Objective 3:** Develop methods to ensure the quality of GT.
- **Objective 4:** Evaluate if GT can improve forecasting of rural tourism demand.

## 1.3 Structure

The structure of the thesis is divided in six chapters. This first chapter serves as introduction, while chapters 2 through 5 are adaptations from research papers which have been published and/or submitted to various international peer-reviewed scientific journals. Finally, chapter 6 presents the conclusions of this thesis.

Chapter 2 is an adapted version of the research paper titled '*Digital footprint for tourism research*'. This paper proposes a classification of data sources which make up the Digital Footprint of a tourist based on a Purchase - Consumption System (PCS) model, which divides the touristic process in to three main stages: pre-trip, during-trip, post-trip. Through this classification, certain trends in the tourism literature can be established. This research paper targets Objective 1 in the thesis.

Chapter 3 is adapted from the research paper '*Is Google Trends a quality data source?*' This paper is published in the scientific journal '*Applied Economics Letters*'. This paper studies the statistical quality of GT data through Karr's criteria. From this analysis, a

practical example based on GT data is presented as a study of how GT data performs in improving the accuracy of forecasting of tourist demand in four Austrian cities.

Chapter 4 presents the research paper '*Addressing Google Trends inconsistencies*'. This paper was published in the scientific journal '*Technological Forecasting and Social Change*'. In this paper, a simulation of the data generating process of GT is proposed. Through this simulation, the sampling methodology of GT is reproduced so that analysis about its inconsistencies can be performed. From this analysis, a method is tested and proposed in order to deal with these inconsistencies. In this research paper, Objective 3 is dealt with.

Chapter 5 presents an adapted version of the research paper '*Can Google Trends predict rural tourism? The case of Spain*'. This paper tests the forecasting capabilities of GT data for rural tourism in the Spanish market. To do so, time series models including GT data as a predictor are compared to classical benchmark time series models in terms of forecasting accuracy. In this research paper, Objective 4 is tackled.

Finally, Chapter 6 presents the conclusions of the thesis. In it, the main contributions are discussed, as well as the implications of the results obtained and the limitations faced. In the end, lines for future work are presented. In Figure 1.1 the structure of the thesis is presented graphically.

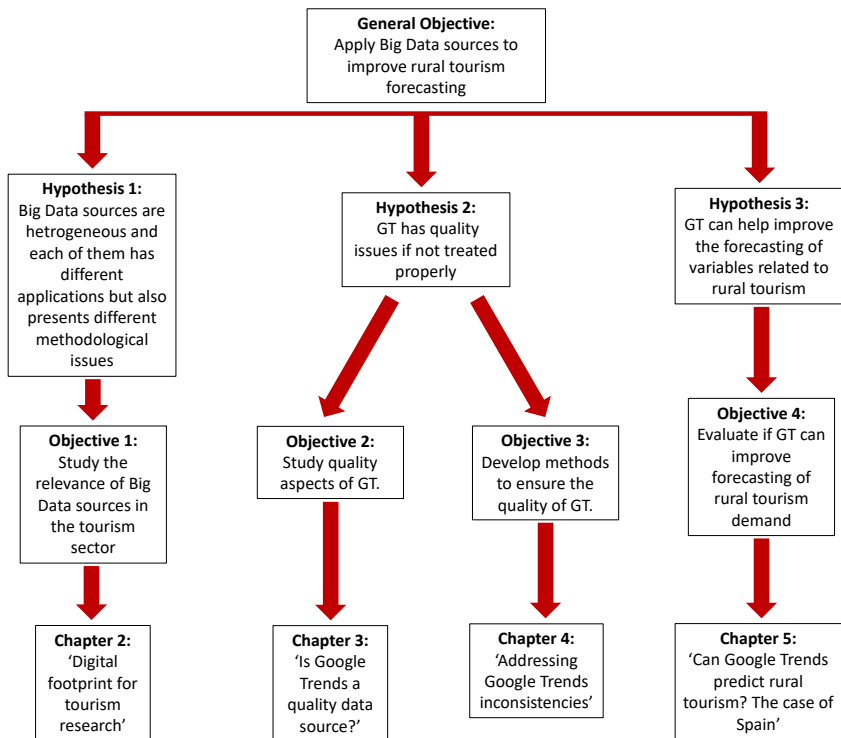


Figure 1.1: Structure of the thesis, hypotheses and objectives.





## 2. Digital footprint for tourism research

*Dedicado a Francisco Ferrer Falcón*

### Digital footprint for tourism research

#### Abstract

**Purpose** - This paper aims to present a comprehensive analysis of Digital Footprint sources used to understand and predict the main variables associated with tourist behavior.

**Design/methodology/approach** Utilizing the Purchase - Consumption System (PCS) model specific to leisure travel behavior, Digital Footprint sources are classified into three stages: pre-trip, during-trip, and post-trip. A literature review is conducted to map the use of Digital Footprint sources in tourism research and establish relationships between different data sources.

**Findings** - The research reveals that internet sources, particularly Social Media, are predominantly used in examining tourist decisions made at the destination. However, the paper identifies underexplored potential in sources like Search Engine Data and Twitter data. The study also underscores the trend of growing interest in studying the post-trip stage, facilitated by the availability of Online Review Data and advancements in natural language processing techniques.

**Originality/value** - This study presents a novel classification of Digital Footprint sources in tourism research, and provides a novel perspective on the potential predictors of tourists' choices, drawing attention to the underutilized sources and new research possibilities. This contribution serves as a guide for researchers to select appropriate data sources and provides a foundation for future investigations in the rapidly evolving field of Digital Footprint analysis in tourism.

**Keywords**— Big Data; PCS; Tourism; Data Sources; Classification

## 2.1 Introduction

Tourism is a rapidly growing industry, as evidenced by the record-breaking 1.46 billion international tourist arrivals in 2019, as reported by the United Nations World Tourism Organization. Furthermore, despite the setback caused by the pandemic, the industry is now rebounding swiftly. This sector plays a significant role in the global economy, contributing to over 10% of the world's GDP and creating millions of employment opportunities (UNWTO, 2021).

In recent years, the increasing availability of digital data has revolutionized the way in which tourist behavior is studied. Technological advances have made it possible to record and analyze every click and interaction that individuals have on electronic devices. This is all possible thanks to the rise of Big Data, which allows for the storage of enormous quantities of these interactions, constituting a Digital Footprint that reveals traits of individual and social behavior. This phenomenon has opened up new research possibilities, as researchers today have access to much more data at a much faster pace than ever before (Blazquez and Domenech, 2018; Girardin et al., 2008). This allows for the analysis of the dynamics of human behavior in detail, presenting opportunities for researchers to gain insights into how individuals interact with technology and with each other.

This Digital Footprint is particularly relevant in tourism research, as it allows to understand the spatiotemporal patterns of tourist decisions, and therefore, it can be used to reveal the preferences of tourists, forecast touristic variables and even plan touristic policies.

In the pre-Big Data era, tourist data were primarily collected through frontier counts, registration at accommodation establishments or sample surveys (Witt and Witt, 1995). However, the rise of Big Data has allowed for newer, more comprehensive approaches to studying tourism. Nowadays, it is possible to study the online image of a city by collecting the opinions of thousands of tourists (Marine-Roig and Clavé, 2015) or to capture almost every trace of data that a tourist leaves behind with a *tourist kit* (Angeloni, 2016).

However, the usefulness of Big Data in tourism research depends on the specific source of the data, as each source can reveal information about different variables related to tourist behavior. To classify these sources, the Purchase-Consumption System (PCS) model applied to Travel and Leisure by Woodside and King (2001) is employed, as it provides a comprehensive representation of the decision-making process that a tourist goes through during all stages of a trip. By using this framework the main sources of tourists' Digital Footprints are connected to their cognitive process.

The main contribution of this paper is threefold: Firstly, Digital Footprint sources are reviewed, including data access methods, variables within the data source, and examples of applications. Secondly, the classification of sources that constitute the Digital Footprint according to the PCS model applied to leisure and tourism from Woodside and King (2001) is presented. This classification aims to identify which data sources are best suited to study the variable of interest. Finally, the potential for complementarity between different data sources is discussed, as some sources have been used with a common purpose and even compared to determine which source works best, as in the

case of Twitter, Instagram, and Flickr (Tenkanen et al., 2017).

The rest of the article is structured as follows: Section 2.2 introduces the purchase-consumption system model for travel and leisure. Section 2.3 reviews the Digital Footprint sources considered in the literature. Section 2.4 presents the mapping between Digital Footprint sources and stages in the purchase-consumption system model. Section 2.5 discusses the main relationships, trends, and opportunities of Digital Footprint sources. Finally, Section 2.6 draws some concluding remarks.

## 2.2 A Purchase-Consumption System for Tourism

Our work is grounded in the updated model of travel and tourism PCS model by Woodside and King (2001). In general, a PCS model represents *'the sequence of mental and observable steps that a consumer undertakes to buy and use several products for which some of the products purchased leads to a purchase sequence involving other products'*. The Woodside and King (2001) model is a revised version of the PCS model applied to travel and tourism, which identifies potential variables that affect travelers' choices and classifies them into three stages: pre-trip, during-trip and post-trip. During all stages, these variables are interactive and any of them may impact upon another one. The three stages and the related variables are represented in Figure 2.1.

Stage 1 is the pre-trip stage and deals with the decisions affecting the thinking and planning processes that occur before going on a trip. Several theoretical constructs such as images, attitudes and perceived risk are relevant in this stage. For example, images of a touristic product are associated to both cognitive and emotional interpretations which are especially important in the early stages of product evaluation.

Stage 1 includes variables that deal with these theoretical constructs and that influence the main decisions that tourists will make later.

These variables relate to outside influences on tourists, such as the family or the marketing exposure, and also to the use and interpretation of information.

Then, these influences create a problem or an opportunity for the tourist and are considered to be effective or not depending on how the customers handle the information they receive.

This idea, represented in Figure 2.1 with the variables five through eight, reflects the tourists' mental process when dealing with external information.

Stage 2, which is where the main decisions for a trip are taken, is divided in three sequential levels. Stage 2.1 covers the choices of destination, activities and attractions, which are considered to be the three most important decisions in the trip, since any of these three variables can be the main reason to make the trip.

Stage 2.2 encompasses decisions on the accommodation, the mode of transportation and the route to the destination, which are usually dependent on the choices made in Stage 2.1. In Stage 2.3, the selection of self-gifts and other non-durable purchases are included, as well as dining out and the

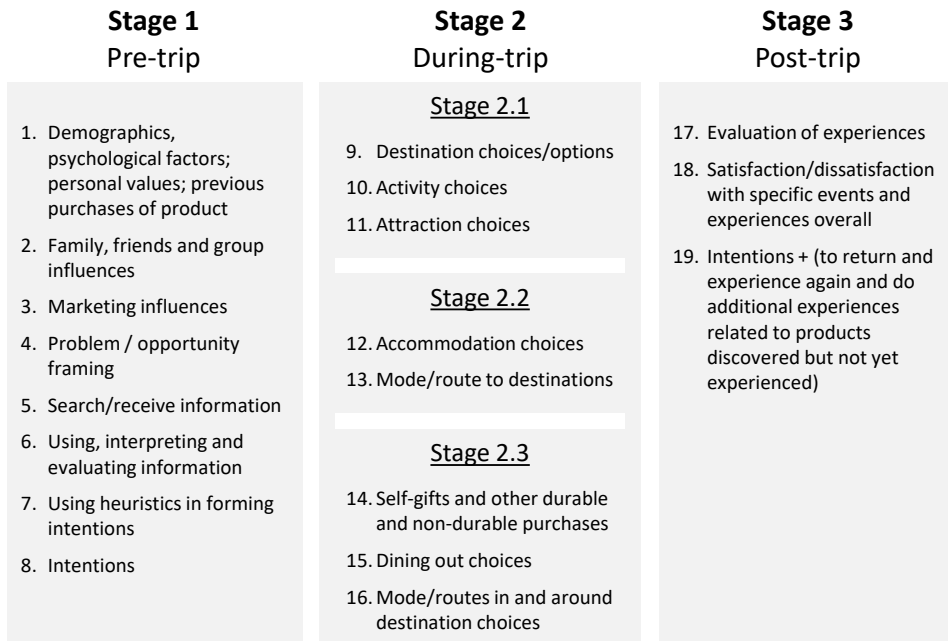


Figure 2.1: The Purchase-Consumption System model for travel and tourism, adapted from Woodside and King (2001), which includes three stages: pre-trip, during-trip, and post-trip, with potential variables affecting traveler’s choices.

modes and routes in and around the destination. These are normally made once the tourist has arrived and settled at the destination, while the previous choices are usually pre-planned.

Finally, Stage 3 refers to the outcomes of the trip: the experiences and the evaluations that occur after the trip (or once some activities have ended).

These evaluations are based on feelings about the trip and can be either positive or negative, but in any case, they will have an impact on future travel choices. It is important to distinguish between the evaluation of experiences and the posterior satisfaction or not with them. The customer first evaluates the experiences either in isolation or as a whole, and then, there will be satisfaction if the expectations are met and dissatisfaction otherwise. Then, the resulting feelings are the basis for returning or not to the same destination.



## 2.3 Digital Footprint Sources in Travel and Tourism

There is a wide range of data sources that can be used to monitor and forecast both tourism flows and tourist behavior. These sources can be categorized into those that pertain to user behavior on the internet and those that originate from other sources. In this section, the sources are analyzed by focusing on the type of data that is commonly retrieved from each source and the methods through which this data is accessed.

### 2.3.1 Internet Sources

Internet sources refer to data that is collected from digital platforms such as websites, Social Media and search engines, and can provide valuable insights into the behavior of users on these platforms, where every click can be traced and recorded (Girardin et al., 2008). This constitutes a huge amount of data, with a significant part of it being publicly available.

There are two main methods to access this type of information. The first one is through the use of Application Programming Interfaces (API). APIs are a combination of definitions and protocols that allow products and services to communicate with each other, making it possible to obtain the necessary data. The second method is Web scrapping, a technique used through different software to gather information from websites. Web scrapping simulates a human being visiting those websites and it helps to index and classify sites into structured data that can be classified and stored.

#### Facebook

is a Social Media platform used by individuals, companies and institutions. In it, they can share text, pictures and other multimedia information with other users, with whom they can interact through comments or likes in their posts or via direct messaging.

The studies primarily rely on information extracted from the Facebook pages of Destination Management Organizations (DMOs), which is subsequently analyzed. Post likes, comments, and shares are the primary sources of data (Gunter et al., 2019; Önder et al., 2020). However, some other studies also obtain other types of information such as the number of weekly posts (Lee et al., 2021), the total number of fans of the DMO page (Mariani et al., 2018), the posting day of the week (Villamediana et al., 2019) or geo-tagged information from pictures uploaded to Facebook (Dalal, 2017).

To access this data, most research studies use the API provided by Facebook (Gunter et al., 2019; Önder et al., 2020; Mariani et al., 2018; Lee et al., 2021; Dalal, 2017), which allows access to part of the published content, since not all of it is public. Other researchers, however, opt to retrieve their data manually (Park et al., 2016a; Villamediana et al., 2019).

### Twitter

is a micro-blogging Social Media platform that allows users and companies to create profiles and interact by sharing their thoughts, opinions and/or photos in posts, namely “tweets”, limited to a maximum of 280 characters. Tweets are public and some of them are geo-tagged.

Most studies that use Twitter as their main data source typically focus on the content of the tweets themselves (Philander and Zhong, 2016; Willson et al., 2021; Park et al., 2016b; Sontayasara et al., 2021; Mishra et al., 2021; Kirilenko and Stepchenkova, 2017; Gulati, 2022; Carvache-Franco et al., 2022). However, several studies also take in to account the geo-tagged information of tweets (Zervoudakis et al., 2021; Stepchenkova et al., 2013; Provenzano et al., 2018; Padilla et al., 2018; Chua et al., 2016; Brandt et al., 2017; Bordogna et al., 2016; Bhatt and Pickering, 2021; Abbasi et al., 2015; Kim et al., 2021). Additionally, some studies use the number of comments, likes and retweets (Kwok et al., 2022) or the posting time of tweets (Lu and Zheng, 2021) as valuable sources of information.

Overall, accessing data from Twitter for research purposes is mainly done through the use of Twitter’s API (Sontayasara et al., 2021; Provenzano et al., 2018; Philander and Zhong, 2016; Park et al., 2016b; Padilla et al., 2018; Chua et al., 2016; Bordogna et al., 2016; Bhatt and Pickering, 2021; Abbasi et al., 2015), which offers free access to a limited amount of tweets. Access to the full tweet archive is a paid service. However, to simplify data collection, some authors opt to use hashtags (Carvache-Franco et al., 2022; Zervoudakis et al., 2021), keywords (Stepchenkova et al., 2013; Kirilenko and Stepchenkova, 2017; Mishra et al., 2021) or a combination of both (Gulati, 2022) as filters. Another method, albeit less common, is web-scraping (Lu and Zheng, 2021; Kim et al., 2021).

Finally, some researchers use commercial third-party services such as Salesforce Social Studio to obtain their data (Willson et al., 2021).

### Instagram

is a free online content application where users and companies can share pictures, videos, reels and stories. Posts can be geo-tagged, and users can interact with each other through comments, likes, and direct messaging.

In the literature, there is a wide range of data types used from Instagram. Several authors focus on the content of the images uploaded to the platform and their attributes such as the colors (Yu and Egger, 2021; Yu et al., 2020) or the image theme (Aramendia-Muneta et al., 2021). Other studies use the attributes of the image combined with other types of information like the geo-tagged data of the posts (Rossi et al., 2018; Paül i Agustí, 2021), engagement-related metrics like the number of likes and comments, or both (Paül i Agustí, 2018; Palazzo et al., 2021). In addition to image-related data, some studies analyze other data from Instagram, such as textual data of posts (Filieri et al., 2021) or the geo-tagged data by itself (Ma et al., 2020).

Unlike other Social Media platforms, Instagram's data is not typically accessed through its API due to its limitations (Rossi et al., 2018; Paül i Agustí, 2018). Instead, other methodologies such as web scrapping are more commonly used. For instance, authors such as Palazzo et al. (2021) and Yu and Egger (2021) use hashtags to filter the data and then scrape it.

However, others authors resort to manual collection. To do so, they search for posts in a specific location and then collect the necessary data (Paül i Agustí, 2018, 2021; Aramendia-Muneta et al., 2021).

Finally, some authors choose to use third-party services such as Picodash (Filieri et al., 2021; Palazzo et al., 2021) or Octoparse to obtain their data (Yu et al., 2020).

### **Flickr**

is an online service where users upload, organize and share their visual content. Users can also add metadata to their photos and videos, including location data through geo-tagging.

The studies that use Flickr mostly focus on a combination of the posting time and the geo-tagged data of images uploaded to the platform (Wood et al., 2013; Girardin et al., 2007; Tenkanen et al., 2017; Popescu and Grefenstette, 2009; Miah et al., 2017). Other studies focus on the content of the images and their attributes, such as the theme of the picture, or the artistic period of the element referenced in the image (Donaire and Galí, 2011).

When using Flickr as a data source, most authors retrieve the data through the API (Wood et al., 2013; Girardin et al., 2007; Tenkanen et al., 2017; Popescu and Grefenstette, 2009; Miah et al., 2017), while other authors prefer to do it manually (Donaire and Galí, 2011).

### **Search Engine data**

The data on the use of search engines is widely employed to understand tourist behavior. The main tool to access search engine data is Google Trends, which gives information on the relative popularity of keywords in Google Search. The data is publicly available as time series.

Authors can access this data by means of a web interface (Jackman and Naitram, 2015; Gunter et al., 2019; Park et al., 2017; Bangwayo-Skeete and Skeete, 2015) or an API, which can be accessed through specific packages for software like R or Python, and in this way, the selected series can be downloaded. Yet, the use of the API is not as common in tourism as it is in other fields such as economics (Barreira et al., 2013; Eichenauer et al., 2022) or health (Kandula and Shaman, 2019).

The data can be obtained with daily, weekly (Bangwayo-Skeete and Skeete, 2015; Havranek and Zeynalov, 2021) or monthly frequencies (Dergiades et al., 2018; Gunter et al., 2019). Moreover, this tool not only monitors keyword popularity, but also search categories (e.g., travel) (Dinis et al., 2017), which represent clusters of related keywords (Cebrián and Domenech, 2023b).

Results can be filtered by geographic area at the country (Park et al., 2017; Rivera, 2016), region (Padhi and Pati, 2017; Siliverstovs and Wochner, 2018; Yang et al., 2015) or city level (Li et al., 2017; Önder, 2017). This allows for the analysis of not only time series, but also cross-sectional and panel

data.

It is worth noting that there is a vast literature using Google Trends as a source to study tourism-related topics. However, due to the lack of consistent reporting of search settings such as frequency or type of access to data, it is difficult to provide a comprehensive classification of how these studies use Google Trends (Jun et al., 2018).

### **Website Traffic statistics**

The Internet traffic that tourism-related websites receive is a valuable source of information, which can be monitored through specific traffic monitoring services, such as Google Analytics. These services provide website owners with time-series data about their visitors, including information on the pages visited, session duration, and visitor location.

Unlike Google Trends, this information is only available to the website owner, who can access the traffic monitoring service via a free plan, or a paid plan which includes more advanced features. Google Analytics is the most frequently used service, with authors utilizing a wide variety of data, such as the number of visitors (Danila and Gaceu, 2009), the source of traffic (Plaza, 2011; Dinis et al., 2016), the number of new and returning visitors (Gunter and Önder, 2016; Plaza, 2011), the average time spent on a page or a session (Dinis et al., 2016; Gunter and Önder, 2016) and the visitor country of origin (Dinis et al., 2016).

Access to this data can be obtained through the user interface or through the API, but most authors choose to access it through the user interface (Danila and Gaceu, 2009; Dinis et al., 2016; Gunter and Önder, 2016; Plaza, 2011).

### **Online Review data**

refers to the reviews and opinions about certain experiences that tourists post on different platforms, which help other potential visitors make their decisions. These reviews are a way for tourists to express their feelings, sentiments and moods.

Most authors focus on the content of the reviews as their main source of data (Marine-Roig and Clavé, 2015; Cheng and Jin, 2019; Kim et al., 2017). Others focus both on the content of the reviews and their ratings (Zhu et al., 2020), or even the ratings of specific attributes of hotels, such as the cleanliness or the quality of the service (Liu et al., 2017). Meanwhile, other authors like Batista E Silva et al. (2018) use the location of the rooms as the main source of data.

These data are normally accessed through web scraping of websites like TripAdvisor (Liu et al., 2017), Booking (Batista E Silva et al., 2018), VirtualTourist (Kim et al., 2017) or even travel blogs (Marine-Roig and Clavé, 2015). Yet, other authors choose to use third-party services like Inside AirBnb to collect the data (Cheng and Jin, 2019; Zhu et al., 2020).

### 2.3.2 Non-Internet Sources

Not all the Digital Footprint sources used for tourism research are related to the internet, yet they are useful for researchers as well. Such sources are important because they rely on different technologies and enable exploration of new dimensions of the Digital Footprint of tourists.

#### Mobile Roaming data

consists of the location coordinates of the mobile phone and is passively collected and stored in the log files of mobile network operators. Due to the widespread use of mobile phones, this data creates a vast amount of spatial information that can be used in research.

Typically, researchers focus on the geographical coordinates of the antenna which collects the signal, the time of the call activity, the country in which the phone is registered, and a device ID number (Ahas et al., 2007, 2008; Qian et al., 2021; Raun et al., 2016; Tiru et al., 2010).

Access to this data is obtained through contracts with private operators (Ahas et al., 2007, 2008; Qian et al., 2021; Raun et al., 2016; Tiru et al., 2010), and its availability is limited due to the competitive nature of the mobile communication industry. Operators prefer to keep information private to keep their competitive edge, and strict measures must be taken to ensure data security and protect the trust of subscribers (Ahas et al., 2008).

#### Consumer Card data

refers to the information obtained from customers who use cards with reward schemes. These cards offer certain benefits to their users when they are used, and data is collected about the consumers and their transactions made. This may include purchases, travel, visited attractions, and restaurant choices.

In studies utilizing consumer card data, customers are identified through the ID of their cards. For transport cards, variables such as the date, departure and arrival stations are recorded, among others (Asakura et al., 2012; Xue et al., 2014). For non-transport consumer cards, some studies focus on the locations where consumers use their card (Zeni et al., 2009; Newing et al., 2014), while others focus on the tourist activities chosen by consumers (Scuderi and Dalle Nogare, 2018).

However, this data is often private, as it belongs to the institution issuing the card, and therefore it requires collaboration between the researchers and the institutions. While institutions are often part of the public administration (Scuderi and Dalle Nogare, 2018; Xue et al., 2014; Asakura et al., 2012; Zeni et al., 2009), data from private parties may also be used (Newing et al., 2014).

#### Specific Gathering Devices data

refer to data collected by devices that are specifically designed for tracking the location of tourists.

The data gathered from these devices usually includes the tourists coordinates and timestamps for each recorded location (Becco et al., 2013; Bfl et al., 2012; McKercher et al., 2012; Shoval and Isaacson, 2007; Sørensen and Sundbo, 2014; Zheng et al., 2017, 2019). In some cases, additional

features, such as the Media Access Control (MAC) address (Yoshimura et al., 2014), which work as an identifier for the device, or the Class of Device (COD) (Delafontaine et al., 2012; Versichele et al., 2012, 2014), are also recorded.

