

Mauricio Sepúlveda Cárdenas / Jaime Miranda Fierro /  
Mauricio Hidalgo Barrientos

# Generation of Domestic Tourism Travel Time Series Using Big Data From Mobile Phone Data

## Abstract

Domestic tourism is challenging to measure because it often blends with other forms of population movement. Additionally, traditional methods of measurement are resource-heavy and time-consuming. This paper proposes rules and processes for deriving time series of domestic tourism flows for overnight trips from passive mobile phone data (PMD) in Chile. The data generated are monthly, covering the period from January 2015 to March 2019. To verify the data's accuracy, we examined the tourism time series of Valparaíso, the most visited domestic region in the country. This analysis showed significant correlations with other time series. To gather insights about the time series, forecasting models are built using four different techniques. The SARIMA model provides a forecast with  $MAPE \leq 6.0\%$  and highlights the importance of is lag 1 and 12 for several models. In conclusion, the constructed time series displays notable correlations and matches with data from other sources, while the forecasting models behave as expected for tourism data. This indicates that the method is promising for generating reliable domestic tourism time series, saving costs and time, and increasing the frequency of updates.

**Keywords:** time series, mobile phone data, domestic tourism, forecasting, overnight trips

## 1. Introduction

Domestic tourism constitutes a significant economic driver; however, obtaining precise measurements remains a challenge. It encompasses the activities of resident visitors within the country, whether as part of domestic trips or outbound tourism (Tang et al., 2025; United Nations World Tourism Organization [UNWTO], 2008). Traditional approaches, typically reliant on surveys and censuses, are both costly and time intensive. Through the analysis of mobile device movement patterns, valuable insights into the behaviours of domestic tourists—including their origins, destinations, and durations of stay—can be acquired. Comprehending the dynamics of domestic tourism enables policymakers and industry stakeholders to make well-informed decisions regarding infrastructure development, marketing strategies, and resource allocation.

Tourism is a major and expanding sector in Chile. In 2019, it directly added CLP 6.5 billion to the economy, accounting for 3.3% of the total GDP. Tourism is also the second largest service export in the Chilean economy, and the market is forecast to grow at a compound annual growth rate of 5.26% from 2024 to 2029, reaching a market volume of US\$3,873.00 million in 2029 (Statista, 2024).

In Chile, the income generated by domestic tourism is estimated at CLP 3.9 trillion in 2018 (Organisation for Economic Co-operation and Development [OECD], 2024). Domestic tourism forms the backbone of tourism demand in Chile and in 2022 surpassed pre-pandemic levels, with 61.1 million domestic overnight visitors, 13.7% more than in 2019. Early figures for 2023 indicate that domestic tourism demand may weaken (7% below 2022 levels) but remain 22% above 2019 levels (OECD, 2024).

---

**Mauricio Sepúlveda Cárdenas**, Corresponding Author, Universidad San Sebastian, Recoleta, Santiago, Chile;  
ORCID ID: <https://orcid.org/0000-0002-5522-3194>; e-mail: [mauricio.sepulveda@uss.cl](mailto:mauricio.sepulveda@uss.cl)

**Jaime Miranda Fierro**, Universidad San Sebastian, Recoleta, Santiago, Chile; ORCID ID: <https://orcid.org/0000-0001-5707-9774>;  
e-mail: [Jaime.miranda@uss.cl](mailto:Jaime.miranda@uss.cl)

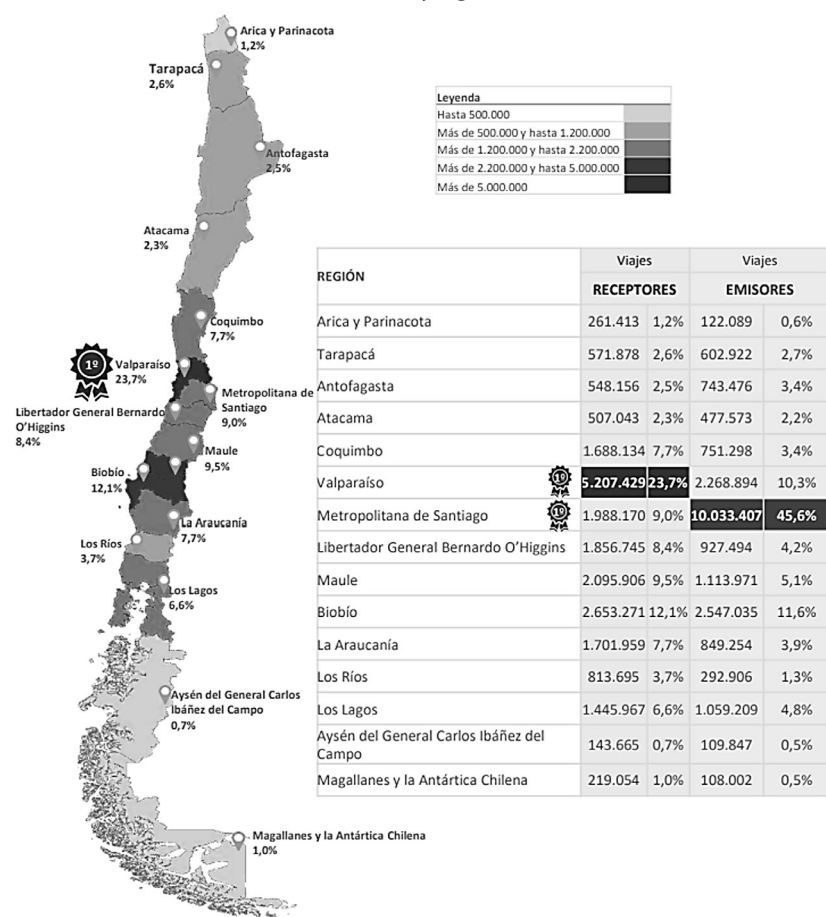
**Mauricio Hidalgo Barrientos**, Universidad Finis Terrae, Providencia, Región Metropolitana, Chile;  
ORCID ID: <https://orcid.org/0000-0003-3191-3673>; e-mail: [mhidalgo@uft.cl](mailto:mhidalgo@uft.cl)