Data from Specific Gathering Devices is often collected through two alternative methods. In the first, tourists are asked to carry GPS receivers and return them at the end of their visit (Shoval and Isaacson, 2007; Sørensen and Sundbo, 2014; Zheng et al., 2017, 2019), activities (Bíl et al., 2012; McKercher et al., 2012), or trip (Becco et al., 2013). In the second, the researchers place technology-specific access points (e.g., Bluetooth) on strategic points of touristic locations and collect the signal of devices that pass through those points (Delafontaine et al., 2012; Versichele et al., 2012, 2014; Yoshimura et al., 2014).

## 2.4 Digital Footprint Sources by Stage in the PCS Model

Once the main sources have been examined, this section classifies them into the stage in the PCS model for which they are relevant according to the analyzed literature. To do so, the variables in the PCS model are analyzed by focusing on the main sources of data used to study the variable, the purpose of the studies which use the source and the methodologies employed in them.

### 2.4.1 Stage 1: Pre-trip

#### **Demographics, psychological factors, personal values, and previous purchases of product**

are explored through various sources. Google Analytics enables the collection of data on website visitors, including the age, gender, country of origin, and more. In this vein, Dinis et al. (2016) study the visitor profile of the Portuguese region of Antalejo using Google Analytics data from the official promotional website.

However, the most common method for obtaining information about the profile of tourists is through the study of their Social Media profiles. This approach has been used in several studies, such as Palazzo et al. (2021), which analyzed the profiles of Instagram ‘influencers’ involved in promoting sustainable tourism. Similarly, Tenkanen et al. (2017) and Wood et al. (2013) examined the visitors’ country of origin to National Parks by analyzing their Flickr profiles.

Other sources, such as Consumer Card Data (Newing et al., 2014) or Mobile Roaming Data, have also been used to study visitors’ country of origin (Ahas et al., 2007, 2008; Raun et al., 2016; Tiru et al., 2010).

Additionally, Social Media platforms like Instagram can be used to explore the psychological factors affecting touristic decision-making. Yu and Egger (2021) studied the relationship between color schemes used in touristic pictures posted on Instagram profiles and user engagement, as measured by the engagement rate (the ratio of likes and comments divided by the user’s followers).

They argued that different color schemes convey different meanings and, therefore, impact human perception.

Most of these studies rely on descriptive analysis (Tenkanen et al., 2017; Wood et al., 2013; Ahas et al., 2007, 2008; Raun et al., 2016; Tiru et al., 2010; Palazzo et al., 2021). However, other methodologies such as machine learning approaches (Yu and Egger, 2021) or the analysis of reports from Google Analytics (Dinis et al., 2016), have also been employed.

### **Marketing Influences**

have been mainly studied through Facebook and Instagram. Studies that use Facebook focus on how effectively touristic organizations promote their destinations (Mariani et al., 2018; Park et al., 2016a; Villamediana et al., 2019; Lee et al., 2021). User engagement is typically used to measure the effectiveness of communications, which is computed through statistics such as the number of likes, comments, shares, and the posting time of the publication, or the number of publications.

Regarding the methodology used, these studies typically obtain a sample of posts from the selected touristic organizations' profiles and collect some engagement-related indicators. These metrics are then analyzed using various methods, including panel data analysis (Mariani et al., 2018), regression analysis (Lee et al., 2021), content analysis (Villamediana et al., 2019; Park et al., 2016a), or network analysis and Pearson correlations (Park et al., 2016a).

Similarly, studies that use Instagram focus on how this platform can be better used to promote touristic destinations. They pay attention to the attributes of the pictures (Aramendia-Muneta et al., 2021), such as the main theme, time of day, and even the colors shown in the picture (Yu et al., 2020). Other authors like Palazzo et al. (2021) focus on the role played by 'influencers' in the promotion of destinations. The influence of these factors is again measured by the number of likes and comments on the selected posts.

To analyze these factors, researchers typically create a sample of posts by selecting different accounts (Aramendia-Muneta et al., 2021; Yu et al., 2020) or hashtags (Palazzo et al., 2021). Then, the information may be analyzed through textual analysis (Palazzo et al., 2021), or content analysis (Aramendia-Muneta et al., 2021; Yu et al., 2020). Finally, the effects are computed with regression analysis (Yu et al., 2020; Aramendia-Muneta et al., 2021).

### **Search / Receive Information**

is commonly studied through Google Analytics, Facebook, or Twitter. For instance, Plaza (2011) explored the online behavior of users visiting a cultural tourism webpage by examining metrics such as the number of pages visited per session, and the percentage of return visitors. On the other hand, Kwok et al. (2022) explored how hospitality companies communicated during the COVID-19 pandemic and how users responded to their messages on Facebook and Twitter.

Regarding the research methods, the former employed regression analysis to obtain their results, while the latter used content analysis to code messages and then performed a descriptive analysis.

## 2.4.2 Stage 2: During-trip

### Stage 2.1

#### Destination Choices

can be studied through a range of sources. Google Trends is commonly used to improve the forecasting (Bangwayo-Skeete and Skeete, 2015; Dergiades et al., 2018; Gunter et al., 2019; Park et al., 2017; Önder, 2017) or nowcasting (Jackman and Naitram, 2015) of tourist arrivals to specific destinations or groups of destinations, as well as to check the correlation between the GT series and the number of resident overnights at specific locations (Dinis et al., 2017).

The methodology employed in these studies typically involves selecting relevant search terms and using the index for those terms as a regressor to predict tourist flows. The forecasts or nowcasts are usually generated using time series analysis, primarily with auto-regressive models (Bangwayo-Skeete and Skeete, 2015; Dergiades et al., 2018; Gunter et al., 2019; Park et al., 2017; Jackman and Naitram, 2015; Önder, 2017), although descriptive analysis is also used in some cases (Dinis et al., 2017).

Mobile Roaming Data is another source employed to explore destination choices. In these studies, tourism flows to a destination are analyzed by tracking call activity from foreign mobile phones (Ahas et al., 2007, 2008; Raun et al., 2016; Tiru et al., 2010). Anonymous data is recovered from operators who collect it through their network of base stations, which track the activity of mobile phones in the area. Descriptive analysis is then used to estimate tourist flows to different destinations.

Twitter also proves useful in this context. For instance, Provenzano et al. (2018) employs Twitter data to explore mobility patterns in the EU, while other studies such as Kim et al. (2021); Carvache-Franco et al. (2022) examine the popularity of coastal destinations or the visitation rates to national parks (Tenkanen et al., 2017).

Regarding the methodology, the geolocation of the tweets or their content is used to aggregate the tweets in different areas. Afterwards, descriptive (Carvache-Franco et al., 2022; Provenzano et al., 2018; Tenkanen et al., 2017) or regression analyses (Kim et al., 2021) are conducted.

Flickr is also utilized to forecast tourist demand (Miah et al., 2017) and predict visits to National Parks (Tenkanen et al., 2017; Wood et al., 2013). In all these studies, the first step involves analyzing the geo-tagged data of the pictures uploaded to Flickr and then aggregating them by location. Various methods of analysis are subsequently employed, ranging from descriptive analysis (Wood et al., 2013; Tenkanen et al., 2017) to time-series analysis (Miah et al., 2017).

The studies by Gunter et al. (2019) and Önder et al. (2020) aim to predict monthly tourist arrivals to four Austrian cities by using the 'Likes' on posts of the Facebook pages of the DMOs as regressors. To do so, they use auto-regressive models to check the usefulness of these sources.

In contrast, Batista E Silva et al. (2018) study spatiotemporal patterns of tourists in the European Union through Online Review Data. Specifically, they analyze online reviews from Booking and TripAdvisor to establish the average daily number of overnight tourists, or 'tourist density', of different destinations in Europe through descriptive analysis.



Finally, Palazzo et al. (2021) study destination choices by collecting content from the profile of Instagram 'influencers', by analyzing the amount of posts by 'influencers' in a certain destination through descriptive analysis.

### Activity Choices

are mainly studied through Consumer Card and Instagram data. In studies using Consumer Card data, tourists are given cards upon arrival at the city (Scuderi and Dalle Nogare, 2018) or event (Zeni et al., 2009). A handful of locations are then strategically chosen to record the entrance of tourists. As a result, their choice of activities can be analyzed. In these studies, descriptive analyses are employed to monitor the frequency of each activity.

For Instagram, Rossi et al. (2018) use geo-tagged pictures from the Venice area, manually assigning them to one of six categories, which represent the different types of tourist activities in the city. The distribution of activities is then explored through descriptive analysis.

Meanwhile, Abbasi et al. (2015) use the geo-tagged information from tweets in the Sidney area, using the Latent Dirichlet Allocation (LDA) to cluster selected words from the tweets and establish six types of activities: shopping, entertainment, eating, work, social and study. Tweets similar enough to any of the clusters are then assigned to the corresponding activity type.

### Attraction Choices

are often studied using Flickr and Instagram data. In both cases, geo-tagged data from pictures is collected from a specific geographic area.

For Flickr, the popularity of the attraction is often established based on the frequency of pictures with geo-tagged data matching the attraction (Girardin et al., 2007; Donaire and Galí, 2011; Miah et al., 2017), or by studying the average time spent at the attraction, based on the time stamps of pictures taken (Popescu and Grefenstette, 2009). Descriptive analysis is commonly used (Girardin et al., 2007; Donaire and Galí, 2011), but some studies also use algorithms to create clusters of pictures and explore the preferred attractions (Miah et al., 2017).

Similarly, for Instagram, geo-tagged data is collected from pictures in the geographical area of interest. Afterward, the study of attraction choices can be done in different ways. For example, Paül i Agustí (2018) study how tourism is promoted in Montevideo by checking for overlap between user-generated content and promoted attractions in travel guides and brochures. Rossi et al. (2018) group the pictures into six categories and analyze the most popular attractions for each category. Paül i Agustí (2021) explore differences in behavior between genders by examining the frequency and the choice of attractions. Again, descriptive analysis is typically employed.

## Stage 2.2

### Mode/Route to destination

Mobile Roaming Data can be applied in different ways to study the mode and route to a destination. For instance, Ahas et al. (2008) track the first call activity of Russian and Latvian tourists in Estonia to estimate the point of entry. They use this data to explain the tendencies of these tourists when visiting Estonia. Similarly, Qian et al. (2021) study the entry and exit of tourists in Shanghai by tracking their coordinates to determine their transportation hub of before or after visiting certain tourist attractions.

Other authors retrieve geo-tagged data from Flickr to study travel routes. Girardin et al. (2007) focus on the tourism in the province of Florence. They retrieve pictures taken before and after the trip and, if these pictures were taken within 48 hours of a picture taken in Florence, they are considered part of the travel route. This allows the authors to establish entry and exit routes to the province.

In a different study, Ma et al. (2020) retrieve posts from Instagram related to the 2017 Great American Solar Eclipse and use them to study the movement patterns of tourists from their origin location to the observation point of choice. The home location of tourists is determined by examining their Instagram profiles.

Similarly, Bordogna et al. (2016) retrieve geo-tagged tweets from travelers to Lombardy, Italy, to study the most popular airports among foreigners, based on the amount of tweets sent from each airport.

In all these studies, descriptive analysis are performed through various methods such as correlations (Qian et al., 2021), cluster analysis (Ma et al., 2020), spatiotemporal analysis (Girardin et al., 2007; Ahas et al., 2008), or frequency of tweets (Bordogna et al., 2016).

## Stage 2.3

### Self-Gifts and other durable and non-durable purchases and Dining out choices

are primarily studied through Consumer Card Data. The first step is to select a group of establishments through which the spending of tourists will be tracked. Then, when the tourists use their Consumer Cards, the entry and/or spending at the location is recorded.

For example, Newing et al. (2014) use data from a loyalty card scheme to explore grocery expenditure among tourists in Cornwall (South West England). In another study, Zeni et al. (2009) distributed Consumer Cards to participants at festivals in Trento and tracked their spending choices by monitoring the locations in which the tourists used their cards. Descriptive analyses were used to examine spending habits.

### Mode/routes in and around destination choices

are frequently studied through the data from Specific Gathering Devices. Two main methodologies are used. The first involves asking visitors to carry GPS devices voluntarily, which track their location every few seconds or minutes. At the end of the day or the stay, the visitors are asked to return the GPS device. Using the data from the GPS devices, trajectories for each tourist are constructed,

and mobility patterns are established through descriptive analysis and mapping of the individual trajectories. Several studies have used this method, such as (Becco et al., 2013; Mc Kercher et al., 2012; Shoval and Isaacson, 2007; Sørensen and Sundbo, 2014; Zheng et al., 2017, 2019).

The second methodology involves placing Bluetooth access points around strategic locations of the city (Delafontaine et al., 2012; Versichele et al., 2012, 2014) or the attraction under study (Yoshimura et al., 2014), capturing Bluetooth data through them. These access points constantly scan for devices entering their area, and when they detect a device, they record its identifier (Media Access Control address, MAC) as well as the time stamp. This data allows the researchers to construct the trajectory of the individuals using the devices by using mapping techniques and/or descriptive analysis.

Tourist patterns can also be studied through Twitter data. Its geo-tagged data together with the timestamp of the tweet can be used to detect tourist routes. This is done by selecting a study area and defining boundaries for the geo-tagged data. The next step is to identify the tweets from tourists by checking their profiles (Chua et al., 2016) or using third-party software (Bordogna et al., 2016). Once tourists are identified, their tweets are ordered by time-stamp to construct the route. These studies employ techniques such as mapping (Chua et al., 2016) and clustering (Bordogna et al., 2016) in combination with descriptive analysis.

Xue et al. (2014) used machine learning techniques to identify tourists from public transport data in Singapore. The Consumer Card data from public transport in Singapore allowed the researchers to track the tourists' mode of transport, origin and destination of the ride among others. Mapping techniques were used to analyze their mobility patterns.

Girardin et al. (2007) studied flow patterns of tourists within the Province of Florence using geo-tagged pictures from Flickr to construct travel routes and map the travel patterns of different types of tourists.

Finally, Danila and Gaceu (2009) explored the use of rental car services by tourists in Romania by studying the webpage of a car rental company through Google Analytics, obtaining the number of monthly webpage visits and comparing them to the number of clients to check how webpage activity translates into clients.

### 2.4.3 Stage 3: Post-trip

#### **Satisfaction/dissatisfaction with specific events and experiences overall**

Twitter is the main source for studying opinions and sentiments of tourists. Most studies adopt a similar methodology: They first select a tourist destination or attraction (Stepchenkova et al., 2013; Padilla et al., 2018; Bhatt and Pickering, 2021), an economic sector such as hospitality (Philander and Zhong, 2016) or gastronomy (Park et al., 2016b) or a specific event such as COVID-19 (Sontayasara et al., 2021; Lu and Zheng, 2021; Mishra et al., 2021), the Australian bushfires of 2019-2020 (Willson et al., 2021) or a sporting event (Kirilenko and Stepchenkova, 2017) to collect tweets. Subsequently,

sentiment analysis is performed on the selected tweets to estimate the overall public sentiment towards the object of study.

However, while most studies focus on the opinions themselves (Willson et al., 2021; Sontayasara et al., 2021; Philander and Zhong, 2016; Park et al., 2016b; Mishra et al., 2021; Gulati, 2022; Bhatt and Pickering, 2021), others explore the attributes that may influence the tourists' opinions, such as the affective states that tourists associate with the location (Stepchenkova et al., 2013) or spatioemporal and temporal effects (Padilla et al., 2018; Kirilenko and Stepchenkova, 2017). All of these studies present their results through descriptive analyses.

Online Review Data is another relevant source. The structure of studies employing this data source are similar to those using Twitter Data. First, the authors obtain online travel reviews for the selected destination (Marine-Roig and Clavé, 2015; Kim et al., 2017) or for the accommodations at the destination (Cheng and Jin, 2019; Liu et al., 2017; Zhu et al., 2020). They then evaluate the reviews to obtain an understanding of the public's opinion.

In Online Review Data, there is a distinction between studies that focus strictly on tourists' opinions (Marine-Roig and Clavé, 2015; Kim et al., 2017), those that focus on the attributes of the accommodations that influence these opinions (Cheng and Jin, 2019; Liu et al., 2017) and those that investigate the relationship between the guests' feelings and the ratings they express in their reviews (Zhu et al., 2020).

While most of these studies use descriptive statistics (Marine-Roig and Clavé, 2015; Cheng and Jin, 2019; Liu et al., 2017; Kim et al., 2017), some employ regression analysis (Zhu et al., 2020).

## 2.5 Discussion

Table 2.1 summarizes the main results of the analysis, showcasing the diverse range of sources employed to study each stage and variable.

Regarding the sources used in the studies, it is observed that certain sources are specifically tailored to examine particular variables such as 'Specific Gathering Devices Data' to 'Mode/Route in Destinations', or 'Search Engines Data' to 'Destination Choices'. However, most sources are apparently useful in studying variables across different stages. In this vein, Instagram stands out as a versatile source, contributing to the analysis of nine different variables in the PCS model. Additionally, Flickr, Online Review Data, and Consumer Card Data exhibit applicability across all three stages of the model.

Regarding the variables of the PCS model, there are also differences in terms of the number of sources used to study them. For example, variables such as 'Marketing Influences' or 'Post-Trip Feelings' are studied through only two data sources. On the other hand, variables such as 'Demographics', 'Destination Choices' or 'Mode/Route in Destinations' have been studied using five or more distinct sources.

Table 2.1 : Matrix representing the association between the sources of Digital Footprint and the stage in the PCS model where they have been utilized, as observed in the analyzed literature. The symbol 'X' indicates that some papers were found for a given combination.

Stage	Variables	FB	TW	IN	FL	SED	WTS	ORD	MRD	CCD	SGD
1	Demographics			X	X		X	X			X
	Marketing Influences	X		X							
	Search Information	X		X			X				
2.1	Destination Choices	X	X	X	X	X		X	X		
	Activity Choices			X						X	
	Attraction Choices			X	X	X					
2.2	Mode/Routes to Destination		X	X	X				X		
2.3	Self-gifts & Dining out choices									X	
	Mode/Route in Destinations		X		X		X			X	X
3	Post-Trip Feelings		X					X			

Legend: FB = Facebook; TW = Twitter; IN = Instagram; FL = Flickr; SED = Search Engine Data; WTS = Website Traffic Statistics; ORD = Online Review Data; MRD = Mobile Roaming Data; CCD = Consumer Card Data; SGD = Specific Gathering Devices Data.

When analyzing the reviewed papers over time, it becomes evident that there has been a significant increase in the number of studies focusing on the Digital Footprint of tourists in recent years. Particularly, there is a clear trend in the literature towards utilizing Internet sources specifically. This shift is effectively illustrated in Figure 2.2, where the growing prominence of internet-related sources is evident when compared to non-internet sources.

Figure 2.3 focuses on the use of Social Media-based sources. Over time, there has been a decline in the popularity of Flickr in favor of the likes of Facebook, Instagram and Twitter. Facebook has proven valuable for studying marketing influences, while Tweets have been effective in capturing post-trip sentiments, among other uses. Instagram has gained an increased interest in the research literature, demonstrating also broader applicability. The predominance of Social Media sources can be partly attributed to their accessibility for data extraction through APIs or web scrapping. Moreover, their extensive user base makes them an attractive source for research purposes.

In addition to Social Media, Google Trends has emerged as a widely used tool for predicting demand for specific tourist destinations, thanks to the ease of accessing its data. Furthermore, online review data is vastly employed to gather the post-trip feelings and opinions from tourists.

However, it is important to acknowledge that Social Media and other internet sources do not represent the entire society, as certain segments, such as non-users or individuals who do not actively engage, are underrepresented (Gayo-Avello, 2013). To address these limitations, the integration of non-internet sources, such as Specific Gathering Devices and Mobile Roaming Data, can provide valuable additional information and help fill the gaps in the data.

In contrast, non-internet sources could provide less biased samples. For instance, the use of Bluetooth sensors in Specific Gathering Devices allows researchers to obtain information from a broader range of individuals, including those who are not active on Social Media. Similarly, Mobile Roaming Data only requires the tourist to carry their own phone and can be useful in gathering information from a broader range of individuals. However, the richness of the information collected through these sources may be significantly more limited than other sources with numerous of metrics for each tourist. As a main drawback of non-internet sources, it should be noted the sensitively higher cost associated with obtaining data, compared to Social Media usage or Google Trends series.

Regarding the distribution of the studied stages in the tourist decision-making process, Figure 2.4 highlights that Digital Footprint sources are predominantly used for examining Stage 2, which involves decisions made by tourists while at the destination. The figure also indicates a growing interest in the post-trip stage in recent years, largely driven by the availability of Online Review Data and advancements in natural language processing techniques.

Finally, our analysis reveals untapped potential in certain Digital Footprint sources. Search Engine Data, primarily used for studying destination choices, has the potential to describe activity and accommodation choices, as well as dining out choices. Through strategic keyword selection, valuable insights could be obtained regarding the preferred type of accommodation among tourists

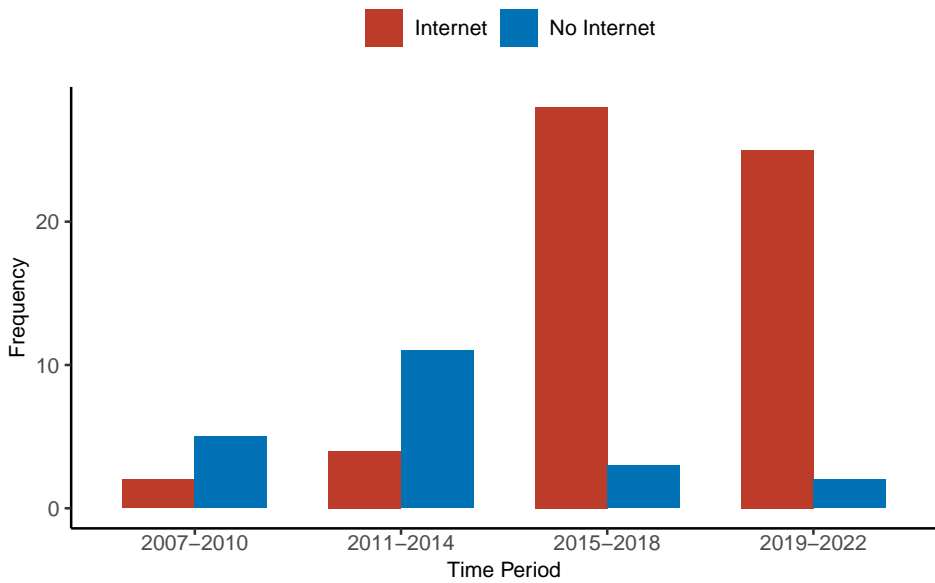


Figure 2.2: Evolution of the use of Digital Footprint sources for analyzing tourist behavior, comparing the use of internet to non-internet sources.

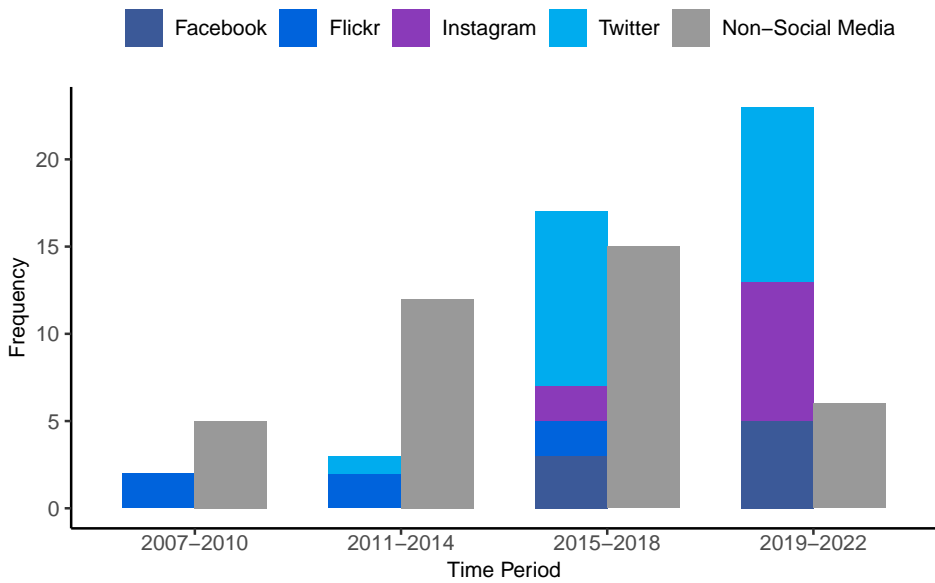


Figure 2.3: Evolution of the use of Digital Footprint sources for analyzing tourist behavior, comparing Social Media sources to non-Social Media sources.



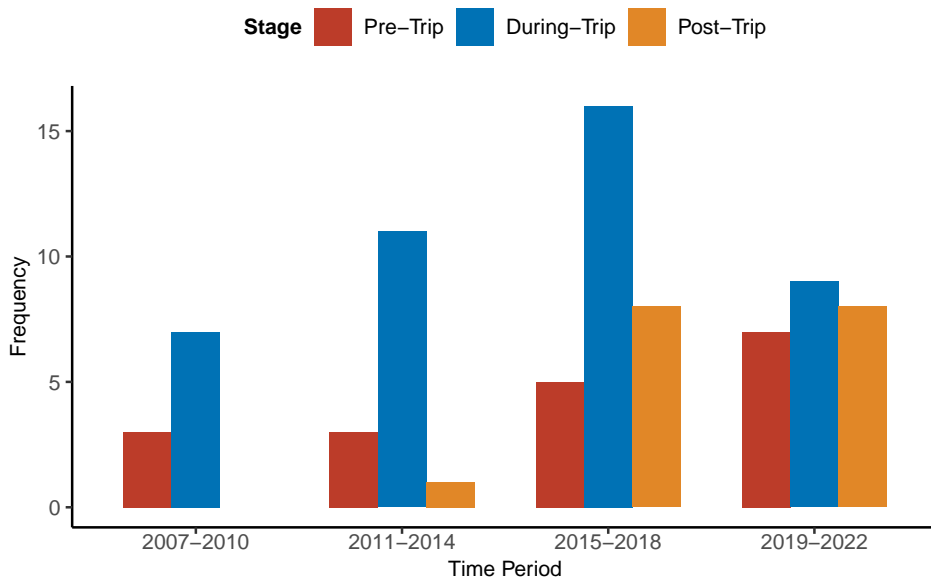


Figure 2.4: Evolution of the use of Digital Footprint sources for analyzing tourist behavior by stages in the Purchase-Consumption System model.

visiting specific destinations, and the popularity of various activities and restaurants within the city. Additionally, Twitter data offers further opportunities for investigating activity and attraction choices, thanks to the geo-position capabilities of user posts.

## 2.6 Conclusions

The rise of Big Data has made it possible to store the interaction between humans and technologies, resulting in what is known as the Digital Footprint. The Digital Footprint left by potential and actual tourists provides valuable insights and predictive capabilities for understanding their behavior. This paper contributes to the understanding of the various sources of Big Data and their utility in describing different aspects of the tourist decision-making process.

By reviewing and classifying these sources within the framework of the PCS model for leisure and travel, we have identified the current trends in the field. There is a notable focus on the decisions made by tourists while at the destination (Stage 2) and a preference for utilizing Internet sources, particularly Social Media sources.

However, our analysis also highlights unexplored potential in sources such as Search Engine Data and Twitter data. These sources offer opportunities to study activity and accommodation choices, as well as dining out choices and attraction preferences, respectively.

Overall, this research underscores the significance of the Digital Footprint in understanding and predicting tourist behavior. As the field continues to evolve, it is important to explore and harness the full potential of diverse data sources to gain deeper insights into the decision-making processes of tourists.

Our work has some limitations that should be acknowledged. First, the literature review was not systematic, which could have provided a more comprehensive understanding of the field. Instead, we relied on targeted keyword searches to identify relevant articles. This approach might have excluded some relevant research that was published in less common journals or not captured by our searches. Therefore, our selection of papers was subjective to some extent. Second, the classification of papers into variables of the PCS model also has some subjectivity. Since most reviewed papers did not explicitly refer to the PCS model, they were classified according to our own criteria.

Despite these limitations, the classification of data sources presented in this study offers valuable insights into the current trends and preferences in utilizing Digital Footprint sources for studying tourist behavior. It serves as a guide for researchers in selecting appropriate data sources based on their research interests and provides a foundation for future investigations in this rapidly evolving field.

## **Acknowledgements**

This work was partially supported by grants PID2019-107765RB-I00 and PEJ2018-003267-A-AR, funded by MCIN/AEI/10.13039/501100011033 and by “ESF Investing in your future”.





## 3. Is Google Trends a Quality Data Source?

*Dedicado a Victoria Gutiérrez Díez*

### Is Google Trends a Quality data source?

---

Chapter 3 is an adapted version of this published research paper:

- Year: 2023
  - Title: Is Google Trends a quality data source?
  - Authors: Eduardo Cebrián & Josep Domenech
  - Journal: Applied Economics Letters
  - Volume: 30
  - Pages: 811 - 815
  - DOI: 10.1080/13504851.2021.2023088
- 

#### Abstract

Google Trends (GT) has become a popular data source among researchers in a wide variety of fields. In economics, its main use has been to forecast other economic variables such as tourism demand, unemployment or sales. This paper questions the quality of these data by discussing the main data quality aspects according to the literature. Our analysis evidences some non-negligible issues related to the measurement accuracy of GT, which potentially affects the results obtained with GT data and therefore the decisions made with this information. These issues are illustrated with an example in which some queries to GT are repeated on six different days.

**Keywords**— Google Trends, Data Quality, Measurement Error, Online Data

### 3.1 Introduction

The rise in popularity of digital media has brought an enormous growth in the number of data sources related to the Digital Footprint left by businesses and consumers (Blazquez and Domenech, 2018). Such online data include sources such as social networking sites, corporate websites, and search engines, which have been used in a wide variety of research topics ranging from medicine (Pelat et al., 2009) or politics (Mellon, 2014) to finance (Preis et al., 2013).

Despite its increasing use in the literature, the quality of these non-traditional data sources has been largely overlooked. Data quality is a multi-dimensional concept which refers to the capability of data to be used quickly and effectively to inform and evaluate decisions. Issues with data quality, such as high measurement error, may impact on model parameter estimates and create economic inefficiencies (Bound et al., 2001).

Google Trends (GT) is a tool that provides reports on the popularity of certain searches in the Google search engine. Among the non-traditional data sources, GT is one of the most widely used in the empirical economic literature. It has demonstrated to be a good proxy for investor's attention (Da et al., 2011), even during the COVID-19 outbreak (Shear et al., 2021; Costola et al., 2021). It is also widely applied in other applied economics topics, ranging from unemployment to tourism demand (Choi and Varian, 2012; Jun et al., 2018). However, its quality as a data source has not been assessed.

This paper addresses this gap by discussing the data quality aspects of GT following the framework proposed by Karr et al. (2006). Our analysis detects that GT data have some non-negligible quality issues, which are evidenced in an illustrative example.

### 3.2 Google Trends

Google Trends is a freely available tool developed by Google that provides reports with the popularity of searches in Google Search. Reports, which include time-series data, are available for any user-selected time period, from 2004 to the present day and can also be restricted to focus on searches done in a certain language or from a specific location.

The searches whose popularity is reported by GT may be specified as terms, entities or categories. Terms refer to the text or keywords included in the search box.

An entity is an abstraction to refer to a single semantic unit, such as a place, a person, an object, an event, or a concept. Since entities refer to the semantics, they are independent from which terms are used to refer to them (i.e., synonyms), or even the language used. Using entities also avoids the problem of polysemic terms because GT identifies them by their ID in Freebase, which is a collaborative knowledge base.

Google classifies all searches into categories, such as Finance or Sports. These can be used to filter out unrelated searches in GT reports for terms or entities. If no term or entity is selected, the report includes all the searches that fall in that category. This way, it is possible to study the popularity

of all searches regarding one specific category.

The main GT output is the Search Volume Index (SVI), which is a time series representing the evolution of the popularity of a given search. This relative index is scaled to represent the highest popularity with an SVI value of 100. Notice that this normalization depends on the particular query to GT, so it depends on the specific search, period, language and geographical area that was selected. This means that it is not possible, for instance, to compare SVIs from different regions because values are relative to the total number of searches in each region.

### 3.3 Quality of Data Sources

The quality of data is, according to Karr et al. (2006), a wide multidimensional concept affecting different perspectives of the data source. Quality dimensions can be grouped into three different “hyperdimensions” of the data source: i) The process, which is related to the methods to generate, assemble, describe and maintain the data; ii) the data, which refers to the data itself contained in the data source; and iii) the use, which is related to how the source is used. The evaluation presented in this paper focuses on the Data hyperdimension.

The analysis of the data quality can also be applied to different levels of the data source: i) the database, ii) the tables composing the database, and iii) the records composing the tables in the database. Unlike traditional data sources, GT is not a database with a set of tables, but a set of records returned as a response to a given user request. For this reason, it is not possible to apply data quality concepts to the database or table levels of GT.

Following Karr et al. (2006), the main quality dimensions of data at the record level are: accuracy, completeness, consistency and validity. These dimensions refer not only to the values of each attribute in the record, but also to the intra-record relationships. Below, we describe these quality dimensions and apply them to GT data.

#### **Accuracy.**

It is related to whether or not the attribute value reports the true value. That is, this dimension is concerned with values measuring what they are expected to measure. Some statistical errors associated with the data, such as coverage biases, sampling defects or non responses, may characterize how accurate a source is.

GT presents an issue in terms of accuracy, derived from the fact that the reports are generated from a sample of searches made by users (Choi and Varian, 2012). The sampling methods are not disclosed by Google, so it is not possible to quantify the sampling error. Although Google recognizes that results may vary just a few per cent day to day due to this, the variation could be significant, as Section 3.4 evidences.

The popularity of searches reported by GT is often considered as an indirect method for measuring the attention to a given event or topic. Although the actual value of this interest is generally not known,

researchers should bear in mind the coverage bias inherent to GT. First, because it only represents the population with frequent access to the Internet (Steinmetz et al., 2014). Although it has increased over the years, it is still far from full coverage, especially in certain countries and group ages. Second, because GT can only collect what was searched for in Google Search. Google is the reference engine for general purpose searches. However, the increasing popularity of specialized sites or apps (such as Skyscanner or Booking) may affect the accuracy of GT for measuring interest in some topics.

### **Completeness.**

A record is complete when it includes values for all attributes. That is, records have no missing values.

GT includes data for all the observations, although it does not mean that a value is provided for each time period. Particularly, the value “0” is reported when the search did not reach a minimum threshold of popularity. The frequency of these missing values depends on the popularity of the search in the specific region of interest. Since “0” values precisely represent low popularity, the lack of completeness does not generally represent an important issue with GT data.

### **Consistency.**

It refers to the situation in which the relationships among the attribute values in the same record are valid. A lack of consistency is, for instance, a starting date after the end date.

GT reports the evolution of the search popularity in a two-attribute table: date and SVI. Since any relationship between values of both attributes is acceptable, no consistency issues may arise.

### **Validity.**

An attribute value is valid when it is within a pre-established domain of acceptable values. For example, a person’s age can only be a positive number. Ensuring attribute value validity is not enough for ensuring accuracy, although it is a necessary condition.

Data in GT reports are generally valid. Dates have well-formed values and SVI is usually between 0 and 100, as expected. However, there exist certain situations in which the SVI returns a non-integer value, particularly “< 1”. This means that the search in that time period had enough volume to appear in the report, but less than 1/100<sup>th</sup> than the period with the highest popularity. GT uses this notation to avoid confusion with the “0” value (which means missing data).

As in the case of completeness, this can be treated and does not represent an important issue. However, it highlights the lack of resolution of the SVI, as it only reports integer numbers.

## **3.4 Empirical Evidence**

This section illustrates some of the accuracy issues detected above with a simple experiment. It consists of repeating the same query to GT on six different days and comparing the results.



Table 3.1: Google Trends parameters in the experimental setting. Four searches, one per each search term, were explored.

Parameter	Values
Search terms	Graz, Salzburg, Innsbruck, Vienna
Time period	2010/06/01 – 2017/02/28
Category	Travel
Language	English
User location	Worldwide

### Searches.

This experiment was designed to reproduce the same searches as in Gunter et al. (2019), which aimed to forecast tourist arrivals to four Austrian cities. They consisted of four searches of the name of the main Austrian cities defined as search terms. Moreover, for all these searches, a constant time frame is used. The characteristics of the searches are shown in Table 3.1.

### Repetitions.

The four queries were submitted to GT on February 4, 2021, and repeated after one day, and weekly up to four weeks. This way, the results for each city were collected 6 times, resulting in 24 time series.

### Results.

Figure 3.1 represents the time series returned by GT on different dates. For the sake of clarity, only three of the six collection dates are shown here. As one can observe, the same queries do not always provide the same set of results. Notice that all these are queries with the exact same configuration, so one would expect that the same set of results is returned at all times. Although the oscillations in the time series are similar, the differences are far from being negligible. This is especially noticeable in the case of the “Graz” search term, where the blue line diverges quite often from the other two lines.

To quantify this dissimilarity, both the Pearson and the Spearman correlation coefficients,  $r$  and  $r_s$  respectively, between the GT results on February 5, 2021 and all the repetitions are computed. Table 3.2 shows that  $r$  ranges from 0.79 to 0.94, while Table 3.3 shows that  $r_s$  ranges from 0.74 to 0.92. In both cases, a decreasing trend is found in some of the series, Although the time series can be considered as highly correlated, they are far from the perfect correlation one would expect from a digital source. Therefore, this evidences that the data reported by GT is not completely accurate and includes some non-negligible measurement error.

To quantify how the measurement error could affect forecasts, some autoregressive distributed

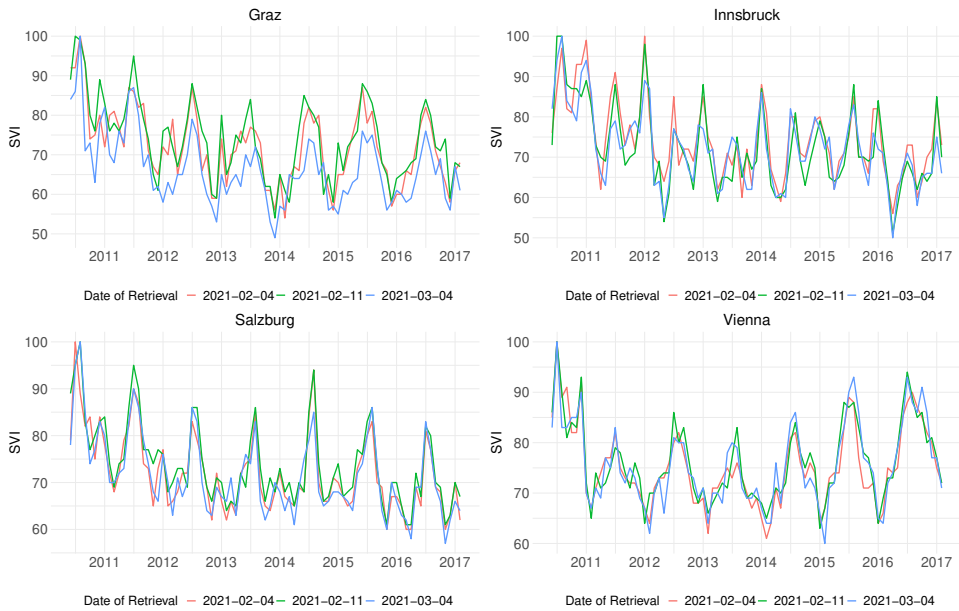


Figure 3.1: GT reports of searches for four Austrian cities collected at three different dates.

lag (ARDL) models were trained following the specification described by Gunter et al. (2019). The models for all four cities were estimated six times, one per each retrieval of Google Trends data. After checking that the in-sample and out-of-sample errors are similar to those reported by Gunter et al. (2019), a monthly out-of-sample forecast was generated for the last year of data with an advance ( $h$ ) of 1, 3 and 6 months. The range of these forecasts is shown in Table 3.4.

Considering only the one-month ahead forecast, the difference between the highest and the lowest estimation of arrivals ranges from 2196 tourists in Innsbruck to 5949 tourist arrivals in Salzburg. The cases of Graz and Salzburg are particularly relevant because the forecast differences can reach up to above 5% of the monthly average of tourist arrivals.

The variability of forecasts derives from the lack of accuracy of GT data, as also observed in Table 3.2. The source of the inaccuracy is probably related to the internal process used by Google to compute the SVI, including here the fact that Google does not use the whole set of searches to compute it, but only a small sample with unknown characteristics.

Table 3.2: Pearson Correlation coefficient of the GT data on February 4, 2021 with GT data returned on different dates.

Retrieval date	Graz	Innsbruck	Salzburg	Vienna
February 5, 2021	0.9084	0.9057	0.9404	0.9184
February 11, 2021	0.9075	0.8920	0.9378	0.9219
February 18, 2021	0.8201	0.8612	0.9030	0.9247
February 25, 2021	0.7936	0.8541	0.9317	0.9081
March 4, 2021	0.8304	0.8655	0.9152	0.9190

Table 3.3: Spearman Correlation coefficient of the GT data on February 4, 2021 with GT data returned on different dates.

Retrieval date	Graz	Innsbruck	Salzburg	Vienna
February 5, 2021	0.9151	0.8702	0.9243	0.8910
February 11, 2021	0.8921	0.8809	0.9175	0.9102
February 18, 2021	0.8320	0.8370	0.8930	0.8895
February 25, 2021	0.7412	0.8405	0.9055	0.8915
March 4, 2021	0.8184	0.8588	0.8812	0.9043

### 3.5 Conclusions

Google Trends has become a very popular data source among researchers of a wide variety of fields over the last decade. After analyzing the main quality dimensions of GT, some data quality issues arose. Those related to the accuracy of the data were considered as particularly relevant, as the lack of accuracy could become a significant source of bias, if it is not corrected. And, when data are used to estimate econometric models, it may affect parameter estimates that eventually would lead to making wrong economic or political decisions.

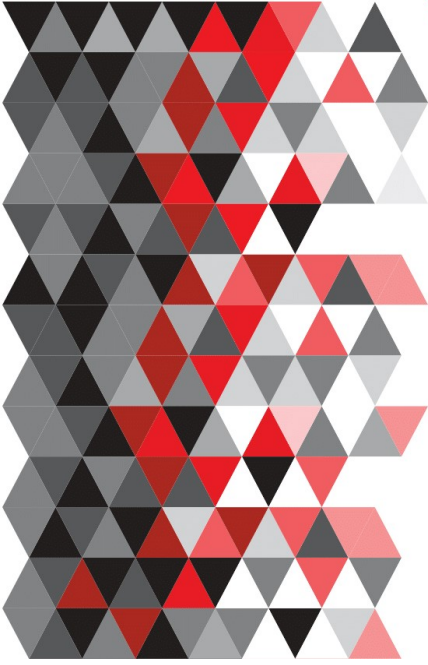
Our results highlight that the lack of accuracy of GT data is not negligible. Although these do not invalidate GT as a data source for social and economic analyses, little is known regarding the scope and the determinants of the inaccuracies. Future research works should explore and measure these issues in a wide variety of contexts to allow researchers to take remedial actions.

Table 3.4: Difference in forecast arrivals by retrieving GT data on different days and for different forecasting horizons ( $h$ ).

City	Difference in forecast arrivals			Average arrivals (Mar 2016 - Feb 2017)
	h=1	h=3	h=6	
Graz	2,681	2,719	2,555	52,586
Innsbruck	2,196	2,175	2,228	77,944
Salzburg	5,949	6,268	6,300	137,708
Vienna	3,894	4,161	5,862	577,195

## Acknowledgements

This work was partially supported by grants PID2019-107765RB-I00 and PEJ2018-003267-A-AR, funded by MCIN/AEI/10.13039/501100011033 and by “ESF Investing in your future”.



## 4. Addressing Google Trends inconsistencies

*Dedicado al Dr. Alfonso Díez Minguela*

### Addressing Google Trends inconsistencies

---

Chapter 4 is an adapted version of this published research paper:

- Year: 2024
  - Addressing Google Trends inconsistencies
  - Authors: Eduardo Cebrián & Josep Domenech
  - Journal: Technological Forecasting and Social Change
  - Volume: 202
  - Pages: 123318
  - DOI: 10.1016/j.techfore.2024.123318
- 

#### Abstract

Google Trends reports the evolution of the popularity of Internet searches. Its main output is the Search Volume Index (SVI), a relative measure of the popularity of a term computed using a sample of the searches. Due to the sampling, the SVI series are not entirely consistent, as the same query produces different results that can widely change from day to day. This paper investigates the nature of these inconsistencies by modeling and simulating the data-generating process. Simulations are applied to describe how a typical time series is distorted due to the sampling process and to quantify how averaging extractions smoothes the series. Finally, a relationship between term popularity, series dispersion, and the averaged extractions is derived, so recommendations for constructing consistent SVIs can be provided.

**Keywords**— Google Trends, Consistency, Popularity, Online Data

## 4.1 Introduction

Google Trends (GT) is a freely available tool developed by Google that allows users to obtain reports of the evolution of the popularity of searches made through the Google Search engine. In the last decade, GT has become popular in the scientific literature because its reports can be used to measure the population's interest on any topic. Moreover, this data can be easily accessed and is continuously updated. Its use is widely spread across a variety of research fields such as medicine (Pelat et al., 2009; Lippi et al., 2022; Díaz et al., 2023), politics (Mellon, 2014; Jelnov and Jelnov, 2022; Casteli Gattinara et al., 2022; Mavragani and Tsagarakis, 2016), economics (Choi and Varian, 2009; Vicente et al., 2015; Bantis et al., 2023; Castelnuovo and Tran, 2017) or tourism (Rivera, 2016; Havranek and Zeynalov, 2021; Yang et al., 2022; Dergiades et al., 2018).

The main output of GT are time series data representing the Search Volume Index (SVI), a relative measure of the popularity of a term. To compute the popularity, Google does not consider the whole set of searches received in a given time period, but a sample with unknown characteristics. Due to the sampling error, the reports are not completely consistent, as the same query can produce different time series which change from day to day (Choi and Varian, 2012; Preis et al., 2013). The importance of these inconsistencies is often minimized (Choi and Varian, 2012; Dilmaghani, 2019) although Cebrián and Domenech (2023b) reported that variations in GT data may be significant enough to hinder the interpretability and reproducibility of the models estimated with them.

To alleviate the sampling error, a usual solution is to extract data from GT a certain number of times and consider the average SVI instead. In this vein, Carrière-Swallow and Labbé (2013) take the average of 50 extractions. Other authors, however, use 10 (Saxa, 2015), 14 (Barreira et al., 2013), or 20 (Borup et al., 2022). Despite the variety of procedures, there is no discussion about the optimal (or, at least, desirable) number of extractions and how effective the averaging method is.

This paper investigates the nature of the inconsistencies in GT data. To understand them, a model of the data-generating process is proposed and simulated. This model is applied to describe how a typical time series is distorted due to the sampling process and how averaging extractions smoothes the series. After that, a relationship between term popularity, series dispersion, and the number of extractions is derived so recommendations for constructing consistent SVIs can be provided.

The contribution of this paper is threefold: i) it proposes a model to understand how the inconsistencies of GT data are created; ii) the model is simulated to describe the error associated with the GT sampling for a stylized seasonal search term; and iii) it quantifies the relationship between the popularity of search terms, the number of extractions, and the expected error derived from GT sampling.

The remainder of the paper is structured as follows. Section 4.2 reviews some literature about the methodology issues in GT. Section 4.3 introduces the details of the sampling process in GT. Section 4.4 describes the simulation model and some simulation results. Section 4.5 quantifies the relationship between the inconsistencies of GT data and the number of extractions. Finally, Section 4.6

presents some concluding remarks.

## 4.2 Related Work

The increasing adoption of GT in research for tracking public interest and behavior has surfaced various methodological issues. This section aims to review the main research works that specifically focus on addressing the methodological challenges associated with the use of Google Trends.

### Selection of search terms

The selection of terms is usually one of the initial steps in working with GT. It must be carefully planned because it is a factor that could introduce bias in the study's results (Nuti et al., 2014). A distinction needs to be made between the volume of search queries and the actual event being studied. GT may reflect irrational user behaviors, such as the rapid dissemination of emotional reactions or negative news, rather than substantive, rational activities (Jun et al., 2018). For instance, in the case of Google Flu Trends, their predictions failed because the elevated search volumes for flu-related terms were influenced more by media coverage than by an actual increase in flu cases (Butler, 2013).

Furthermore, identifying the correct semantic interpretation is often problematic. Queries based on keywords are language-specific and can be ambiguous, as polysemic words can distort measurements (Arora et al., 2019), thus favoring the use of categories and topics in GT reports (Nicolas Woloszko, 2020). Additionally, Mavragani et al. (2018) mention challenges specific to non-English languages, particularly those with more complex alphabets.

Further complicating matters is the often poor documentation of how GT is utilized in academic research, leading to a lack of reproducibility in many studies (Nuti et al., 2014). This deficiency calls for greater transparency to enhance the reliability of GT-based research. However, there is yet no consensus on documenting queries and search strategies, adding an extra layer of complexity to this methodological challenge (Arora et al., 2019).

### Changing patterns of total searches

SVI reports the volume of specific search terms relative to total searches. The main interest is often the numerator of this ratio. However, the denominator, representing total search volume, is subject to substantial fluctuations, which are commonly overlooked.

Nicolas Woloszko (2020) indicates that variations in total search volume can introduce biases, as internet usage evolves. It should be emphasized that GT data is restricted to the population with internet access who utilize Google as their search engine. This population has significantly expanded over the years, changing its composition as well (Narita and Yin, 2018; Bokelmann and Lessmann, 2019).

To deal with the changing trend of total searches, Bokelmann and Lessmann (2019) suggest to compare the term of interest with another term, aiming to isolate the specific search interest from

the general search trend. Nicolas Woloszko (2020), instead, employs Principal Component Analysis (PCA) and filter out long-term trends using an HP filter.

Indeed, the broader influences on total search volume can arise from various fronts. The attention to the composition of total searches is particularly relevant when analyzing long time series and studying developing countries. However, abrupt shifts in search behavior can also happen in relation with other observed events like the COVID-19 pandemic, thus altering the time-series data (Knipe et al., 2021).

### **Changes in Google's algorithms**

Algorithmic changes in GT and Google search engine are another critical consideration when using GT for academic research. Such alterations can introduce abrupt breaks in GT series that require adjustments for valid interpretation (Bantis et al., 2023; Bokelmann and Lessmann, 2019; Nicolas Woloszko, 2020). Notable instances of changes in GT algorithms occurred in January 2011, 2016 and 2022, which can impact the estimations derived from GT data

Not only the algorithms in GT may affect the results, but also the algorithms in Google search engine. In the context of Google Flu Trends, Lazer et al. (2014) noted that the search engine provides users with suggested search terms, influencing the search behavior of users. As a result, changes in the recommendation system not only affect the search data reported by GT but also underscores the dynamic and endogenously cultivated nature of search behavior by the service provider, Google.

### **Comparability across terms**

Comparing SVIs for different terms also represents a methodological challenge in GT research. One important obstacle is that each series is normalized to its peak popularity, which results in non-comparable SVIs. Comparison is only feasible when the terms are queried simultaneously in GT. However, this approach carries its limitations, as terms with lower popularity are often reduced to zero since GT reports values in integers.

To address these limitations, Malagón-Selma et al. (2023) suggest using a chain of terms wherein more popular terms serve as a reference for less popular ones. Similarly, Springer et al. (2023) recommend using terms with maximum interest, such as “coronavirus”, as a benchmark for assessing the relative popularity of other terms.

When comparing terms, Narita and Yin (2018) emphasizes that GT measures searches, not people or activities. Search terms linked to high-frequency activities like real-time financial investment can inflate their SVIs, giving the appearance of higher public interest than is actually the case. This is particularly noticeable when compared to SVIs for search terms related to less immediate activities like car purchases or travel planning. Such nuances caution against making direct comparisons of SVIs across different types of search queries without understanding the context and frequency of these searches.



### **Inconsistencies across time frequencies**

Google Trends provides data in multiple time frequencies, including hourly, daily, weekly, and monthly formats. However, the data consistency across these time frames is not guaranteed, as the chained higher resolution observations do not have the same trend as lower resolution data points (Nicolas Woloszko, 2020; Eichenauer et al., 2022). These inconsistencies can undermine the validity of time-sensitive analyses, such as real-time event tracking.

Particularly, Nicolas Woloszko (2020) points out that weekly series do not align consistently with monthly series, necessitating calibration for comparative analyses. Similarly, Eichenauer et al. (2022) highlight that the raw data varies when comparing different time frequencies, such as daily versus monthly data. To reconcile these inconsistencies, the authors propose a methodology to integrate data across different frequencies while preserving the long-term trend.

### **Inconsistencies derived from sampling**

These inconsistencies are derived from the fact that GT reports data based on a changing sample of queries. As a result, identical queries might not always return the same set of results (Narita and Yin, 2018; Medeiros and Pires, 2021; Rovetta, 2021; Cebrián and Domenech, 2023b). Although this issue has been recognized for some time (Choi and Varian, 2012; Da et al., 2011; Carrière-Swallow and Labbé, 2013; Combes and Bortoli, 2016), many research works using GT do not address these inconsistencies. For instance, Bangwayo-Skeete and Skeete (2015) use GT data to predict tourist demand to five touristic destinations in the Caribbean, but do not explicitly mention any treatment of the GT data to account for the inconsistencies. Similarly, Hu et al. (2018) use GT data to improve the prediction of the direction of a stock index. In the medical field, Walker et al. (2020) study a potential relationship between the number of searches for loss of smell and the number of COVID cases and fatalities. None of these works provide information about a treatment for the inconsistencies from GT, probably because they use a single extraction.

Some other authors identify these inconsistencies, but they do not consider them relevant enough to affect their results. Choi and Varian (2012) nowcast certain economic indicators such as unemployment claims or automobile sales in the US. They state that GT data is computed through a sampling method and the results change daily, but disregard it as a problem. Da et al. (2011) propose a measure of investor attention based on GT and extract SVIs for a sample of stocks. They describe that the impact of the sampling error is small as they obtain correlations of above 97% among the different SVIs of the same stock. Stephens-Davidowitz and Varian (2015) state that GT sampling usually gives reasonably precise estimates and suggest that researchers will generally not need more than a single extraction.

Finally, there are some other research works that identify these inconsistencies as an important source of error and consider multiple GT requests of the same time series. Among them, their conclusions on the magnitude of the problem widely vary. D'Amuri and Marcucci (2017) took

24 different extractions for the search term “jobs” and reported cross-correlations of at least 0.99 between extractions. Cebrián and Domenech (2023b) extracted queries related to Austrian cities on 6 different occasions and found correlations between 0.79 and 0.94. Saxa (2015) took 10 different extractions for mortgage-related terms 10 different and obtained correlations between 0.78 and 0.85. Carrière-Swallow and Labbé (2013) used the average standard deviation to measure the sampling error and reported values above 15% for the term “Chevrolet” after 50 extractions. Barreira et al. (2013) studied how GT data can improve forecasts of unemployment and car sales in Spain, Portugal, France, and Italy. To do so, they averaged extractions on 14 different days, reporting average standard deviations ranging from 3.5% to 7.6% for the different search terms used.

Similarly, to alleviate these inconsistencies, Eichenauer et al. (2022) average at least 12 samples, Simran and Sharma (2023) use 7, Nicolas Woloszko (2020) takes 6 samples and Tudor and Sova (2023) average 5 extractions. Both Nicolas Woloszko (2020) and Eichenauer et al. (2022) exclude series with large variance among extractions, although they do not offer explicit criteria for that.

Taking several samples from GT data has an additional inconvenience due to its cache system. If data is requested within 24 hours of the original request, the system will return the same sample (Raubenheimer, 2022; Stephens-Davidowitz and Varian, 2015). This can prolong the data collection process and reduce the freshness of data. Previous research has proposed methods to circumvent this issue, such as making requests with varying overlapping time periods (Raubenheimer, 2022), or adding disjunctions with random words in the search term (Askitas, 2015).

To the best of our knowledge, there is no method for determining how many extractions should be averaged to alleviate the sampling error. To understand the intricacies of how the SVIs are produced and the effects of averaging multiple extractions of the same GT time series, this paper uses a simulation model to generate the SVIs (and its sampling error) in a similar way to Google Trends. The simulation model is used afterward to provide recommendations on how many extractions to consider, depending on the acceptable error.

### 4.3 Google Trends sampling

GT does not compute the SVI from the whole set of searches that Google received but from a sample (Da et al., 2011; Böhme et al., 2020; D’Amuri and Marcucci, 2017). This sample is used to create the SVI time series, in a process similar to the one illustrated in Figure 4.1 (Narita and Yin, 2018). The upper part of the figure represents the general process followed by the data. The lower part gives a simple example of the GT sampling process and how it affects the computation of the SVI for a query.

The SVI computation process departs from the whole set of searches that Google has received since 2004 (*Total Searches* in Figure 4.1). From this set, GT draws a random sample. This introduces a sampling error that is unknown because both the sample and the population size are not disclosed

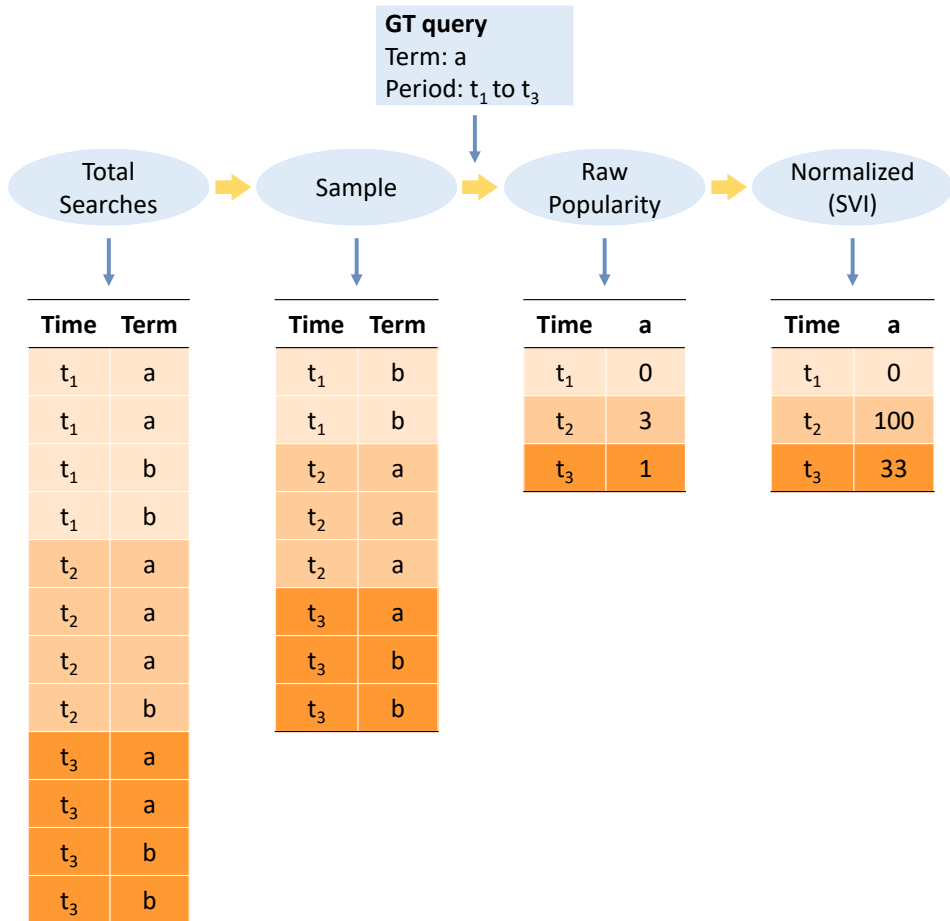


Figure 4.1: Process GT follows to compute an SVI time series. The upper part represents the general process followed by the data. The lower part gives a simple example of how the sampling affects the computation of the SVI for a term. The example assumes that the GT report for the term  $a$  from  $t_1$  to  $t_3$  is requested.

by Google. When a user requests the GT report for a given query<sup>1</sup> and time period, the sample is filtered to keep only those searches matching the request. This filtered sample is used to compute the frequencies by time period<sup>2</sup>, which are Search Volume time series measured in quantity of searches (represented as *Raw popularity* in Figure 4.1). Finally, the time series are normalized by setting the SVI to 100 in the most popular period and proportionally scaling the values in other periods. These values are finally rounded to integers (*Normalized (SVI)* in Figure 4.1).

The sampling error is illustrated in the lower part of Figure 4.1. In the example provided, the GT report for term *a* is requested. If the SVI were computed considering the *Total Searches* set, the series (67, 100, 67) would have been returned. Instead, (0, 100, 33) values are reported due to the sampling error.

The sample used for computing the SVI is not static. GT changes the sample used from day to day. Moreover, one should note that each reported SVI series cannot be considered an independent sample due to the normalization process.

## 4.4 Simulating GT sampling

This section provides some simulations of the GT process described in Section 4.3 to check how the SVI time series change depending on the popularity of the search query and what the effect of averaging multiple extractions is.

### 4.4.1 Modeling

To simulate the SVI generation process as in Figure 4.1, the first step is to model an actual *Total Searches* set from which the sample will be drawn. Since tourism forecasting is one of the popular applications of GT, we assume that the popularity of the search query of interest follows a time series with trend and seasonal components. Equation 5.1 presents the general function modeling the total searches of the query of interest *y*. It includes a sine wave function with a cycle of 12 time periods, representing the seasonality component in monthly data. The sine wave function is chosen because of two main reasons: (i) it has proven effective in modeling cyclic behavior such as those in tourism (Song et al., 2019; Chu, 2004; Chan, 1993; Wong, 1997), and (ii) it is smooth, so distortions to its

---

<sup>1</sup>Queries can be defined using terms/keywords, topics or categories. For the sake of clarity, the description of the methodology and results mainly refer to terms. However, the principles and findings discussed are applicable and extend to topics and categories as well.

<sup>2</sup>Data point frequencies can vary from minutes to months depending on the selected time period. Users cannot select the frequency directly; instead, Google Trends determines it based on the chosen length of the time period. For instance, a search spanning the past 24 hours will yield data points every 8 minutes, whereas a search covering the previous 6 years provides monthly data points.

shape can be easily visualized.

$$Y_t = \beta_0 + \beta_1 t + \beta_2 \sin\left(\frac{2\pi}{12}t\right) \quad (4.1)$$

In Equation 5.1,  $Y_t$  is the number of searches of the query  $y$  at time  $t$ ,  $\beta_1$  is the parameter defining the strength of the linear trend, and  $\beta_2$  defines the strength of the seasonal component.

For each time period  $t$ , the frequency of the query  $y$  in the sample ( $y_t$ ) follows a binomial distribution with parameter  $n$ , which equals the sample size, and  $p$ , which equals the proportion of searches of query  $y$  among all the searches received by Google at that time period. Therefore, the expected number of occurrences in the sample of query  $y$  at time  $t$  is:

$$E[y_t] = n \cdot p_t \quad (4.2)$$

Notice that  $p_t$  varies in time, being this variation the change in popularity of the query.

#### 4.4.2 Scenarios

Different simulation parameters are considered to illustrate the inconsistencies in a variety of situations. A total of six scenarios are simulated, by combining two patterns of time series with three popularity levels.

##### Search patterns.

The patterns of the time series are defined by the parameters of Equation 5.1. Table 4.1 shows the sets of parameters for the two considered patterns. In the first one, the seasonal component dominates the trend, while in the second, the trend component is relatively stronger. Parameter sets have been adjusted to produce values consistent with SVIs (i.e., between 0 and 100) in a 60-period simulation where the time frame remains the same through the whole simulation process. The resulting *Total searches* are represented in Figure 4.2.

##### Query popularity.

The different levels of popularity are defined by changing parameters in Equation 4.2. A term with ‘High popularity’ has an average expected frequency of 2000. A term with ‘Medium popularity’ has an average expected frequency of 20, while the ‘Low popularity’ term has an average expected frequency of 2.

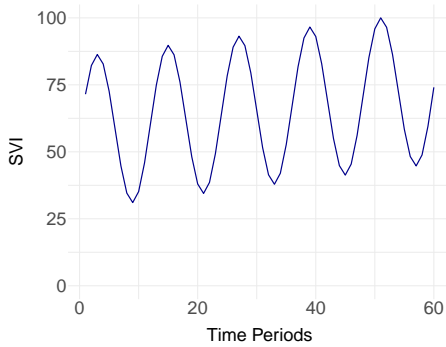
#### 4.4.3 Simulation results

Simulations are conducted for 60 periods, equivalent to five years of monthly data. To illustrate how averaging GT extractions work, the random process of generating the SVI for each term has been repeated up to 20 times, each one representing one extraction.

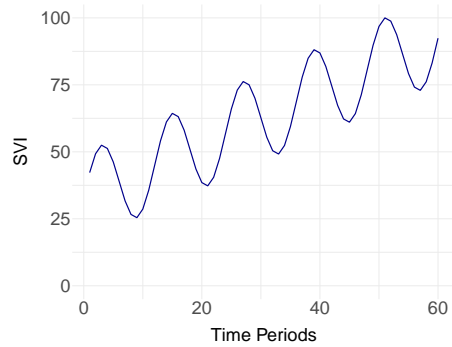
Figure 4.3 and Figure 4.4 represent the simulation for the seasonal-dominant and trend patterns, respectively. Each figure shows the results of 1 (left), 10 (center), and 20 (right) SVI extractions.

Table 4.1: Sets of parameters used in the simulation.

Parameter	Seasonal-dominant pattern	Trend-dominant pattern
$\beta_0$	56.99	33
$\beta_1$	0.28	1
$\beta_2$	28.49	16.47



(a) Seasonal-dominant pattern



(b) Trend-dominant pattern

Figure 4.2: Actual SVI for a Seasonal-dominant pattern (a) and a Trend-dominant pattern (b)

Light blue lines represent each individual extraction, while dark blue lines illustrate the average of all the extractions in each plot. Plots in the top row refer to a term with ‘High popularity’, plots in the middle row refer to a ‘Medium popularity’ term, and plots in the lower row refer to a ‘Low popularity’ term. Each plot legend includes the Pearson’s correlation coefficient ( $r$ ) as well as the Spearman’s correlation coefficient ( $r_s$ ) of the time series in dark blue with the actual popularity of the term (i.e.,  $Y_t$  in Equation 5.1).

The sampling error contained in one extraction is considerable, as can be observed in Figure 4.3 and Figure 4.4 (especially in plots d and g). This can also be seen as the lower values of the correlation coefficients  $r$  and  $r_s$  between the real distribution and one extraction, which is accentuated when the term is less popular (graphically, when moving from top to bottom plots). One single extraction is unable to capture the increasing trend of the original series in the terms with ‘Low popularity’ (subfigures g), and even in the term with ‘Medium popularity’ and seasonal-dominant pattern (Figure 4.3d), despite a Pearson correlation coefficient  $r$  as high as 0.80. This is not an issue, however, in ‘Highly popular’ terms (subfigs. a).

When averaging more extractions (i.e., center and right columns in Figure 4.3 and Figure 4.4), the noise is reduced, as the patterns of the dark lines get more similar to the actual values (Figure 4.2).

This can also be numerically observed as an increase in the correlation coefficients  $r$  and  $r_s$ . However, the improvement is not uniform for the different terms, as ‘Low popularity’ terms are noticeably more noisy than the others.

Hence, the number of extractions to alleviate the sampling issues of GT depends on the popularity of the search term.

A side effect of averaging multiple extractions is that the range of the estimated SVI changes. Since every single extraction is normalized to 100, the average of extractions can only reach that value if all the extractions have their maximum in the same time period. This is more evident in the case of a ‘Low popularity’ term: The range of 1 extraction is [0, 100] (Fig. 3g), but the range of the average of 20 extractions is [12, 56] (Fig. 3i), while the range of the actual values is [30, 100] (Figure 4.2a). This implies that the average of an increasing number of extractions does not converge to the actual value of the SVI. Similar behavior can be observed in Figure 4.4 with the trend-dominant pattern.

The alteration of the SVI range is relevant because it affects the model estimations made with the GT data. That is, the parameters of any regression model estimated with averaged extractions will change depending on the number of extractions used. However, this is a minor issue because it could be easily fixed by re-scaling the maximum value to 100.

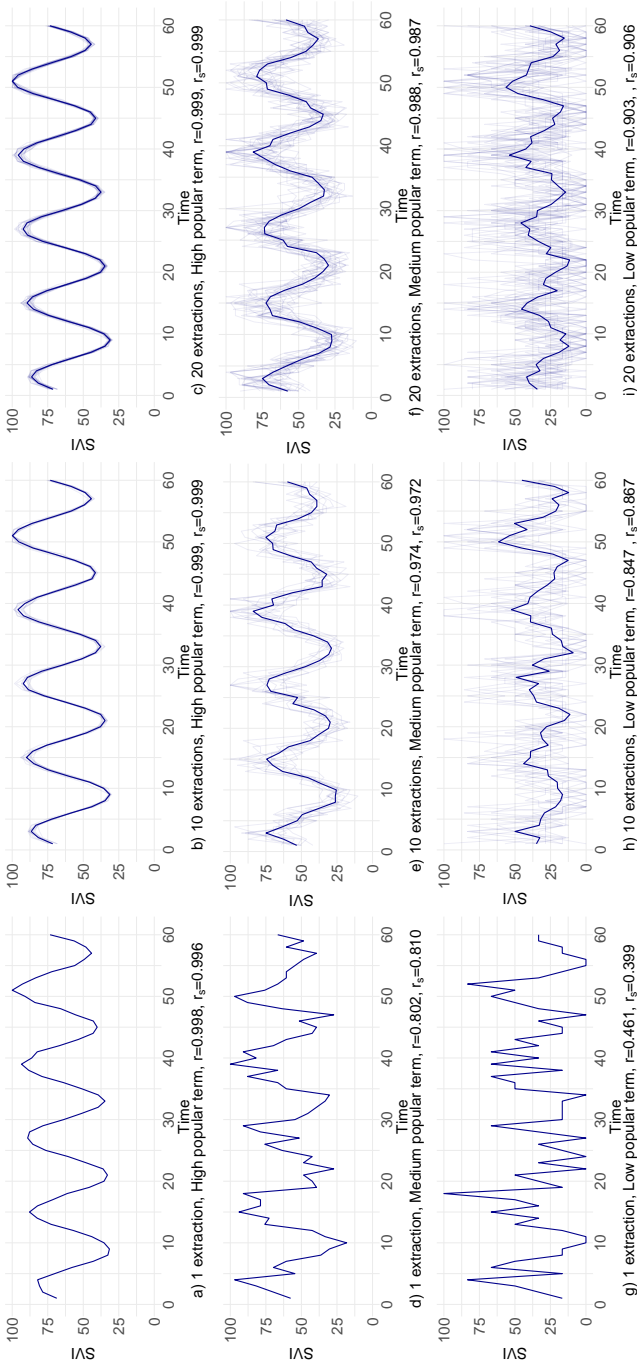


Figure 4.3: Simulation of GT data generation process for a term with seasonal-dominant pattern. Each individual extraction is shown in light blue. The darker line represents the average of all extractions.  $r$  and  $r_s$  are the Pearson and Spearman Correlation coefficients respectively between the dark line and the actual values.



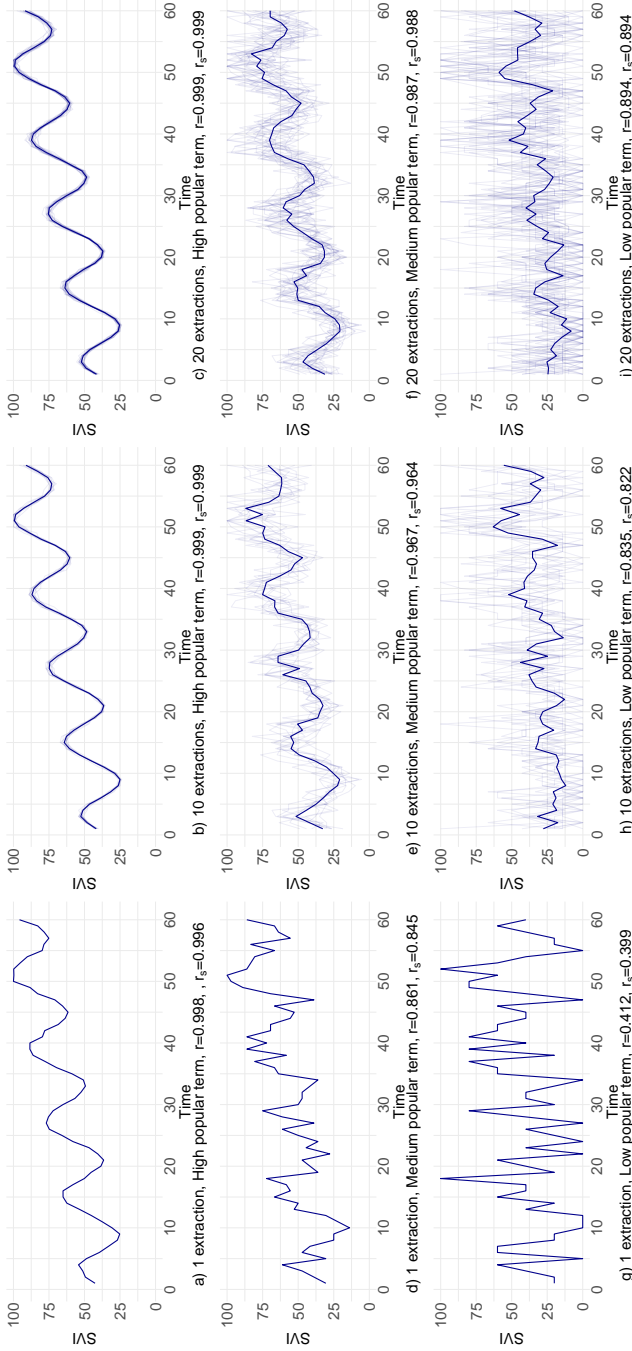


Figure 4.4: Simulation of GT data generation process for a term with trend-dominant pattern. Each individual extraction is shown in light blue. The darker line represents the average of all extractions.  $r$  and  $r_s$  are the Pearson and Spearman Correlation coefficients respectively between the dark line and the actual values.

## 4.5 Alleviating Google Trends inconsistencies

The results above have evidenced that the sampling error decreases with the search popularity of the term, although this can be alleviated by averaging multiple extractions. This section examines the relation between the inconsistencies and the number of extractions, modulated by the popularity of the search term. To do so, Subsection 4.5.1 describes a way to account for the popularity of a term out of a simulation environment. Subsection 4.5.2 introduces how the inconsistencies are measured, Subsection 4.5.3 proposes an equation to quantify the relationship, and Subsection 4.5.4 compares the simulated results with empirically obtained data.

### 4.5.1 A Measure for Popularity

The popularity of a term defined in Equation 4.2 cannot be observed because Google does not disclose it. However, the variability among different extractions, which is negatively correlated with the term popularity, can be quantified from a sample of extractions. Therefore, the mean of the standard deviation of the different extractions ( $\bar{s}$ ) is an indirect measure of term unpopularity. It is defined as:

$$\bar{s} = \frac{1}{T} \sum_{t=1}^T s_t \quad (4.3)$$

where  $T$  is the length of the time series, and  $s_t$  is the standard deviation in each period, defined as:

$$s_t = \sqrt{\frac{\sum_{i=1}^k (y_{i,t} - \bar{y}_t')^2}{k-1}} \quad (4.4)$$

where  $y_{i,t}$  is the SVI for period  $t$  provided by GT in the  $i$ -th extraction,  $k$  is the number of extractions, and  $\bar{y}_t'$  is the average of SVIs for time period  $t$ :

$$\bar{y}_t' = \frac{1}{k} \sum_{i=1}^k y_{i,t} \quad (4.5)$$

Notice that the maximum of  $\bar{y}_t'$  may not be 100, as evidenced in Subsection 4.4.3. To be comparable with the series retrieved in a single extraction, it must be re-scaled to the  $[0, 100]$  range as:

$$\bar{y}_t' = \bar{y}_t \frac{100}{\max(\bar{y})} \quad (4.6)$$

**4.5.2 A Measure for Inconsistency**

GT series are inconsistent because they deviate from the SVI that it would be produced from the whole population of searches. This way, the Mean Absolute Percentage Error (MAPE) of the average of extractions measures how inconsistent it is. The MAPE is defined as:

$$MAPE = \frac{1}{T} \sum_{t=1}^T \left| \frac{y_t^* - y_t'}{y_t^*} \right| \cdot 100 \tag{4.7}$$

where  $y_t^*$  is the SVI computed from the whole population and  $y_t'$  is the re-scaled SVI computed from the average of extractions, as defined in Equation 4.6. Notice that  $y_t^*$  cannot be observed because it is not disclosed by Google, but is known in the simulation environment.

**4.5.3 Reducing inconsistencies by averaging extractions**

The relation between the MAPE and the number of extractions averaged to compute the SVI was examined by simulating the six scenarios described in Subsection 4.4.2. To this end, we computed the MAPE after simulating and averaging between 1 and 3000 extractions. Figure 4.5 shows the simulation results in a log-log scale after 1000 repetitions of the whole process.

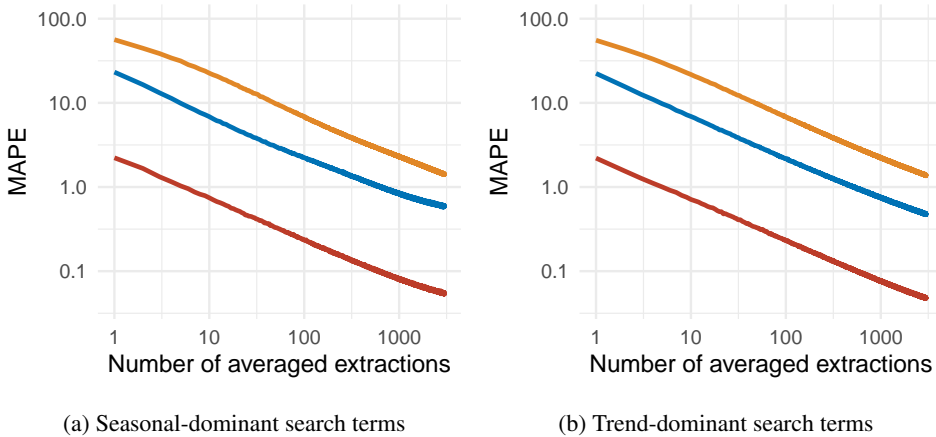


Figure 4.5: Relationship between MAPE and the number of averaged extractions (log-log scale) under six different scenarios: two search patterns (seasonal and trend-dominant) combined with three popularity levels (High, red lines; Medium, blue lines; Low, orange lines)

Figure 4.5a and Figure 4.5b represent the scenarios of the seasonal-dominant and trend-dominant patterns, respectively. Red, blue, and orange lines represent the high, medium, and, low popularity

terms, respectively. As one can observe, both plots are almost identical. This means that the pattern of the search term does not significantly affect the relationship between the term popularity and the inconsistencies of the GT series. Regarding the term popularity, plots show that averaging 10 extractions of the highly popular term has a MAPE slightly below 1%. To reach this error with the medium popularity term, it would be required to average around 1000 extractions. It can also be observed that the MAPE of a single extraction of the highly popular term is comparable to the average of 100 extractions of the medium popularity term.

The data from the results shown in Figure 4.5 were used to estimate an equation quantifying the relationship between the term popularity and the MAPE of the SVI computed after averaging a number of extractions. After testing different specifications (see Appendix A), the predicted MAPE can be approximated as:

$$\widehat{MAPE} = 1.3728 \cdot \frac{\bar{s}}{\sqrt{k}} + 0.0034 \cdot \frac{\bar{s}^3}{\sqrt{k}} \quad (4.8)$$

where  $k$  is the number of averaged extractions, and  $\bar{s}$  is the term unpopularity, measured as the mean of the standard deviation of the different extractions. Equation 4.8 can be algebraically rearranged to express the number of extractions as an explicit function of the term unpopularity and the acceptable error.

$$k = \frac{(1.3728 \cdot \bar{s} + 0.0034 \cdot \bar{s}^3)^2}{MAPE^2} \quad (4.9)$$

Equation 4.9 allows any researcher to plan how many extractions are required to keep the error associated with GT sampling within the acceptable limits, which may depend on the specific purpose.

Table 4.2: Summary of Google Trends terms extracted from the literature with their respective periods, geography, mean standard deviation ( $\bar{s}$ ).

Term	Period	Geography	$\bar{s}$	Reference
Insolvenz	1-1-2020 24-4-2020	Germany	1.70	Eichenauer et al. (2022)
Insolvenz	1-1-2020 24-4-2020	Austria	15.02	Eichenauer et al. (2022)
Graz	1-6-2010 28-2-2017	World	3.58	Gunter et al. (2019)
Puerto Rico Hotels	1-1-2004 31-1-2014	USA	10.98	Rivera (2016)

Table 4.3: Summary of Google Trends terms extracted from the literature with their mean standard deviation ( $\bar{s}$ ) and the necessary amount of extractions to obtain a 1% MAPE according to Equation 4.9.

Term	$\bar{s}$	Amount of extractions	Rounded Amount
Insolvenz	1.70	5.52	6
Insolvenz	15.02	1033.06	1034
Graz	3.58	25.711	26
Puerto Rico Hotels	10.98	383.146	384

#### 4.5.4 Empirical validation

To validate Equation 4.8 with real data, several GT queries from previous research studies were replicated through R-Studio and analyzed. These are summarized in Table 4.2, while the necessary amount of extractions necessary to obtain a 1% MAPE according to the formula in Equation 4.9 is provided in Table 4.3. For the validation, each GT series was extracted 100 times, and its average considered as the 'true' value ( $y_i^*$ ) for the series.

To put the GT queries into perspective, the term 'Insolvenz' in Germany has a mean standard deviation ( $\bar{s}$ ) of 1.70, similar to the high-popularity term in the simulations ( $\bar{s} = 1.759$ ), and therefore a 1% MAPE would be obtainable with just 6 extractions. Likewise, the term 'Puerto Rico Hotels' has a mean standard deviation similar to the simulated medium-popularity term ( $\bar{s} = 12.327$ ) and a MAPE of 1% would be obtainable with 384 extractions. The term with a popularity closer to the low-popularity term ( $\bar{s} = 22.634$ ) is 'Insolvenz' when searched in Austria, and for which a MAPE of 1% could be obtained after 1034 extractions. More examples of the  $\bar{s}$  of terms used in other studies compared to the ones in this paper can be found in Appendix B.

For each term, an increasing number of extractions were averaged and its deviation from the 'true' value measured with MAPE. Figure 4.6 compares the empirical MAPE with the theoretical value derived from Equation 4.8. The red lines show empirical MAPE for one instance of GT series averaging, while the shaded areas represent the 95% confidence interval estimated using bootstrapping

with 5,000 repetitions. As one can observe, the theoretical estimations not only fall within the shaded area, but also capture the general trend of the empirical errors. Figure 4.6 also shows that single GT data extraction, especially for less popular terms, can be unreliable. Even with multiple extraction averaging, there is a significant variation in the MAPE, indicating caution when using GT data.

## 4.6 Conclusions

Google Trends has become a popular data source among researchers. However, the data it provides could be inconsistent due to the sampling process used by Google to compute the SVIs. This paper has modeled and simulated this process to understand the nature of the inconsistencies and propose remedial actions.

Our simulation results showed that the inconsistencies are related to the popularity of the searches. While the absolute amount of searches cannot be retrieved from Google Trends, its popularity can be indirectly measured with the mean of the standard deviations of repeated extractions.

Averaging multiple GT extractions may alleviate the described inconsistencies, but only to a limited extent. Our main results are summarized in Equation 4.9. It allows any researcher to find the number of extractions required to keep the associated error within acceptable limits. Results also suggest that studies based on low popular queries are subject to a high error and hence reduced accuracy. This is consistent with the observations made by Nicolas Woloszko (2020) and Eichenauer et al. (2022), where queries with few searches result in highly variant time series. Therefore, GT data for these terms must be treated with caution.

## Acknowledgements

This work was partially supported by grants PID2019-107765RB-I00 and PEJ2018-003267-A-AR, funded by MCIN/AEI/10.13039/501100011033 and by “ESF Investing in your future”.

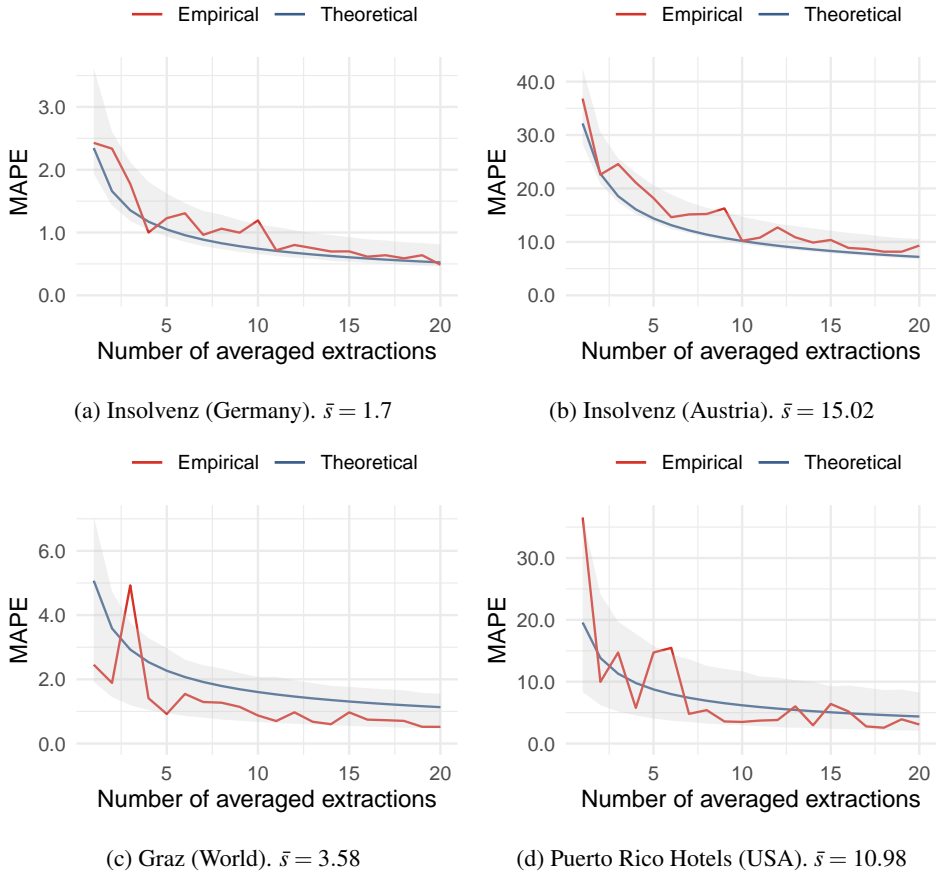


Figure 4.6: Comparison of empirical Mean Absolute Percentage Error (MAPE) with theoretical values from Equation 4.8 for selected Google Trends terms. Red lines indicate empirical MAPE for a single averaging of GT series, while shaded regions represent 95% confidence intervals determined through bootstrapping. Each subfigure corresponds to a specific term and search location, together with its associated mean standard deviation ( $\bar{s}$ ).

## Appendix A. Model estimation for the relationship between term popularity and GT sampling error

Table 4.4 presents the estimations of different specifications for modeling the relationship between the term popularity, the number of extractions and the error associated with GT sampling. Model 1 offers the best fit among the four different specifications. Although Model 4 also presents an adjusted  $R^2$  of 0.9974, its quadratic term is non-significant. Therefore Model 1 is chosen and presented in the main text as Equation 4.8.

Table 4.4: Models tested for the relationship between term popularity and MAPE of the SVI

Variable	Model 1	Model 2	Model 3	Model 4
$(\bar{s})/k$	-	0.1136***	-	-
$(\bar{s})/\sqrt{k}$	1.3728***	0.5205***	2.7043***	1.3728***
$(\bar{s})^2/\sqrt{k}$	-	-	-	0.0000
$(\bar{s})^3/\sqrt{k}$	0.0034***	-	-	0.0034***
Adjusted $R^2$	0.9974	0.9969	0.9541	0.9974
p-value	0.0000	0.0000	0.0000	0.0000

*Dependent Variable: MAPE. Estimation results obtained by ordinary least squares. \*\*\*Statistical Significance at the 1% level, \*\*Statistical Significance at the 5% level, \*Statistical Significance at the 10% level.*

## Appendix B

This Appendix presents the mean of the standard deviations of different extractions ( $\bar{s}$ ) for the simulated scenarios (see Table 4.5) and for a selection of GT queries used in the literature (see Table 4.6). These are calculated after repeating 30 times the extractions from GT. Since changes in the sampling method used by GT are not documented, it is not possible to assure that the deviations reported here are the same as the deviations when the respective authors accessed the data.




Table 4.5: Mean of the standard deviation ( $\bar{s}$ ) of the simulation scenarios.

<b>Popularity</b>	<b>Pattern</b>	<b>Frequency</b>	<b>Periods</b>	<b><math>\bar{s}</math></b>
High	Seasonal-dominant	Monthly	60	1.759
High	Trend-dominant	Monthly	60	1.636
Medium	Seasonal-dominant	Monthly	60	12.327
Medium	Trend-dominant	Monthly	60	12.226
Low	Seasonal-dominant	Monthly	60	22.634
Low	Trend-dominant	Monthly	60	22.619

Table 4.6: Mean of the standard deviation ( $\bar{s}$ ) of 30 different extractions for a selection of GT queries used in the literature.

Term	Frequency	Period	Area	$\bar{s}$
<b>Eichenauer et al. (2022)</b>				
Insolvenz	Daily	1-1-2020 24-4-2020	Germany	1.702
Insolvenz	Daily	1-1-2020 24-4-2020	Austria	15.024
Insolvenz	Daily	1-1-2020 24-4-2020	Switzerland	19.071
<b>Gunter et al. (2019)</b>				
Graz	Monthly	1-6-2010 28-2-2017	World	3.581
Innsbruck	Monthly	1-6-2010 28-2-2017	World	2.124
Salzburg	Monthly	1-6-2010 28-2-2017	World	1.92
Vienna	Monthly	1-6-2010 28-2-2017	World	1.626
<b>Preis et al. (2013)</b>				
Credit	Monthly	5-1-2004 22-2-2011	USA	1.818
War	Monthly	5-1-2004 22-2-2011	USA	1.401
Culture	Monthly	5-1-2004 22-2-2011	USA	2.761
Politics	Monthly	5-1-2004 22-2-2011	USA	0.771
Bubble	Monthly	5-1-2004 22-2-2011	USA	1.766
Consumption	Monthly	5-1-2004 22-2-2011	USA	1.002
<b>Rivera (2016)</b>				
Puerto Rico Hotels	Monthly	1-1-2004 31-1-2014	USA	10.984
Puerto Rico Flights	Monthly	1-1-2004 31-1-2014	USA	8.304
San Juan Hotels	Monthly	1-1-2004 31-1-2014	USA	11.759
Puerto Rico Resorts	Monthly	1-1-2004 31-1-2014	USA	7.430
Puerto Rico Vacations	Monthly	1-1-2004 31-1-2014	USA	7.428
Puerto Rico Vacation	Monthly	1-1-2004 31-1-2014	USA	9.061
Puerto Rico Tourism	Monthly	1-1-2004 31-1-2014	USA	11.256
Puerto Rico Travel	Monthly	1-1-2004 31-1-2014	USA	11.308
Puerto Rico Hotel Deals	Monthly	1-1-2004 31-1-2014	USA	11.008

Notes: All the searches in this work are restricted to the category 67: "Travel".



## 5. Can Google Trends predict rural tourism? The case of Spain

*Dedicado a la Dra. Guadalupe Serrano Domingo*

### Can Google Trends predict Rural Tourism? The case of Spain

#### Abstract

Rural tourism is a sector which is becoming increasingly more important not only in Spain but around the world, and yet the prediction of rural tourism flows is hardly being addressed despite its importance. In the tourism literature, Google Trends (GT) has emerged as one of the most common sources to predict touristic flows. However, GT has not been reported to predict rural tourism yet. This paper investigates whether if GT can help improve predictions on the monthly rural overnight stays by national residents in Spain (2012M01 – 2020M02). To do so, the forecasting accuracy of models which include a variable of GT information is compared with that of classic time-series benchmark models. Unlike what happens with other destinations, the results show that models which employ a GT variable are unable to outperform classical time-series models. Some discussion is provided to understand the different causes of this outcome.

**Keywords**— Rural Tourism; Google Trends; Tourism Demand; Forecasting

#### 5.1 Introduction

Over the last few decades the ascent of the Internet as well as that of the online booking platforms has greatly increased the accessibility and the appeal of rural tourism. In fact, by 2021 rural areas comprised 43.8% of accommodation beds and contributed to 37% of overnight stays in the European

## Chapter 5. Can Google Trends predict rural tourism? The case of Spain 76

Union (EPRS, 2023). In Spain, tourism is a particularly important sector in the economy, as it represented a 12.4% of GDP in 2019 (UNWTO, 2023a). Moreover, from 2012 to 2019 there had been a significant increase of above 20% (INE, 2023) of rural tourism overnight stays by national residents.

Rural tourism can be a great source for the development of rural areas as it can help create sustainable, local employment and can increase the levels of profits in these areas (Guaita Martínez et al., 2019; Wijijayanti et al., 2020).

For that reason, scientific literature has focused on exploring what are the drivers that can foster this development (Yang et al., 2021; Liu et al., 2020; Kumar et al., 2022; Rid et al., 2014) as well as understanding the issues that may hinder this growth process (Wang et al., 2013; Rosalina et al., 2021).

However, the precise prediction of tourism flows can aid management as well as policy makers in these rural areas in making decisions about the planning and strategies in tourism (Xi and Donglai, 2022). In the tourism literature, one way of predicting touristic flows is by including searches from Google Trends (GT) (Bangwayo-Skeete and Skeete, 2015; Carrière-Swallow and Labbé, 2013; Havranek and Zeynalov, 2021) and to the best of our knowledge, this approach has not been reported in rural tourism.

Therefore, this paper addresses this literature gap by comparing the forecasting performance of models that include different information from GT with some other time-series benchmark models to predict rural tourism overnight stays in Spain (2012M01 – 2020M02).

The rest of this paper is structured as follows: Section 5.2 provides a literature review of relevant articles which have used GT to improve predictions of variables from different topics. Section 5.3 introduces the extraction and treatment of the data, as well as the in-sample fit. Section 5.4 describes the forecasting results obtained and Section 5.5 presents some concluding remarks.

### 5.2 Literature Review

Tourism holds significant potential for revitalizing rural economies by providing diversified income sources and enhancing the socioeconomic fabric of rural communities (Guaita Martínez et al., 2019; Wijijayanti et al., 2020). The economic impact of rural tourism is multifaceted, including stimulation of economic growth, alleviation of poverty, and improvement of living standards in local communities (Wilson et al., 2001; Liu et al., 2023). Such growth often emerges from the clustering of tourism-related activities, fostering cooperation and partnerships between local actors (Wilson et al., 2001; Kumar et al., 2022). This collaborative framework not only boosts local economies but also ensures the sustainability of tourism initiatives, creating a robust economic ecosystem (Nooripoor et al., 2021).

Beyond economic benefits, tourism contributes significantly to the sociocultural and environmental well-being of rural areas. It offers avenues for cultural preservation, social stability, and

## **Chapter 5. Can Google Trends predict rural tourism? The case of Spain 77**

---

community pride while enhancing the quality of life for residents (Liu et al., 2023). The sociocultural benefits are diverse, ranging from the revitalization of local customs and crafts to the promotion of cultural heritage and community identity (Lane and Kastenholz, 2015). Environmentally, tourism can lead to improved conservation efforts, promoting biodiversity and sustainable land use practices. This shift towards tourism allows rural communities to maintain their natural and scenic beauty, which is vital for attracting tourists and ensuring long-term sustainability (Lane and Kastenholz, 2015; Rosalina et al., 2021; Jepson and Sharpley, 2015). Thus, the integration of tourism into rural economies provides a holistic approach to rural development, addressing economic, sociocultural, and environmental dimensions, and fostering a sustainable future for rural communities.

One of the key elements of tourism planning and development consists in predicting touristic flows. In this regard, the literature for forecasting of touristic flows is extensive, with search engine data being the most common data source used to improve forecast predictions, specially through time series and econometric techniques (Antolini and Grassini, 2019; Li et al., 2021b; Cebrián and Domenech, 2023a).

The most commonly used search data engine is GT. GT is a tool which provides an index on the relative popularity of searches made in Google's search engine from 2004 to the present day (Cebrián and Domenech, 2023b). The index can be restricted to a certain time and geographic location, and is based on a query share where the point in time with the highest amount of searches for the selected search term is set to a 100 and the rest of points are calculated from it (Choi and Varian, 2012). Moreover, searches can be placed into categories, such as 'Autos and Vehicles' (category 47) or 'Movies' (category 34), which act as filters so that when a category is selected alongside a search term, only the searches related to that category are shown.

In this vein, Bangwayo-Skeete and Skeete (2015) use GT searches of hotels and flights from US, Canada and the U.K. to five Caribbean destinations and find that GT improves forecasting for these touristic flows while Park et al. (2017) successfully improve forecasts of Japanese inflows to Korea by employing models including GT searches.

However, not all authors find that GT always improves forecasting in tourism. For example, Rivera (2016) uses GT data to attempt to improve forecasting predictions of the number of non-resident registrations in hotels in Puerto Rico, and finds that while predictions using linear models with GT perform better in longer forecast horizons, Holt-Winters models predict better in short-term horizons.

In terms of rural tourism, the forecasting of touristic flows is less developed in the literature. For example, Yin (2020) successfully adapts forecasting methods to newly developed areas in rural China, while Xi and Donglai (2022) use an enhanced Quad-Res Net model for forecasting regional flows of rural tourism to Jiayuguan, China and are able to improve forecasting results.

However, both authors point out potential difficulties presented by rural tourism flows, as they might potentially present seasonal effects as well as significant amounts of volatility (Xi and Donglai,

## Chapter 5. Can Google Trends predict rural tourism? The case of Spain 78

2022). Furthermore, there is a deficiency of adequate historical market data, and concerns arise regarding the limited online presence of newly developed rural areas. (Yin, 2020), which in turn, could limit the effect of traditional tourism forecasting methods.

To the best of our knowledge, the use of GT data as a predictor for rural touristic flows has not been reported in the literature so far, and so we identify a literature gap. Do these well-established methods in tourism work as well in rural tourism? Can we predict rural tourist inflows in the same way as general tourist inflows are predicted? Therefore, the contribution of this article is the forecasting of rural overnight stays by national residents in Spain by including Autoregressive Distributed Lag models (ADL) with GT and checking whether if they outperform benchmark time-series models.

### 5.3 Methodology

The Methodology Section is divided in to three parts: First, Subsection 5.3.1 describes the data used in the manuscript. Next, Subsection 5.3.2 deals with the necessary transformations performed to the data as well as the techniques utilized to obtain the different Model specifications. Finally, Subsection 5.3.3 describes the fit of the models specified in the previous part.

#### 5.3.1 Data

##### Rural Overnight Stays

The dependent variable employed is the monthly rural overnight stays by national residents in Spain from 2012M1 to 2020M2. This period is specifically selected to avoid the COVID-19 pandemic, which could introduce a shock in the series which could potentially make the comparison across models harder. The data is extracted from a survey conducted on the occupation of rural tourism accommodations (Encuesta de ocupación en alojamientos de turismo rural) by the INE (Instituto Nacional de Estadística), which is Spain's official statistical institute, which publish open data for Spain regarding any topic.

##### Google Trends data

In order to use data from GT, the first step is the selection of keywords based on which search terms might be relevant to the dependent variable. This is done by extracting a set of different proposed combination of search terms and categories and then computing the correlation of each combination and the dependent variable. The extraction is carried out with the 'trendecon' R-package (Eichenauer et al., 2022) and the data is obtained in a monthly frequency to match that of the dependent variable. Then, those combinations with the highest correlation to the dependent variable are chosen for next steps (Table 5.1). The proposed search terms are 'rural' (rural), 'turismo rural' (rural tourism) and 'agroturismo' (agrotourism). Then, these search terms are filtered through the proposed categories: 67 'travel'; 1389 'agrotourism'; 1005 'ecotourism' and 1391 'vineyards and wine tourism'. Finally, these searchers are all obtained for the geographic area of Spain. Finally, the chosen queries are the term

## Chapter 5. Can Google Trends predict rural tourism? The case of Spain 79

Table 5.1: Search term and category combinations and their correlation to the dependent variable.

Term	Category	Correlation
<b>rural</b>	<b>travel</b>	<b>0.50</b>
<b>rural</b>	<b>agroturismo</b>	<b>0.66</b>
rural	ecoturismo	0.37
rural	vineyards & wine tourism	0.04
turismo rural	travel	0.35
turismo rural	agroturismo	0.29
turismo rural	ecoturismo	0.17
turismo rural	vineyards & wine tourism	-0.01
<b>agroturismo</b>	<b>travel</b>	<b>0.71</b>
<b>agroturismo</b>	<b>agroturismo</b>	<b>0.69</b>
agroturismo	ecoturismo	0.05
agroturismo	vineyards & wine tourism	-0.01

*Notes: Rows in bold indicate the selected queries for the next steps in the study.*

'rural' under the 'travel' and 'agroturismo' categories and the term 'agroturismo' under the 'travel' and 'agroturismo' categories.

Given that the different extractions vary from day to day and can cause non-negligible issues with the accuracy of the data (Cebrián and Domenech, 2023b), we use the methodology of Cebrián and Domenech (2024) for the processing of GT data. First, the selected combinations are extracted 10 times each to compute the mean standard deviation (s.d.) of the different extractions. Based on this mean s.d., the formula is applied for a 1% MAPE (Mean Absolute Percentage Error) and then, it can be solved to obtain the necessary amount of extractions to arrive to a 1% MAPE for each one of the selected combinations. Following this method, the results on Table 5.2 are obtained.

### 5.3.2 Models

Once the search terms have been extracted, they are transformed into logarithms along with the rest of the data. Then, unit root tests are performed to all the variables and both the rural overnight stays as well as the different GT terms are found to have one unit root. For this reason, all the variables are used in first differences. Then, the variables are deseasonalized.

Next, several Autoregressive Distributed Lag Models (ARLM's) are fitted: First, a baseline model with lagged values of the dependent variable as the only predictor, and secondly, a series of models

Table 5.2: Necessary extractions to obtain a 1% MAPE for each search term.

Term	Category	S.D. with 10 ex- tractions	Number of ex- tractions	Rounded num- ber of extrac- tions
rural	travel	1.72	5.64	6
rural	ecoturism	2.50	12.20	13
agroturismo	travel	2.75	14.74	15
agroturismo	ecoturism	3.76	28.55	29

which add lagged values of the different GT terms to the baseline specification. Moreover, all the models contain both a linear and a quadratic trend component so that both linear and non-linear forms of trending behavior might be captured.

Finally, the models are estimated through Ordinary Least Squares (OLS) with heteroskedasticity robust errors using the full sample (2012M1-2020M2). A general specification of the models is represented in Equation 5.1, where  $Y_t$  represents the logarithm of overnight stays,  $Y_{t-i}$  represents the lags of the dependent variable and  $GT_{t-i}$  are the lags for the GT term and  $t$  and  $t^2$  are the linear and quadratic trend components respectively<sup>1</sup>.

$$Y_t = \alpha + \sum_{i=1}^{12} \gamma * Y_{t-i} + \sum_{i=0}^{12} \beta * GT_{t-i} + \delta * t + \zeta * t^2 + \epsilon_t \tag{5.1}$$

**5.3.3 Estimation results**

The results for the in-sample estimations are shown in Table 5.3. The results for the in-sample estimations are shown in Table 5.3. While all the models provide highly significant F-statistics, Model 3 presents the best fit with an adjusted R<sup>2</sup> of 0.563, while Model 5 presents the lowest with an adjusted R<sup>2</sup> of 0.488.

For all the models, at least the first four lags of the dependent variable are significant, but after that, there are differences across models. In, Models 2 and 3 most of the rest of lags of the dependent variable are significant, while only 5 lags are so in Model 4.

Regarding the GT terms, there are also some noteworthy differences. Firstly, Model 5 provides no significant GT term, lagged or not, which also results in the worst fitting Model. Models 2, 3 and 4, which are the three best fits by a decent margin provide in each case, at least three significant GT terms, and in Model 2, 10 of the 13 GT terms are significant at 10%.

Finally, no significant trending behavior is found in any of the models, linear or otherwise.

<sup>1</sup>The methodology in this study is an adaptation of the one employed by Önder (2017).



Table 5.3: In-sample fit model estimations

Model	Model 1	Model 2	Model 3	Model 4	Model 5
Query	-	rural	rural	agroturism	agroturism
Category	-	travel	agroturism	travel	agroturism
Constant	0.010	0.022	0.012	0.011	0.013
ln(y) (-1)	-0.808***	-0.827***	-0.746***	-0.766***	-0.809***
ln(y) (-2)	-0.659***	-0.737***	-0.641***	-0.656***	-0.676***
ln(y) (-3)	-0.599***	-0.778***	-0.587***	-0.653***	-0.677***
ln(y) (-4)	-0.518**	-0.736***	-0.579***	-0.477**	-0.544***
ln(y) (-5)	-0.409*	-0.727**	-0.440**	-0.307	-0.409**
ln(y) (-6)	-0.225	-0.506**	-0.239	-0.098	-0.210
ln(y) (-7)	-0.273	-0.594***	-0.305*	-0.039	-0.250
ln(y) (-8)	-0.291	-0.558***	-0.317*	-0.053	-0.240
ln(y) (-9)	-0.363**	-0.591***	-0.390**	-0.193	-0.340**
ln(y) (-10)	-0.352**	-0.528***	-0.312**	-0.110	-0.310*
ln(y) (-11)	-0.104	-0.216	0.077	-0.025	-0.133
ln(y) (-12)	-0.300**	-0.409***	-0.367***	-0.280**	-0.305***
ln (GT)	-	0.155	0.071**	0.058*	0.024
ln(GT) (-1)	-	0.228**	0.085*	0.082**	0.051
ln(GT) (-2)	-	0.234**	0.072	0.079	0.067
ln(GT) (-3)	-	0.294***	-0.012	0.098*	0.070
ln(GT) (-4)	-	0.196*	0.064	0.062	0.064
ln(GT) (-5)	-	0.330**	0.046	0.040	0.037
ln(GT) (-6)	-	0.204*	0.022	0.025	0.021
ln(GT) (-7)	-	0.258**	0.002	0.038	-0.008
ln(GT) (-8)	-	0.198	0.006	-0.022	-0.027
ln(GT) (-9)	-	0.302***	0.044	0.001	0.016
ln(GT) (-10)	-	0.211*	0.005	-0.023	-0.030
ln(GT) (-11)	-	0.076	-0.031	-0.021	-0.008
ln(GT) (-12)	-	0.191*	0.008	0.053*	0.021
Trend	0.000	0.000	0.000	0.000	0.000
Trend <sup>2</sup>	-0.000	-0.000	-0.000	-0.000	-0.000
Adjusted R <sup>2</sup>	0.501	0.553	0.548	0.563	0.488
F-Statistic	8.265	7.926	8.751	8.795	5.897
P-Value	4.308e <sup>-10</sup>	3.549e <sup>-11</sup>	4.741e <sup>-12</sup>	4.278e <sup>-12</sup>	9.654e <sup>-9</sup>

Notes: Dependent Variable: ln(y). Estimation results obtained by ordinary least squares. \*\*\*Statistical Significance at the 1% level, \*\*Statistical Significance at the 5% level, \*Statistical Significance at the 10% level.

## 5.4 Forecasting Results

The out-of-sample forecasting performance is compared by means of the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Moreover, following Önder (2017), other time series models such as a naïve model and a non-seasonal Holt-Winters model are added as competing models.

Then, all forecasting horizons are calculated using a rolling window methodology and prediction errors are obtained for forecasting horizons 1 through 12. The results of this process are shown in Table 5.4 for forecasting horizons of 1, 2, 3, 6 and 12.

The forecasting results show that the Naïve model outperforms all other models from forecast horizons 1 through 12, with the Holt-Winters model generally ranking second and the baseline third. Among the models which include GT terms, Model 3 provides the best forecasts across all horizons, while Model 5 provides the worst. Yet, even Model 3 does not outrank the baseline model, nor the Holt-Winters or the Naïve models.

In both the baseline as well as the GT models (Models 2,3,4 and 5), a spiking behavior is found in  $h=1$  and  $h=2$  where both the RMSE and the MAE are higher than they are at the rest of forecasting horizons, and from  $h=3$  onwards, both statistics start at a lower value and then steadily increase, which is a more usual behavior. The spikes observed at the first two horizons could be due to a significant amount of noise being present in the prediction, given that the model fit is not incredibly high to begin with.

So all in all, because the models with a GT predictor do not outperform classical time series models such as Naïve or a non-seasonal Holt-Winters model, then it cannot be said that GT data improves predictions for rural overnights stays by national residents in Spain, at least not with this specification.

These results are consistent with those found by Önder (2017) in predicting arrivals for Belgium by using total searches, searches from the U.S. and searches from the U.K, where they find that the Naïve specification performs the best among all competing models for Belgium. Similarly, when Gunter et al. (2019) attempt to predict total tourist arrivals to four Austrian cities, they find that Naïve models generally outperform models which only include lags of the dependent variable and lags of GT as the main predictors. Moreover, Rivera (2016) also finds evidence that his Dynamic Linear Model (DLM) used to forecast non resident registrations on Puerto Rico hotels only outperforms other benchmarks models such as the ones included in this study when the horizon of forecasting is of 6 or above.

Table 5.4: Out of sample forecasting, forecast horizons 1, 2, 3, 6 and 12.

Model	h = 1		h = 2		h = 3		h = 6		h = 12	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Model 2	0.273	0.109	0.301	0.114	0.153	0.094	0.130	0.097	0.165	0.127
Model 3	0.097	0.075	0.098	0.074	0.089	0.068	0.098	0.072	0.128	0.094
Model 4	0.129	0.074	0.146	0.079	0.129	0.075	0.201	0.095	0.308	0.113
Model 5	0.926	0.205	0.903	0.193	0.374	0.117	0.105	0.076	0.243	0.119
Baseline	0.075	0.052	0.081	0.056	0.083	0.058	0.085	0.065	0.114	0.082
Naïve	0.046	0.032	0.047	0.032	0.042	0.030	0.042	0.029	0.044	0.030
Holt-Winters	0.052	0.043	0.054	0.045	0.056	0.047	0.056	0.049	0.066	0.060

### 5.5 Conclusions

The results obtained do not seem to back the initial hypothesis that using information from GT helps improve the predictions of monthly rural overnight stays of national residents in Spain. As seen in Section 5.4, the benchmark models outperform those which use information from GT.

One of the reasons for these results could be that GT does not capture enough of the search information made by tourists since not all users use the Google search engines to find information about their rural tourism activities.

Another reason could be that if they do, they do not search for 'rural tourism' or similar terms but rather for specific locations which they might desire to visit, hence queries which include more general terms such as the ones proposed in this study might be failing to capture the desired behavior. This can be seen in Subsection 5.3.3, where the Models with GT queries do not provide a much better fit than those without it. Moreover, even the models which do provide a better fit, Models 2, 3 and 4, presents adjusted  $R^2$  lower than 0.6, which means that more than 40% of the variability of the dependent variable is not being captured correctly through this specification, even if it has provided successful results to other authors in the past (Önder, 2017), albeit in a different context.

Yet, if users do in fact search for the specific location that they want to visit, using those locations as search terms would make it very difficult to provide an estimate of the searches for rural tourism in all of Spain, for several reasons: First, the elevated cost of tracking down all the locations and search terms that are relevant to the dependent variable. Secondly, because GT does not work well with low popularity queries (Cebrián and Domenech, 2024), it would be very difficult to obtain consistent estimates of the searches made for each location. Moreover, GT might report 0's if the search does not reach a particular popularity threshold (Cebrián and Domenech, 2023b), which would likely be the case for smaller rural destinations, and in this case, a GT index could not be constructed at all. Thirdly, even if the searches could be obtained for all queries, since GT does not provide the raw data of these searches, creating an aggregated index would not be feasible either.

It is also possible that part of the way in which information is transmitted is not directly through the internet but rather through word of mouth or other methods which cannot be captured by means of a search engine.

Finally, it could also be the case that GT does not work as well in predicting rural tourism variables as it does for other forms of tourism or for other kinds of variables for any of the above reasons or for others that may escape our knowledge.

Our work also faces some limitations: First, instead of combining search terms and categories, multiple search terms could have been combined to try to improve the forecasting results. Secondly, the results from this article are only applicable to the case of rural tourism overnight stays by national residents in Spain, and it is possible that results may vary under different conditions or under a different econometric specification.

Hence, future research should explore these findings under different contexts so that it can be

## **Chapter 5. Can Google Trends predict rural tourism? The case of Spain 85**

---

corroborated whether if GT can be useful in improving predictions for rural tourism or not.

### **Acknowledgements**

This work was partially supported by grants PID2019-107765RB-I00 and PEJ2018-003267-A-AR, funded by MCIN/AEI/10.13039/501100011033 and by “ESF Investing in your future”. The authors would like to thank Dr. Guadalupe Serrano and Dr. Carles Bretó for their contributions to this manuscript.





## 6. Conclusions

### 6.1 Main contributions

In the last few years, Big Data has taken a main role in our society because of its multitude of applications and therefore, its understanding has become paramount for both the public and the private sector. The sources which generate these data have allowed researchers to explore new possibilities and methodologies regarding human behavior. For these reasons, understanding how to treat, store, analyze and interpret these kind of data is one of the keys to moving forward as a society.

In this context, this thesis has focused on applying regression and forecasting techniques to gather how and to which extent Big Data can be useful in making decisions about tourism, and specially, rural tourism. To do so, this thesis has focused on four specific objectives.

The first objective of this thesis was to study the relevance of Big Data sources in the tourism sector. For this purpose, a classification of data sources has been proposed in the context of a Purchase-Consumption System (PCS) model applied to Travel and Leisure, which is a comprehensive representation of all the touristic process. In this context, a literature review was performed on the field of tourism. The main goal in this part of the thesis consisted in providing a classification of the literature reviewed based on two dimensions: the data sources in the study used and the part, or the 'stage' of the touristic process in which they are used. Through this investigation, a map of the Digital Footprint of a tourist has been produced, with the purpose of representing how the different data sources are used to understand the variables related to tourist behavior. Through this analysis it has been found that there is a notable focus on the decisions made by tourists while at the destination, particularly destination choices, as well as a preference for the use of Big Data sources from the Internet, specially Social Media. Moreover, the analysis also highlighted potential unexplored value in certain Big Data sources.

In the aforementioned classification, one of the main findings was the growing trend in the use of

Google Trends (GT) as a predictor to improve the forecasting and nowcasting of tourism demand. Hence, the second objective was to analyze the quality of GT data as a data source. To do so, the focus was shifted on to the replication of already existing results in the literature to check their consistency. However, in doing so, sampling issues related to the accuracy of the GT data arose, making it difficult to reproduce the results from other authors as the extractions from a same query changed significantly from day to day. Finally, some evidence has been provided about these accuracy issues and how the may affect forecasting results in a non-negligible manner.

Once these issues had been found, the third objective in this thesis was tackled: developing methods to ensure the quality of GT. In this context, some authors had tried to average different extractions to reduce the lack of accuracy. In the following work, we attempted to prove the validity of said hypothesis. However, because the sampling parameters GT are unknown, yet they introduce error, a way to find a solution was to create a simulation with population and sampling parameters which had been proven to be successful in the literature, and test the extent to which these lack of accuracy in the data could distort GT extraction. Through this experiment, it was found not only the extent to which these issues can go, but more importantly, that averaging extractions is indeed a useful solution to improve these accuracy issues. Finally, a relationship between the popularity of the search term, the number of extractions and its error has been established to help alleviate the issues.

Lastly, following the results from the previous steps, GT has been employed as a predictor to find out if it can improve the accuracy of forecasting of rural tourism demand in Spain. Hence, GT data was first retrieved, and then treated to alleviate its inconsistencies following previous results of this thesis, and then that same data was used as a predictor on ARDL models. Yet, it has been found that it does not improve the forecasting of rural tourism for Spain with respect to benchmark models, which is consistent to the results found in the tourism literature. However, it has not been established yet if this is due to the fact that not all users make use of the Google engine, or that users search for specific locations instead of more generic terms, or if GT data simply does not work well in predicting rural tourism.

## 6.2 Implications

In this thesis among others, a classification of Big Data sources in the touristic process as well as a qualitative and quantitative study of the quality of GT as a Big Data source have been provided. Furthermore, a solution to the accuracy issues found in GT data has been provided as well. Finally, these methods have been applied to rural tourism forecasting.

The relevance of these advances is that they should help create better practices in the handling of Big Data in the field of tourism, and therefore improve the decision-making with the data. These results have implications for the academic, the public and the private sector.

For academics, the classification of data sources into a PCS models offers a comprehensive map



of how different sources are employed in the different parts of the touristic process. In this sense, researchers can find what sources and methodologies have been used successfully in the past to explore their variable of interest. Moreover, the identification and proposal of a method to solve the problems related to the accuracy of GT data provides researchers with both the information necessary to proceed with caution when using GT data, and a way of dealing with GT data so that results in the studies can be consistent and reproducible.

In the public sector, this thesis provides a free to use, consistent data source that can help forecast the behavior of variables related to the topic of interest of the policy makers. But the most important implication is how timely this information is. In fact, some of the sources that were classified in the first part of the thesis, specially data from Social Media, can be obtained immediately after they are generated, which provides policy makers with tools to make faster, more timely decisions. In terms of GT, the existence of a method to obtain consistent and accurate data allows to track the interest of the population in a certain topic while also allows to forecast and nowcast variables regarding touristic and economic behavior that might be of relevant to policy-makers.

In the private sector, because of the rise of Big Data, each passing year more and more firms expand and invest in to their data and analytics departments. For this reason, the overall understanding of the touristic behavior becomes necessary in corporate planning for firms in the sector, and that is where the advances of this thesis can provide added value. More specifically, the classification of data sources provides detailed information as to how information from various sources can be used almost instantly to keep up with the trends of the market, as well as anticipate possible changes and take the corresponding actions. Similarly, the handling of GT in a more consistent manner allows the economic agents to reliably explore the interest of the market in the desired topics and it also allows them to use this data to also predict and adapt to future changes.

### 6.3 Limitations

While the research conducted in this thesis hopes to provide valuable knowledge to the touristic sector and apply it specifically to rural tourism, some limitations which have been found during the research process are listed below.

Firstly, none of the sources represented in the classification of data sources is presented in Chapter 2 are completely representative of the population. In the aforementioned Chapter, the data sources were mainly divided between those which use the internet and those which do not. The sources which do use the Internet, obviously do not represent the whole population given that not everyone uses the internet, where significant differences might be found depending on the income, age and country of residence of the person. And, for sources which do not require an internet connection typically other devices such as cards or mobile phones are required, which creates a similar bias since not everybody owns them. For example, Consumer Card Data can only be tracked for people who

own the specific card necessary for the study. Once more, this bias is related to differences in income, age and country of residence.

Secondly, following chapters of the thesis all have dealt with the methodologies and applications of GT to tourism and how it can be used to improve predictions in rural tourism demand. However, the information provided by GT only takes into account the searches made through Google, but not through any other search engine, and even if Google does have the biggest market share, only using Google data still creates a bias among internet users.

Thirdly, because Google does not disclose any of the parameters used during its sampling process when producing GT data, the results and solutions provided in this thesis are obtained through assumed parameters and functional forms that have been proposed in the literature, and that help create a simulation where our results come from. However, none of these parameters have been confirmed by Google itself, and hence, they results might differ if other parameters or functional forms are assumed.

## 6.4 Future Work

Based on the research developing during the writing of this thesis, several research questions arise.

First, the results obtained in Chapter 4 have been obtained assuming a specific functional form of the data, in this case, a sine wave function. However, we are yet to investigate if these results would be consistent if other functional forms were assumed for the data, which may constitute a potential line of research in the future. Moreover, there are plans to expand the work on the Google Trends methodologies by studying the impact that the different types of searches may have on the error of the query. For example, it is unknown whether if categories or if entities work better because they filter out unrelated searches. Conversely, search terms could be more precise. While we have not focused on these questions during the thesis, some basic analysis has been performed, but with no conclusive evidence and therefore, further analysis will be carried out to determine if indeed there is a pattern regarding the type of searches or not.

Moreover, in Chapter 5 it is expressed how Google Trends seemed not to work in improving forecasting in rural tourism. Now, this could be due to the low searches for the search terms chosen, because we did not include searches made through categories, or simply because indeed, Google Trends does not improve rural tourism forecasting. In light of these results, another line of investigation has emerged to give answers to these questions.

The last line of investigation that comes from the work in this thesis is the possible applications of Big Data sources to different areas, and specifically the application of Google Trends to the forecasting and nowcasting of variables from different fields. One of the benefits of Google Trends, as well as Social Media, is that it can provide timely information. So, if it can help predict the effect of public policies it holds potential to become a very relevant tool for policy makers. Therefore, the

last line of investigation originating from the thesis will focus on the capabilities of Big Data sources to predict the effects of public policies.





## Bibliography

- Abbasi, A., Rashidi, T. H., Maghrebi, M., and Waller, S. T. (2015). “Utilising Location Based Social Media in Travel Survey Methods: Bringing Twitter Data into the Play”. In *Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN’15*, New York, NY, USA. Association for Computing Machinery. <https://doi.org/10.1145/2830657.2830660>.
- Ahas, R., Aasa, A., Mark, U., Pae, T., and Kull, A. (2007). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management*, 28(3), 898 – 910. <https://doi.org/10.1016/j.tourman.2006.05.010>.
- Ahas, R., Aasa, A., Roose, A., Mark, U., and Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3), 469 – 486. <https://doi.org/10.1016/j.tourman.2007.05.014>.
- Angeloni, S. (2016). A tourist kit ‘made in Italy’: An ‘intelligent’ system for implementing new generation destination cards. *Tourism Management*, 52, 187 – 209. [10.1016/j.tourman.2015.06.011](https://doi.org/10.1016/j.tourman.2015.06.011).
- Antolini, F. and Grassini, L. (2019). Foreign arrivals nowcasting in Italy with Google Trends data. *Quality Quantity*, 53(5), 2385–2401. <https://doi.org/10.1007/s11135-018-0748-z>.
- Aramendia-Muneta, M. E., Olarte-Pascual, C., and Ollo-López, A. (2021). Key Image Attributes to Elicit Likes and Comments on Instagram. *Journal of Promotion Management*, 27(1), 50–76. <https://doi.org/10.1080/10496491.2020.1809594>.
- Arora, V. S., McKee, M., and Stuckler, D. (2019). Google Trends: Opportunities and limitations in health and health policy research. *Health Policy*, 123(3), 338–341. <https://doi.org/10.1016/j.healthpol.2019.01.001>.

- Asakura, Y., Iryo, T., Nakajima, Y., and Kusakabe, T. (2012). Estimation of behavioural change of railway passengers using smart card data. *Public Transport*, 4(1), 1 – 16. <https://doi.org/10.1007/s12469-011-0050-0>.
- Askitas, N. (2015). Google search activity data and breaking trends. *IZA World of Labor*, 206. <https://doi.org/10.15185/izawol.206>.
- Askitas, N. and Zimmermann, K. F. (2015). The internet as a data source for advancement in social sciences. *International Journal of Manpower*, 36, 2 – 12. <https://doi.org/10.1108/IJM-02-2015-0029>.
- Awan, M. J., Rahim, M. S. M., Nobanee, H., Munawar, A., Yasin, A., and Zain, A. M. (2021a). Social Media and Stock Market Prediction: A Big Data Approach. *Computers, Materials Continua*, 67(2), 2569–2583. <https://doi.org/10.32604/cmc.2021.014253>.
- Awan, U., Shamim, S., Khan, Z., Zia, N. U., Shariq, S. M., and Khan, M. N. (2021b). Big data analytics capability and decision-making: The role of data-driven insight on circular economy performance. *Technological Forecasting and Social Change*, 168, 120766. <https://doi.org/10.1016/j.techfore.2021.120766>.
- Bangwayo-Skeete, P. F. and Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management*, 46, 454–464. <https://doi.org/10.1016/j.tourman.2014.07.014>.
- Bantis, E., Clements, M. P., and Urquhart, A. (2023). Forecasting GDP growth rates in the United States and Brazil using Google Trends. *International Journal of Forecasting*. <https://doi.org/10.1016/j.ijforecast.2022.10.003>.
- Barreira, N., Godinho, P., and Melo, P. (2013). Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends. *NETNOMICS: Economic Research and Electronic Networking*, 14(3), 129–165. <https://doi.org/10.1007/s11066-013-9082-8>.
- Batini, C., Rula, A., Scannapieco, M., and Viscusi, G. (2015). From Data Quality to Big Data Quality. *Journal of Database Management*, 26(1), 60–82. <https://doi.org/10.4018/JDM.2015010103>.
- Batista E Silva, F., Marín Herrera, M. A., Rosina, K., Ribeiro Barranco, R., Freire, S., and Schiavina, M. (2018). Analysing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and Big Data sources. *Tourism Management*, 68, 101–115. <https://doi.org/10.1016/j.tourman.2018.02.020>.

- Becco, J. A., Huang, W.-J., Hallo, J. C., Norman, W. C., McGehee, N. G., McGee, J., and Goetcheus, C. (2013). GPS Tracking of Travel Routes of Wanderers and Planners. *Tourism Geographies*, 15(3), 551 – 573. <https://doi.org/10.1080/14616688.2012.726267>.
- Bellini, V., Guidolin, M., and Pedio, M. (2020). “Can Big Data Help to Predict Conditional Stock Market Volatility? An Application to Brexit”. In *2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, 398–409. [https://doi.org/10.1007/978-3-030-64583-0\\_36](https://doi.org/10.1007/978-3-030-64583-0_36).
- Bhatt, P. and Pickering, C. M. (2021). Public perceptions about Nepalese national parks: A global Twitter discourse analysis. *Society & Natural Resources*, 34(6), 685–702. <https://doi.org/10.1080/08941920.2021.1876193>.
- Blazquez, D. and Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, 130, 99–113. <https://doi.org/10.1016/j.techfore.2017.07.027>.
- Bokelmann, B. and Lessmann, S. (2019). Spurious patterns in Google Trends data - An analysis of the effects on tourism demand forecasting in Germany. *Tourism Management*, 75, 1–12. <https://doi.org/10.1016/j.tourman.2019.04.015>.
- Bordogna, G., Frigerio, L., Cuzzocrea, A., and Psaila, G. (2016). “Clustering geo-tagged tweets for advanced Big Data analytics”. In *2016 IEEE International Congress on Big Data (BigData Congress)*, 42 – 51. <https://doi.org/10.1109/bigdatacongress.2016.78>.
- Borup, D., Christian, E., and Schütte, M. (2022). In search of a job: Forecasting employment growth using Google Trends. *Journal of Business & Economic Statistics*, 40(1), 186–200. <https://doi.org/10.1080/07350015.2020.1791133>.
- Bound, J., Brown, C., and Mathiowetz, N. (2001). Measurement error in survey data. In Heckman, J. J. and Leamer, E., editors, *Handbook of Econometrics*, volume 5, 3705–3843. Elsevier.
- Boyd, D. and Crawford, K. (2013). Critical questions for Big Data. *Information, Communication Society*, 15(5), 662–679. <https://doi.org/10.1080/1369118X.2012.678878>.
- Brandt, T., Bendler, J., and Neumann, D. (2017). Social media analytics and value creation in urban smart tourism ecosystems. *Information & Management*, 54(6), 703 – 713. <https://doi.org/10.1016/j.im.2017.01.004>.
- Brewis, C., Dibb, S., and Meadows, M. (2023). Leveraging big data for strategic marketing: A dynamic capabilities model for incumbent firms. *Technological Forecasting and Social Change*, 190, 122402. <https://doi.org/10.1016/j.techfore.2023.122402>.

- Buhalis, D. and Volchek, K. (2021). Bridging marketing theory and big data analytics: The taxonomy of marketing attribution. *International Journal of Information Management*, 56, 102253. <https://doi.org/10.1016/j.ijinfomgt.2020.102253>.
- Butler, D. (2013). When Google got flu wrong. *Nature*, 494, 155–156. <https://doi.org/10.1038/494155a>.
- Bydon, M., Schirmer, C. M., Oermann, E. K., Kitagawa, R. S., Pouratian, N., Davies, J., Sharan, A., and Chambless, L. B. (2020). Big Data Defined: A Practical Review for Neurosurgeons. *World Neurosurgery*, 133, e842–e849. <https://doi.org/10.1016/j.wneu.2019.09.092>.
- Bíl, M., Bilová, M., and Kubeček, J. (2012). Unified GIS database on cycle tourism infrastructure. *Applied Geography*, 33(6), 1554 – 1561. <https://doi.org/10.1016/j.tourman.2012.03.002>.
- Böhme, M. H., Gröger, A., and Stöhr, T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, 142, 102–347. <https://doi.org/10.1016/j.jdeveco.2019.04.002>.
- Carrière-Swallow, Y. and Labbé, F. (2013). Nowcasting with Google Trends in an emerging market. *Journal of Forecasting*, 32(4), 289–298. <https://doi.org/10.1002/for.1252>.
- Carvache-Franco, O., Carvache-Franco, M., and Carvache-Franco, W. (2022). Coastal and marine topics and destinations during the COVID-19 pandemic in Twitter’s tourism hashtags. *Tourism and Hospitality Research*, 22(1), 32–41. <https://doi.org/10.1177/1467358421993882>.
- Casteli Gattinara, P., Froio, C., and Pirro, A. L. (2022). Far-right protest mobilisation in Europe: Grievances, opportunities and resources. *European Journal of Political Research*, 61, 1019–1041. <https://doi.org/10.1111/1475-6765.12484>.
- Castelnuovo, E. and Tran, T. D. (2017). Google It Up! A Google Trends-based Uncertainty index for the United States and Australia. *Economics Letters*, 161, 149–153. <https://doi.org/10.1016/j.econlet.2017.09.032>.
- Cebrián, E. and Domenech, J. (2023a). Digital footprint for tourism research. *Submitted*. <https://doi.org/10.21203/rs.3.rs-3591204/v1>.
- Cebrián, E. and Domenech, J. (2023b). Is Google Trends a quality data source? *Applied Economics Letters*, 30(6), 811–815. <https://doi.org/10.1080/13504851.2021.2023088>.
- Cebrián, E. and Domenech, J. (2024). Addressing Google Trends inconsistencies. *Technological Forecasting and Social Change*, 202, 123318. <https://doi.org/10.1016/j.techfore.2024.123318>.



- Chan, Y.-M. (1993). Forecasting tourism: A sine wave time series regression approach. *Journal of Travel research*, 32(2), 58–60. <https://doi.org/10.1177/00472875930320020>.
- Chang, W. and Grady, N. (2019). “NIST Big Data Interoperability Framework: Volume 1, Definitions”. [https://www.nist.gov/publications/nist-big-data-interoperability-framework-volume-1-definitions?pub\\_id=918927](https://www.nist.gov/publications/nist-big-data-interoperability-framework-volume-1-definitions?pub_id=918927) [Online; accessed: 26-02-2024].
- Cheng, M. and Jin, X. (2019). What do Airbnb users care about? An analysis of on-line review comments. *International Journal of Hospitality Management*, 76, 58 – 70. <https://doi.org/10.1016/j.ijhm.2018.04.004>.
- Choi, H. and Varian, H. (2009). Predicting initial claims for unemployment benefits. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=4d91786f9f88e0ec8dd5a25ca7c08f4d8e693b53> [Online; accessed: 29-04-2024].
- Choi, H. and Varian, H. (2012). Predicting the present with google trends. *Economic Record*, 88, 2 – 9. <https://doi.org/10.1111/j.1475-4932.2012.00809.x>.
- Chu, F.-L. (2004). Forecasting tourism demand: a cubic polynomial approach. *Tourism Management*, 25(2), 209–218. [https://doi.org/10.1016/S0261-5177\(03\)00086-4](https://doi.org/10.1016/S0261-5177(03)00086-4).
- Chua, A., Servillo, L., Marcheggiani, E., and Moore, A. V. (2016). Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy. *Tourism Management*, 57, 295 – 310. <https://doi.org/10.1016/j.tourman.2016.06.013>.
- Combes, S. and Bortoli, C. (2016). “Nowcasting with Google Trends, the more is not always the better”. In *1st International Conference on Advanced Research Methods in Analytics (CARMA 2016)*, 15–22. <https://doi.org/10.4995/CARMA2016.2016.4226>.
- Costola, M., Iacopini, M., and Santagiustina, C. R. (2021). Google search volumes and the financial markets during the COVID-19 outbreak. *Finance Research Letters*, 42, 101884. <https://doi.org/10.1016/j.frl.2020.101884>.
- Cox, M. and Ellsworth, D. (1997). “Application-controlled demand paging for out-of-core visualization”. In *Proceedings of the IEEE Visualization Conference*, 235–244. <https://doi.org/10.1109/VISUAL.1997.663888>.
- Da, Z., Engelberg, J., and Gao, P. (2011). In search of attention. *Journal of Finance*, 66, 1461–1499. <https://doi.org/10.1111/j.1540-6261.2011.01679.x>.
- Dalal, A. (2017). “Tourist destination recommendation system based on user Facebook profile”. <https://norma.ncirl.ie/2882/> [Online; accessed: 29-04-2024].

- D'Amuri, F. and Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, 33(4), 801–816. <https://doi.org/10.1016/j.ijforecast.2017.03.004>.
- Danila, D. and Gaceu, L. (2009). Online evaluation method for assessing the variation of the number of tourists interested in car rental. *Bulletin of the Transilvania University of Brasov*, 2(51), 75. [https://webbut.unitbv.ro/index.php/Series\\_II/article/view/1591](https://webbut.unitbv.ro/index.php/Series_II/article/view/1591) [Online; accessed: 29-04-2024].
- De Mauro, A., Greco, M., and Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, 65(3), 122–135. <https://doi.org/10.1108/LR-06-2015-0061>.
- Deepa, N., Pham, Q.-V., Nguyen, D. C., Bhattacharya, S., Prabadevi, B., Gadekallu, T. R., Maddikunta, P. K. R., Fang, F., and Pathirana, P. N. (2022). A survey on blockchain for big data: Approaches, opportunities, and future directions. *Future Generation Computer Systems*, 131, 209–226. <https://doi.org/10.1016/j.future.2022.01.017>.
- Del Giudice, M., Chierici, R., Mazzucchelli, A., and Fiano, F. (2021). Supply chain management in the era of circular economy: the moderating effect of big data. *The International Journal of Logistics Management*, 32(2), 337–356. <https://doi.org/10.1108/IJLM-03-2020-0119>.
- Delafontaine, M., Versichele, M., Neutens, T., and Van de Weghe, N. (2012). Analysing spatiotemporal sequences in Bluetooth tracking data. *Applied Geography*, 34, 659 – 668. <https://doi.org/10.1016/j.apgeog.2012.04.003>.
- Dergiades, T., Mavragani, E., and Pan, B. (2018). Google Trends and tourists' arrivals: Emerging biases and proposed corrections. *Tourism Management*, 66, 108 – 120. <https://doi.org/10.1016/j.tourman.2017.10.014>.
- Dijcks, J.-P. (2013). “Oracle: Big Data for the Enterprise”. <https://docplayer.net/252186-An-oracle-white-paper-june-2013-oracle-big-data-for-the-enterprise.html> [Online; accessed: 26-02-2024].
- Dilmaghani, M. (2019). Workopolis or The Pirate Bay: what does Google Trends say about the unemployment rate? *Journal of Economic Studies*, 46(2), 422–445. <https://doi.org/10.1108/JES-11-2017-0346>.
- Dinis, G., Costa, C., and Pachecho, O. (2017). Similarities and correlation between resident tourist overnights and Google Trends information in Portugal and its tourism regions. *Tourism & Management Studies*, 13(3), 15 – 22. <https://tmstudies.net/index.php/ectms/article/view/940/2351> [Online; accessed: 29-04-2024].

- Dinis, M. G. F., Costa, C. M. M., and Pacheco, O. R. (2016). Profile of the national Alentejo online visitor: analysis of the official tourism website using Google Analytics. *Tourism and Hospitality International Journal*, 6(1), 23 – 34. <https://www.cabidigitallibrary.org/doi/full/10.5555/20183224012> [Online; accessed: 29-04-2024].
- Donaire, J. A. and Galí, N. (2011). La imagen turística de Barcelona en la comunidad de Flickr. *Cuadernos de Turismo*(27), 291–303. <https://revistas.um.es/turismo/article/view/139961> [Online; accessed: 29-04-2024].
- Dumbill, E. (2013). Making Sense of Big Data. *Big Data*, 1(1), 1–2. <https://doi.org/10.1089/big.2012.1503>.
- Díaz, F., Henríquez, P., Hardy, N., and Ponce, D. (2023). Population well-being and the COVID-19 vaccination program in Chile: evidence from Google Trends. *Public Health*, 219, 22–30. <https://doi.org/10.1016/j.puhe.2023.03.007>.
- Eichenauer, V. Z., Indergand, R., Martínez, I. Z., and Sax, C. (2022). Obtaining consistent time series from Google Trends. *Economic Inquiry*, 60(2), 694–705. <https://doi.org/10.1111/ecin.13049>.
- EPRS (2023). “Rural tourism”. [https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/751464/EPRS\\_BRI\(2023\)751464\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/751464/EPRS_BRI(2023)751464_EN.pdf) [Online; accessed: 15-01-2024].
- Favaretto, M., De Clercq, E., Schneble, C. O., and Elger, B. S. (2020). What is your definition of Big Data? Researchers’ understanding of the phenomenon of the decade. *PLOS ONE*, 15(2). <https://doi.org/10.1371/journal.pone.0228987>.
- Filieri, R., Yen, D. A., and Yu, Q. (2021). I Love London: An exploration of the declaration of love towards a destination on Instagram. *Tourism Management*, 85, 104291. <https://doi.org/10.1016/j.tourman.2021.104291>.
- Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from twitter data. *Social Science Computer Review*, 31, 649 – 679. <https://doi.org/10.1177/0894439313493979>.
- Ghasemaghahi, M. and Calic, G. (2020). Assessing the impact of big data on firm innovation performance: Big data is not always better data. *Journal of Business Research*, 108, 147–162. <https://doi.org/10.1016/j.jbusres.2019.09.062>.
- Gidebo, H. B. (2021). Factors determining international tourist flow to tourism destinations: A systematic review. *Journal of Hospitality Management and Tourism*, 12(1), 9–17. <https://doi.org/10.5897/JHMT2019.0276>.

- Girardin, F., Calabrese, F., Fiore, F. D., Ratti, C., and Blat, J. (2008). Digital Footprinting: Uncovering Tourists with User-Generated Content. *IEEE Pervasive Computing*, 7(4), 36–43. <https://doi.org/10.1109/MPRV.2008.71>.
- Girardin, F., Dal Fiore, F., Blat, J., and Ratti, C. (2007). In *Understanding of Tourist Dynamics from Explicitly Disclosed Location Information.*, volume 58. [https://www.girardin.org/fabien/publications/girardin\\_dalfiore\\_blat\\_ratti\\_lbs2007\\_final.pdf](https://www.girardin.org/fabien/publications/girardin_dalfiore_blat_ratti_lbs2007_final.pdf) [Online; accessed: 29-04-2024].
- Goldstein, I., Spatt, C. S., and Ye, M. (2021). Big Data in Finance. *The Review of Financial Studies*, 34(7), 3213–3225. <https://doi.org/10.1093/rfs/hhab038>.
- Guaita Martínez, J. M., Martín Martín, J. M., Salinas Fernández, J. A., and Mogorrón-Guerrero, H. (2019). An analysis of the stability of rural tourism as a desired condition for sustainable tourism. *Journal of Business Research*, 100, 165–174. <https://doi.org/10.1016/j.jbusres.2019.03.033>.
- Gulati, S. (2022). Decoding the global trend of “vaccine tourism” through public sentiments and emotions: does it get a nod on Twitter? *Global Knowledge, Memory and Communication*, 71(8/9), 899–915. <https://doi.org/10.1108/GKMC-06-2021-0106>.
- Gunter, U. and Önder, I. (2016). Forecasting city arrivals with Google Analytics. *Annals of Tourism Research*, 61, 199 – 212. <https://doi.org/10.1016/j.annals.2016.10.007>.
- Gunter, U., Önder, I., and Gindl, S. (2019). Exploring the predictive ability of LIKES of posts on the Facebook pages of four major city DMOs in Austria. *Tourism Economics*, 25(3), 375 – 401. <https://doi.org/10.1177/1354816618793765>.
- Hasan, M. M., Popp, J., and Oláh, J. (2020). Current landscape and influence of big data on finance. *Journal of Big Data*, 7(1), 21. <https://doi.org/10.1186/s40537-020-00291-z>.
- Havranek, T. and Zeynalov, A. (2021). Forecasting tourist arrivals: Google Trends meets mixed-frequency data. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 27(1), 129–148. <https://doi.org/10.1177/1354816619879584>.
- Hossen, M. I., Goh, M., Hossen, A., and Rahman, M. A. (2020). “A Study on the Aspects of Quality of Big Data on Online Business and Recent Tools and Trends Towards Cleaning Dirty Data”. In *2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, 209–213. <https://doi.org/10.1109/ICSGRC49013.2020.9232648>.
- Hu, H., Tang, L., Zhang, S., and Wang, H. (2018). Predicting the direction of stock markets using optimized neural networks with Google Trends. *Neurocomputing*, 285, 188–195. <https://doi.org/10.1016/j.neucom.2018.01.038>.

- INE (2023). “Alojamientos de turismo rural: encuesta de ocupación e índice de precios. Últimos datos”. [https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736176963&idp=1254735576863](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176963&idp=1254735576863) [Online; accessed: 21-07-2023].
- Jackman, M. and Naitram, S. (2015). Research Note: Nowcasting Tourist Arrivals in Barbados – Just Google it! *Tourism Economics*, 21(6), 1309 – 1313. <https://doi.org/10.5367/te.2014.0402>.
- Jelnov, A. and Jelnov, P. (2022). Vaccination policy and trust. *Economic Modelling*, 108, 105773. <https://doi.org/10.1016/j.econmod.2022.105773>.
- Jepson, D. and Sharpley, R. (2015). More than sense of place? Exploring the emotional dimension of rural tourism experiences. *Journal of Sustainable Tourism*, 23(8-9), 1157–1178. <https://doi.org/10.1080/09669582.2014.953543>.
- Jun, S.-P., Yoo, H. S., and Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of Big Data utilizations and applications. *Technological Forecasting and Social Change*, 130, 69–87. <https://doi.org/10.1016/j.techfore.2017.11.009>.
- Kandula, S. and Shaman, J. (2019). Reappraising the utility of Google Flu Trends. *PLOS Computational Biology*, 15(8), 1–16. <https://doi.org/10.1371/journal.pcbi.1007258>.
- Karr, A. F., Sanil, A. P., and Banks, D. L. (2006). Data quality: A statistical perspective. *Statistical Methodology*, 3(2), 137–173. <https://doi.org/10.1016/j.stamet.2005.08.005>.
- Kassim, M. R. M. (2020). “IoT Applications in Smart Agriculture: Issues and Challenges”. In *2020 IEEE Conference on Open Systems (ICOS)*. <https://doi.org/10.1109/ICOS50156.2020.9293672>.
- Khanna, A. and Kaur, S. (2020). Internet of Things (IoT), Applications and Challenges: A Comprehensive Review. *Wireless Personal Communications*, 114(2), 1687–1762. <https://doi.org/10.1007/s11277-020-07446-4>.
- Khanra, S., Dhir, A., Islam, A. K. M. N., and Mäntymäki, M. (2020). Big data analytics in healthcare: a systematic literature review. *Enterprise Information Systems*, 14(7), 878–912. <https://doi.org/10.1080/17517575.2020.1812005>.
- Kim, G. S., Chun, J., Kim, Y., and Kim, C.-K. (2021). Coastal Tourism Spatial Planning at the Regional Unit: Identifying Coastal Tourism Hotspots Based on Social Media Data. *ISPRS International Journal of Geo-Information*, 10(3). <https://doi.org/10.3390/ijgi10030167>.
- Kim, K., Park, O.-j., Yun, S., and Yun, H. (2017). What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management. *Technological Forecasting and Social Change*, 123(6), 362 – 369. <https://doi.org/10.1016/j.techfore.2017.01.001>.

- Kirilenko, A. P. and Stepchenkova, S. O. (2017). Sochi 2014 Olympics on Twitter: Perspectives of hosts and guests. *Tourism Management*, 63, 54 – 65. <https://doi.org/10.1016/j.tourman.2017.06.007>.
- Kitchin, R. and McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data Society*, 133(1). <https://doi.org/10.1177/20539517166631130>.
- Knipe, D., Gunnell, D., Evans, H., John, A., and Fancourt, D. (2021). Is Google Trends a useful tool for tracking mental and social distress during a public health emergency? A time-series analysis. *Journal of Affective Disorders*, 294, 737–744. <https://doi.org/10.1016/j.jad.2021.06.086>.
- Kopetz, H. and Steiner, W. (2022). *Real-Time Systems: Design Principles for Distributed Embedded Applications*. Springer International Publishing. [https://doi.org/10.1007/978-3-031-11992-7\\_13](https://doi.org/10.1007/978-3-031-11992-7_13).
- Kumar, S., Valeri, M., and Shekhar (2022). Understanding the relationship among factors influencing rural tourism: a hierarchical approach. *Journal of Organizational Change Management*, 35(2), 385–407. <https://doi.org/10.1108/JOCM-01-2021-0006>.
- Kwok, L., Lee, J., and Han, S. H. (2022). Crisis communication on social media: What types of COVID-19 messages get the attention? *Cornell Hospitality Quarterly*, 63(4), 528–543. <https://doi.org/10.1177/19389655211028143>.
- Lane, B. and Kastenholz, E. (2015). Rural tourism: the evolution of practice and research approaches – towards a new generation concept? *Journal of Sustainable Tourism*, 23(8-9), 1133–1156. <https://doi.org/10.1080/09669582.2015.1083997>.
- Laney, D. (2001). “3-d data management: controlling data volume, velocity and variety”. <https://web.archive.org/web/20120304154148/https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf> [Online; accessed: 26-02-2024].
- Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6176), 1203–1205. <https://doi.org/10.1126/science.1248506>.
- Lee, M., Hong, J. H., Chung, S., and Back, K.-J. (2021). Exploring the roles of DMO’s social media efforts and information richness on customer engagement: Empirical analysis on Facebook event pages. *Journal of Travel Research*, 60(3), 670–686. <https://doi.org/10.1177/0047287520934874>.
- Li, W., Chai, Y., Khan, F., Jan, S. R. U., Verma, S., Menon, V. G., Kavita, and Li, X. (2021a). A comprehensive survey on machine learning-based big data analytics for iot-enabled smart healthcare

- system. *Mobile Networks and Applications*, 26(1), 234–252. <https://doi.org/10.1007/s11036-020-01700-6>.
- Li, X., Law, R., Xie, G., and Wang, S. (2021b). Review of tourism forecasting research with internet data. *Tourism Management*, 83, 104245. <https://doi.org/10.1016/j.tourman.2020.104245>.
- Li, X., Pan, B., Law, R., and Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism Management*, 59, 57 – 66. <https://doi.org/10.1016/j.tourman.2016.07.005>.
- Lippi, G., Nocini, R., and Henry, B. M. (2022). Analysis of online search trends suggests that SARS-CoV-2 Omicron (B.1.1.529) variant causes different symptoms. *Journal of Infection*, 84, 76–77. <https://doi.org/10.1016/j.jinf.2022.02.011>.
- Liu, C., Dou, X., Li, J., and Cai, L. A. (2020). Analyzing government role in rural tourism development: An empirical investigation from China. *Journal of Rural Studies*, 79, 177–188. <https://doi.org/10.1016/j.jrurstud.2020.08.046>.
- Liu, J., Li, J., Li, W., and Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 134–142. <https://doi.org/10.1016/j.isprsjprs.2015.11.006>.
- Liu, Y., Teichert, T., Rossi, M., Li, H., and Hu, F. (2017). Big Data for big insights: Investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews. *Tourism Management*, 59, 554 – 563. <https://doi.org/10.1016/j.tourman.2016.08.012>.
- Liu, Y.-L., Chiang, J.-T., and Ko, P.-F. (2023). The benefits of tourism for rural community development. *Humanities and Social Sciences Communications*, 10(1). <https://doi.org/10.1057/s41599-023-01610-4>.
- Lu, Y. and Zheng, Q. (2021). Twitter public sentiment dynamics on cruise tourism during the COVID-19 pandemic. *Current Issues in Tourism*, 24(7), 892–898. <https://doi.org/10.1080/13683500.2020.1843607>.
- Lutfi, A., Alrawad, M., Alsayouf, A., Almaiah, M. A., Al-Khasawneh, A., Al-Khasawneh, A. L., Alshira'h, A. F., Alshirah, M. H., Saad, M., and Ibrahim, N. (2023). Drivers and impact of big data analytic adoption in the retail industry: A quantitative investigation applying structural equation modeling. *Journal of Retailing and Consumer Services*, 70, 103–129. <https://doi.org/10.1016/j.jretconser.2022.103129>.
- Ma, S. D., Kirilenko, A. P., and Stepchenkova, S. (2020). Special interest tourism is not so special after all: Big Data evidence from the 2017 Great American Solar Eclipse. *Tourism Management*, 77, 104021. <https://doi.org/10.1016/j.tourman.2019.104021>.

- Malagón-Selma, P., Debón, A., and Domenech, J. (2023). Measuring the popularity of football players with Google Trends. *PLOS ONE*, 18(8), 1–21. <https://doi.org/10.1371/journal.pone.0289213>.
- Mariani, M. M., Mura, M., and Di Felice, M. (2018). The determinants of Facebook social engagement for national tourism organizations' Facebook pages: A quantitative approach. *Journal of Destination Marketing & Management*, 8, 312 – 325. <https://doi.org/10.1016/j.jdmm.2017.06.003>.
- Marine-Roig, E. and Clavé, S. A. (2015). Tourism analytics with massive user-generated content: A case study of Barcelona. *Journal of destination marketing & management*, 4(3), 162. <https://doi.org/10.1016/j.jdmm.2015.06.004>.
- Mavragani, A., Ochoa, G., and Tsagarakis, K. P. (2018). Assessing the methods, tools, and statistical approaches in Google Trends research: systematic review. *Journal of Medical Internet Research*, 20(11), e9366. <https://doi.org/10.2196>
- Mavragani, A. and Tsagarakis, K. P. (2016). YES or NO: Predicting the 2015 GReferendum results using Google Trends. *Technological Forecasting and Social Change*, 109, 1–5. <https://doi.org/10.1016/j.techfore.2016.04.028>.
- Mayer-Schönberger, V. and Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.
- McKercher, B., Shoval, N., Ng, E., and Birenboim, A. (2012). First and repeat visitor behaviour: GPS tracking and GIS Analysis in Hong Kong. *Tourism Geographies*, 14(1), 147 – 161. <https://doi.org/10.1080/14616688.2011.598542>.
- Medeiros, M. C. and Pires, H. F. (2021). The proper use of Google Trends in forecasting models. *arXiv preprint arXiv:2104.03065*. <https://doi.org/10.48550/arXiv.2104.03065>.
- Mellon, J. (2014). Internet Search Data and Issue Salience: The Properties of Google Trends as a Measure of Issue Salience. *Journal of Elections, Public Opinion and Parties*, 24(1), 45–72. <https://doi.org/10.1080/17457289.2013.846346>.
- Miah, S. J., Vu, H. Q., Gammack, J., and McGrath, M. (2017). A Big Data analytics method for tourist behaviour analysis. *Information & Management*, 54(6), 771–785. <https://doi.org/10.1016/j.im.2016.11.011>.
- Mishra, R. K., Urolagin, S., Jothi, J. A. A., Neogi, A. S., and Nawaz, N. (2021). Deep learning-based sentiment analysis and topic modeling on tourism during COVID-19 pandemic. *Frontiers in Computer Science*, 3. <https://doi.org/10.3389/fcomp.2021.775368>.



- Muryani, P., Mia, F., and Esquivias, M. A. (2020). Determinants of Tourism Demand in Indonesia: A Panel Data Analysis. *Tourism Analysis*, 25(1), 77–89. <https://doi.org/10.3727/108354220X15758301241666>.
- Narita, F. and Yin, R. (2018). In search of information: Use of Google Trends' data to narrow information gaps for low-income developing countries. *IMF Working Papers*, 2018(286). <https://doi.org/10.5089/9781484390177.001>.
- Newing, A., Clarke, G., and Clarke, M. (2014). Exploring small area demand for grocery retailers in tourist areas. *Tourism Economics*, 20(2), 407 – 427. <https://doi.org/10.5367/te.2013.0277>.
- Nicolas Woloszko, N. (2020). Tracking activity in real time with Google Trends. *OECD Economics Department Working Papers*(1634). <https://doi.org/https://doi.org/10.1787/6b9c7518-en>.
- Noolan, L. (2023). The role of culture as a determinant of tourism demand: evidence from European cities. *International Journal of Tourism Cities*, 9(1), 13–34. <https://doi.org/10.1108/IJTC-07-2021-0154>.
- Nooripoor, M., Khosrowjerdi, M., Rastegari, H., Sharifi, Z., and Bijani, M. (2021). The role of tourism in rural development: Evidence from Iran. *GeoJournal*, 86(4), 1705–1719. <https://doi.org/10.1007/s10708-020-10153-z>.
- Nuti, S. V., Wayda, B., Ranasinghe, I., Wang, S., Dreyer, R. P., Chen, S. I., and Murugiah, K. (2014). The use of Google Trends in health care research: A systematic review. *Journal of Elections, Public Opinion and Parties*, 9, 1–49. <https://doi.org/10.1371/journal.pone.0109583>.
- Oancea, B., Necula, M., Salgado, D., and Sanguiao, L. (2019). “ESSnet Big Data ii”. <https://oecd-opsi.org/wp-content/uploads/2021/04/MobilePhoneDataSimulator.pdf> [Online; accessed: 28-02-2024].
- OECD (2022). “OECD Tourism Trends and Policies 2022”. [https://read.oecd-ilibrary.org/industry-and-services/oecd-tourism-trends-and-policies-2022\\_a8dd3019-en#page1](https://read.oecd-ilibrary.org/industry-and-services/oecd-tourism-trends-and-policies-2022_a8dd3019-en#page1) [Online; accessed: 28-02-2024].
- Padhi, S. S. and Pati, R. K. (2017). Quantifying potential tourist behavior in choice of destination using Google Trends. *Tourism Management Perspectives*, 24, 44 – 47. <https://doi.org/10.1016/j.tmp.2017.07.001>.
- Padilla, J. J., Kavak, H., Lynch, C. J., Gore, R. J., and Diallo, S. Y. (2018). Temporal and spatiotemporal investigation of tourist attraction visit sentiment on Twitter. *PLoS One*, 13(6), 1 – 20. <https://doi.org/10.1371/journal.pone.0198857>.

- Palazzo, M., Vollero, A., Vitale, P., and Siano, A. (2021). Urban and rural destinations on Instagram: Exploring the influencers' role in sustainable tourism. *Land Use Policy*, 100, 104915. <https://doi.org/10.1016/j.landusepol.2020.104915>.
- Park, J. H., Lee, C., Yoo, C., and Nam, Y. (2016a). An analysis of the utilization of Facebook by local Korean governments for tourism development and the network of smart tourism ecosystem. *International Journal of Information Management*, 36(6, Part B), 1320 – 1327. <https://doi.org/10.1016/j.ijinfomgt.2016.05.027>.
- Park, S., Lee, J., and Song, W. (2017). Short-term forecasting of Japanese tourist inflow to South Korea using Google Trends data. *Journal of Travel & Tourism Marketing*, 34(3), 357–368. <https://doi.org/10.1080/10548408.2016.1170651>.
- Park, S. B., Jang, J., and Ok, C. M. (2016b). Analyzing Twitter to explore perceptions of Asian restaurants. *Journal of Hospitality and Tourism Technology*, 7(4), 405–422. <https://doi.org/10.1108/JHTT-08-2016-0042>.
- Paül i Agustí, D. (2018). Characterizing the location of tourist images in cities. Differences in user-generated images (Instagram), official tourist brochures and travel guides. *Annals of Tourism Research*, 73, 103–115. <https://doi.org/10.1016/j.annals.2018.09.001>.
- Paül i Agustí, D. (2021). Mapping gender in tourist behaviour based on Instagram. *Journal of Outdoor Recreation and Tourism*, 35, 100381. <https://doi.org/10.1016/j.jort.2021.100381>.
- Pelat, C., Turbelin, C., Bar-Hen, A., Flahault, A., and Valleron, A.-J. (2009). More diseases tracked by using Google Trends. *Emerging Infectious Diseases*, 15(8), 1327–1328. <https://doi.org/10.3201/eid1508.090299>.
- Philander, K. and Zhong, Y. (2016). Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. *International Journal of Hospitality Management*, 55, 16 – 24. <https://doi.org/10.1016/j.ijhm.2016.02.001>.
- Plaza, B. (2011). Google Analytics for measuring website performance. *Tourism Management*, 32(3), 477 – 481. <https://doi.org/10.1016/j.tourman.2010.03.015>.
- Popescu, A. and Grefenstette, G. (2009). “Deducing trip related information from Flickr”. WWW '09, p. 1183–1184, New York, NY, USA. Association for Computing Machinery. <https://doi.org/10.1145/1526709.1526919>.
- Preis, T., Moat, H. S., and Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports*, 3(1), 1684. <https://doi.org/10.1038/srep01684>.

- Provenzano, D., Hawelka, B., and Baggio, R. (2018). The mobility network of European tourists: A longitudinal study and a comparison with geo-located Twitter data. *Tourism Review*, 73(1), 28–43. [10.1108/tr-03-2017-0052](https://doi.org/10.1108/tr-03-2017-0052).
- Qian, C., Li, W., Duan, Z., Yang, D., and Ran, B. (2021). Using mobile phone data to determine spatial correlations between tourism facilities. *Journal of Transport Geography*, 92. <https://doi.org/10.1016/j.jtrangeo.2021.103018>.
- Raubenheimer, J. E. (2022). A practical algorithm for extracting multiple data samples from google trends extended for health. *American Journal of Epidemiology*, 191(9), 1666–1669. <https://doi.org/10.1093/aje/kwac088>.
- Raun, J., Ahas, R., and Tiru, M. (2016). Measuring tourism destinations using mobile tracking data. *Tourism Management*, 57, 202–212. <https://doi.org/10.1016/j.tourman.2016.06.006>.
- Rid, W., Ezeuduji, I. O., and Pröbstl-Haider, U. (2014). Segmentation by motivation for rural tourism activities in The Gambia. *Tourism Management*, 40, 102–116. <https://doi.org/10.1016/j.tourman.2013.05.006>.
- Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. *Tourism Management*, 57, 12–20. <https://doi.org/10.1016/j.tourman.2016.04.008>.
- Rosalina, P. D., Dupre, K., and Wang, Y. (2021). Rural tourism: A systematic literature review on definitions and challenges. *Journal of Hospitality and Tourism Management*, 47, 134–149. <https://doi.org/10.1016/j.jhtm.2021.03.001>.
- Rossi, L., Boscaro, E., and Torsello, A. (2018). “Venice through the lens of Instagram: A visual narrative of tourism in Venice”. In *Companion Proceedings of the The Web Conference 2018, WWW ’18*, p. 1190–1197, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee. <https://doi.org/10.1145/3184558.3191557>.
- Rovetta, A. (2021). Reliability of Google Trends: Analysis of the limits and potential of web infoveillance during COVID-19 pandemic and for future research. *Frontiers in Research Metrics and Analytics*, 6. <https://doi.org/10.3389/frma.2021.670226>.
- Saleh, I., Marei, Y., Ayoush, M., and Abu Afifa, M. M. (2023). Big Data analytics and financial reporting quality: qualitative evidence from Canada. *Journal of the Knowledge Economy*, 21(1), 83–104. <https://doi.org/10.1108/JFRA-12-2021-0489>.
- Salinas Fernández, J. A., Serdeira Azevedo, P., Martín Martín, J. M., and Rodríguez Martín, J. A. (2020). Determinants of tourism destination competitiveness in the countries most visited by

- international tourists: Proposal of a synthetic index. *Tourism Management Perspectives*, 33, 100582. <https://doi.org/10.1016/j.tmp.2019.100582>.
- Saxa, B. (2015). Forecasting mortgages: internet search data as a proxy for mortgage credit demand. *National Bank of the Republic of Macedonia*, 107. [https://www.nbrm.mk/WBStorage/Files/WebBuilder\\_Working\\_Paper\\_4th\\_Research\\_Conference.pdf#page=10](https://www.nbrm.mk/WBStorage/Files/WebBuilder_Working_Paper_4th_Research_Conference.pdf#page=10) [Online; accessed: 29-04-2024].
- Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., and Tufano, P. (2012). “Analytics: The Real-World Use of Big Data”. <https://www.bdvc.nl/images/Rapporten/GBE03519USEN.PDF> [Online; accessed: 26-02-2024].
- Scuderi, R. and Dalle Nogare, C. (2018). Mapping tourist consumption behaviour from destination card data: What do sequences of activities reveal? *International Journal of Tourism Research*, 20(5), 554 – 565. <https://doi.org/10.1002/jtr.2205>.
- Sestino, A., Prete, M. I., Piper, L., and Guido, G. (2020). Internet of Things and Big Data as enablers for business digitalization strategies. *Technovation*, 98, 102173. <https://doi.org/10.1016/j.technovation.2020.102173>.
- Shear, F., Ashraf, B. N., and Sadaqat, M. (2021). Are investors’ attention and uncertainty aversion the risk factors for stock markets? International evidence from the COVID-19 crisis. *Risks*, 9, 1461–1499. <https://doi.org/10.3390/risks9010002>.
- Shoval, N. and Isaacson, M. (2007). Sequence Alignment as a Method for Human Activity Analysis in Space and Time. *Annals of the Association of American Geographers*, 97(2), 282 – 297. <https://doi.org/10.1111/j.1467-8306.2007.00536.x>.
- Siliverstovs, B. and Wochner, D. S. (2018). Google Trends and reality: Do the proportions match?: Appraising the informational value of online search behavior: Evidence from Swiss tourism regions. *Journal of Economic Behavior & Organization*, 145, 1 – 23. <https://doi.org/10.1016/j.jebo.2017.10.011>.
- Simran and Sharma, A. K. (2023). Asymmetric impact of economic policy uncertainty on cryptocurrency market: Evidence from NARDL approach. *The Journal of Economic Asymmetries*, 27, e00298. <https://doi.org/10.1016/j.jeca.2023.e00298>.
- Singh, R. P., Javaid, M., Haleem, A., and Suman, R. (2020). Internet of Things (IoT), Applications and Challenges: A Comprehensive Review. *Diabetes Metabolic Syndrome: Clinical Research Reviews*, 14(4), 521–524. <https://doi.org/10.1016/j.dsx.2020.04.041>.

- Song, H., Qiu, R. T., and Park, J. (2019). A review of research on tourism demand forecasting: Launching the annals of tourism research curated collection on tourism demand forecasting. *Annals of Tourism Research*, 75, 338–362. <https://doi.org/10.1016/j.annals.2018.12.001>.
- Sontayasara, T., Jariyapongpaiboon, S., Promjun, A., Seelpipat, N., Saengtabtim, K., Tang, J., and Leelawat, N. (2021). Twitter sentiment analysis of Bangkok tourism during COVID-19 pandemic using support vector machine algorithm. *Journal of Disaster Research*, 16(1), 24–30. <https://doi.org/10.20965/jdr.2021.p0024>.
- Springer, S., Strzelecki, A., and Zieger, M. (2023). Maximum generable interest: A universal standard for Google Trends search queries. *Healthcare Analytics*, 3, 100158. <https://doi.org/10.1016/j.health.2023.100158>.
- Steinmetz, S., Bianchi, A., Tijdens, K., and Biffignandi, S. (2014). *Improving web survey quality*, chapter 12, 273–298. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118763520.ch12>.
- Stepchenkova, S., Kirilenko, A., and Kim, H. (2013). “Grassroots branding with Twitter: Amazing Florida”. In Cantoni, L. and Xiang, Z. P., editors, *Information and Communication Technologies in Tourism 2013*, 144 – 156, Berlin, Heidelberg. Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-36309-2\\_3](https://doi.org/10.1007/978-3-642-36309-2_3).
- Stephens-Davidowitz, S. and Varian, H. (2015). “A hands-on guide to Google Data”. <https://aeaweb.org/conference/2016/retrieve.php?pdfid=14330&tk=S7QBHGE>. [Online; accessed 16-October-2023].
- Sørensen, F. and Sundbo, J. (2014). Potentials for user-based innovation in tourism: the example of GPS tracking of attraction visitors. In *Handbook of Research on Innovation in Tourism Industries*, 132 – 153. Edward Elgar Publishing, Cheltenham, UK. <https://doi.org/10.18111/9789284422456>.
- Tang, M. and Xu, H. (2023). Cultural Integration and Rural Tourism Development: A Scoping Literature Review. *Tourism and Hospitality*, 4(1), 75–90. <https://doi.org/10.3390/tourhosp4010006>.
- Tavares, J. M. and Leitão, N. C. (2016). The determinants of international tourism demand for Brazil. *Tourism Economics*, 23(4), 834–845. <https://doi.org/10.5367/te.2016.0540>.
- Tenkanen, H., Di Minin, E., Heikinheimo, V., Hausmann, A., Herbst, M., Kajala, L., and Toivonen, T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Scientific Reports*, 7(1), 1 – 11. <https://doi.org/10.1038/s41598-017-18007-4>.

- Tiru, M., Kuusik, A., Lamp, M.-L., and Ahas, R. (2010). LBS in marketing and tourism management: Measuring destination loyalty with mobile positioning data. *Journal of Location Based Services*, 4(2), 120 – 140. <https://doi.org/10.1080/17489725.2010.508752>.
- Tudor, C. and Sova, R. A. (2023). Mining Google Trends data for nowcasting and forecasting colorectal cancer (CRC) prevalence. *PeerJ Computer Science*, 9, e1518. <https://doi.org/10.7717/peerj-cs.1518>.
- UNWTO (2010). “International Recommendations for Tourism Statistics 2008”. [https://unstats.un.org/unsd/publication/seriesm/seriesm\\_83rev1e.pdf](https://unstats.un.org/unsd/publication/seriesm/seriesm_83rev1e.pdf) [Online; accessed: 28-02-2024].
- UNWTO (2019). “UNWTO Tourism Definitions”. <https://www.e-unwto.org/doi/epdf/10.18111/9789284420858> [Online; accessed: 28-02-2024].
- UNWTO (2021). “International tourism highlights, 2020 edition”. <https://doi.org/10.18111/9789284422456> [Online; accessed: 28-02-2024].
- UNWTO (2023a). “Economic contribution and SDG”. <https://www.unwto.org/tourism-statistics/economic-contribution-SDG> [Online; accessed: 21-07-2023].
- UNWTO (2023b). “Tourism and Rural Development: A Policy Perspective”. <https://www.e-unwto.org/doi/10.18111/9789284424306> [Online; accessed: 28-02-2024].
- UNWTO (2023c). “World Tourism Barometer”. [https://webunwto.s3.eu-west-1.amazonaws.com/s3fs-public/2023-11/UNWTO\\_Barom23\\_04\\_November\\_EXCERPT\\_v2.pdf?VersionId=Q3i27HkRVsyU9gSP6yV4NCgxZiPdHrE](https://webunwto.s3.eu-west-1.amazonaws.com/s3fs-public/2023-11/UNWTO_Barom23_04_November_EXCERPT_v2.pdf?VersionId=Q3i27HkRVsyU9gSP6yV4NCgxZiPdHrE) [Online; accessed: 28-02-2024].
- Versichele, M., de Groote, L., Claeys Bouuaert, M., Neutens, T., Moerman, I., and Van de Weghe, N. (2014). Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: A case study of Ghent, Belgium. *Tourism Management*, 44, 67 – 81. <https://doi.org/10.1016/j.tourman.2014.02.009>.
- Versichele, M., Neutens, T., Delafontaine, M., and Van de Weghe, N. (2012). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent festivities. *Applied Geography*, 32(2), 208 – 220. <https://doi.org/10.1016/j.apgeog.2011.05.011>.
- Vicente, M. R., López-Menéndez, A. J., and Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? *Technological Forecasting and Social Change*, 92, 132 – 139. <https://doi.org/10.1016/j.techfore.2014.12.005>.
- Villamediana, J., Küster, I., and Vila, N. (2019). Destination engagement on Facebook: Time and seasonality. *Annals of Tourism Research*, 79, 102747. <https://doi.org/10.1016/j.annals.2019.102747>.

- Vogel, C., Zwolinsky, S., Griffiths, C., Hobbs, M., Henderson, E., and Wilkins, E. (2019). A Delphi study to build consensus on the definition and use of big data in obesity research. *International Journal of Obesity*, 43(12), 2573–2586. <https://doi.org/10.1038/s41366-018-0313-9>.
- Walker, A., Hopkins, C., and Surda, P. (2020). Use of Google Trends to investigate loss-of-smell-related searches during the COVID-19 outbreak. *International Forum of Allergy & Rhinology*, 10(7), 839–847. <https://doi.org/10.1002/alr.22580>.
- Wang, J., Liu, Y., Li, P., Lin, Z., Sindakis, S., and Aggarwal, S. (2023). Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality. *Journal of the Knowledge Economy*. <https://doi.org/10.1007/s13132-022-01096-6>.
- Wang, L.-e., Cheng, S.-k., Zhong, L.-s., Mu, S.-l., Dhruva, B. G. C., and Ren, G.-z. (2013). Rural tourism development in China: Principles, models and the future. *Journal of Mountain Science*, 10(1), 116–129. <https://doi.org/10.1007/s11629-013-2501-3>.
- Weaver, S. and Gahegan, M. (2007). Constructing, visualizing, and analyzing a digital footprint. *Geographical Review*, 97(3), 324–350. <https://doi.org/10.1038/s41366-018-0313-9>.
- Wijijayanti, T., Agustina, Y., Winarno, A., Istanti, L. N., and Dharma, B. A. (2020). Rural Tourism: A Local Economic Development. *Australasian Accounting, Business and Finance Journal*, 14(1), 5–13. <http://dx.doi.org/10.14453/aabfj.v14i1.2>.
- Willson, G., Wilk, V., Sibson, R., and Morgan, A. (2021). Twitter content analysis of the Australian bushfires disaster 2019–2020: Futures implications. *Journal of Tourism Futures*, 7(3), 350–355. <https://doi.org/10.1108/JTF-10-2020-0183>.
- Wilson, S., Fesenmaier, D. R., Fesenmaier, J., and Van Es, J. C. (2001). Factors for Success in Rural Tourism Development. *Journal of Travel Research*, 40(2), 132–138. <https://doi.org/10.1177/004728750104000203>.
- Witt, S. F. and Witt, C. A. (1995). Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting*, 11(3), 447–475. [https://doi.org/10.1016/0169-2070\(95\)00591-7](https://doi.org/10.1016/0169-2070(95)00591-7).
- Wong, K. K. (1997). The relevance of business cycles in forecasting international tourist arrivals. *Tourism Management*, 18(8), 581–586. [https://doi.org/10.1016/S0261-5177\(97\)00073-3](https://doi.org/10.1016/S0261-5177(97)00073-3).
- Wong, K. Y. and Wong, R. K. (2020). “Big Data Quality Prediction on Banking Applications: Extended Abstract”. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 791–792. <https://doi.org/10.1109/DSAA49011.2020.00119>.
- Wood, S. A., Guerry, A. D., Silver, J. M., and Lacayo, M. (2013). Using social media to quantify nature-based tourism and recreation. *Scientific Reports*, 3(1). <https://doi.org/10.1038/srep02976>.

- Woodside, A. G. and King, R. I. (2001). An updated model of travel and tourism Purchase-Consumption Systems. *Journal of Travel & Tourism Marketing*, 10(1), 3 – 27. <https://doi.org/10.1300/J073v10n0102>.
- Xi, C. and Donglai, C. (2022). Application of Improved Algorithm Based on Four-Dimensional ResNet in Rural Tourism Passenger Flow Prediction. *Journal of Sensors*, 2022, 555–571. <https://doi.org/10.1155/2022/9675647>.
- Xiang, Z. and Fesenmaier, D. R. (2017). Big data analytics, tourism design and smart tourism. *Analytics in Smart Tourism Design: Concepts and Methods*, 299–307. [https://doi.org/10.1007/978-3-319-44263-1\\_7](https://doi.org/10.1007/978-3-319-44263-1_7).
- Xue, M., Wu, H., Chen, W., Ng, W. S., and Goh, G. H. (2014). “Identifying tourists from public transport commuters”. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, p. 1779–1788, New York, NY, USA. Association for Computing Machinery. <https://doi.org/10.1145/2623330.2623352>.
- Yang, J., Yang, R., Chen, M.-H., Su, C.-H. J., Zhi, Y., and Xi, J. (2021). Effects of rural revitalization on rural tourism. *Journal of Hospitality and Tourism Management*, 47, 35–45. <https://doi.org/10.1016/j.jhtm.2021.02.008>.
- Yang, X., Pan, B., Evans, J. A., and Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management*, 46, 386 – 397. <https://doi.org/10.1016/j.tourman.2014.07.019>.
- Yang, Y., Fan, Y., Jiang, L., and Liu, X. (2022). Search query and tourism forecasting during the pandemic: When and where can digital footprints be helpful as predictors? *Annals of Tourism Research*, 93, 103365. <https://doi.org/10.1016/j.annals.2022.103365>.
- Yin, L. (2020). Forecast without historical data: objective tourist volume forecast model for newly developed rural tourism areas of China. *Asia Pacific Journal of Tourism Research*, 25(5), 555–571. <https://doi.org/10.1080/10941665.2020.1752755>.
- Yoshimura, Y., Sobolevsky, S., Ratti, C., Girardian, F., Carrascal, J. P., Blat, J., and Sinastra, R. (2014). An analysis of visitors’ behavior in the Louvre Museum: A Study Using Bluetooth data. *Environment and Planning B: Planning and Design*, 41(6), 1113 – 1131. <https://doi.org/10.1068/b130047p>.
- Yu, C.-E., Xie, S. Y., and Wen, J. (2020). Coloring the destination: The role of color psychology on Instagram. *Tourism Management*, 80, 104110. <https://doi.org/10.1016/j.tourman.2020.104110>.
- Yu, J. and Egger, R. (2021). Color and engagement in touristic Instagram pictures: A machine learning approach. *Annals of Tourism Research*, 89, 103204. <https://doi.org/10.1016/j.annals.2021.103204>.



- Zeni, N., Kiyavitskaya, N., Barbera, S., Oztaysic, B., and Mich, L. (2009). "RFID-based action tracking for measuring the impact of cultural events on tourism". In *Information and Communication Technologies in Tourism 2009*, 223 – 235. Springer Vienna. [https://doi.org/10.1007/978-3-211-93971-0\\_19](https://doi.org/10.1007/978-3-211-93971-0_19).
- Zenkner, G. and Navarro-Martinez, S. (2023). A flexible and lightweight deep learning weather forecasting model. *Applied Intelligence*, 53(21). <https://doi.org/10.1007/s10489-023-04824-w>.
- Zervoudakis, S., Marakakis, E., Kondylakis, H., and Goumas, S. (2021). Opinionmine: A Bayesian-based framework for opinion mining using Twitter data. *Machine Learning with Applications*, 3, 100018. <https://doi.org/10.1080/08941920.2021.1876193>.
- Zheng, W., Huang, X., and Li, Y. (2017). Understanding the tourist mobility using GPS: Where is the next place? *Tourism Management*, 59, 267 – 280. <https://doi.org/10.1016/j.tourman.2016.08.009>.
- Zheng, W., Zhou, R., Zhang, Z., Zhong, Y., Wang, S., Wei, Z., and Haipeng 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>.
- Zhu, L., Lin, Y., and Cheng, M. (2020). Sentiment and guest satisfaction with peer-to-peer accommodation: When are online ratings more trustworthy? *International Journal of Hospitality Management*, 86, 102369. <https://doi.org/10.1016/j.ijhm.2019.102369>.
- Önder, I. (2017). Forecasting tourism demand with Google Trends: Accuracy comparison of countries versus cities. *International Journal of Tourism Research*, 19(6), 648 – 660. <https://doi.org/10.1002/jtr.2137>.
- Önder, I., Gunter, U., and Gind, S. (2020). Utilizing Facebook statistics in tourism demand modeling and destination marketing. *Journal of Travel Research*, 59(2), 195 – 208. <https://doi.org/10.1177/0047287519835969>.