Generally, the flow and economic impact of “domestic” tourism are estimated indirectly. The only direct studies of this activity for the domestic tourism market in Chile are those conducted by the National Tourism Service (Servicio Nacional de Turismo [SERNATUR], 2024a) and published by the Chilean Undersecretary of Tourism (Subsecretaría de Turismo de Chile, 2024). This information is gathered through costly field survey studies that are carried out over long periods. In 2018, a nationwide longitudinal study estimated that 11.6 million people residing in Chile took at least one trip with an overnight stay during the year. This represents 74.1% of the country's urban population, amounting to 22.0 million trips by residents in Chile (Subsecretaría de Turismo de Chile, 2018). Forty-two per cent of trips occur in high season, which spans 90 days of the year (the second half of December, January, February, and the first half of March). The expense and time involved in developing this survey-based study highlight the need for lower-cost, faster, and more accurate alternatives.

Studies of domestic tourism in Chile regard overnight stays as a crucial element. Figure 1 summarises the flows of inbound and outbound domestic trips with overnight stays in each region of the country for 2018. The Valparaíso Region is notable for having the highest number of inbound domestic tourism trips, accounting for 23.7% of this type of flow in the year studied.

The Metropolitan Region (RM) is the nation's capital area and the main hub of domestic tourism. The RM holds a third of the nation's population and is the nearest region to the tourist centers of the Valparaíso Region, at nearly 120 km.

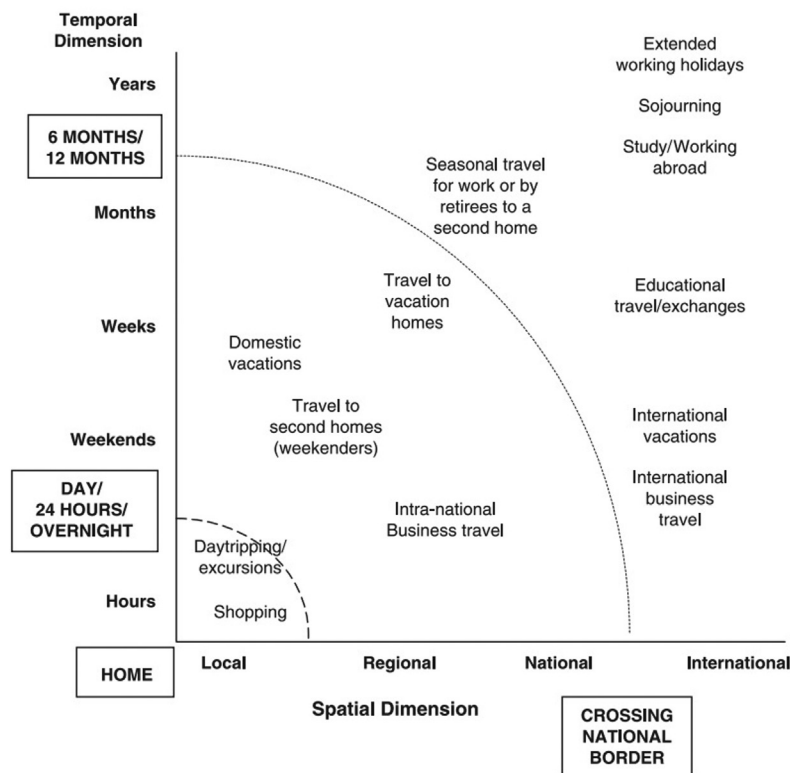
**Figure 1**  
*Inbound and outbound domestic tourism by region*



Source: Subsecretaría de Turismo de Chile (2018).

Recognizing movements as domestic tourism can sometimes be challenging and complex. Regarding mobility, Hall (2005) argues that tourism mobility should be understood within a spatial and temporal framework to distinguish different types of mobility. Figure 2 illustrates this idea by placing domestic holidays within timeframes of several weeks and on a regional or national scale. Movements lasting between six and twelve months might be considered migration. Weekend trips could relate to business or short stays at second homes. This suggests that only some factors, like time and space, differentiate forms of mobility and tourism. This understanding is essential for using mobile device data to identify domestic trips by estimating overnight stays outside the trip's origin area. The trip frequency also helps determine the purpose; regular weekly or monthly trips often indicate study, work, or business travel.

**Figure 2**  
*Temporary mobility in space and time*



Source: Hall (2005).

Consequently, domestic or internal tourism trips are difficult to determine, given the complexity of separating them from other types of population travel and from foreign tourists moving within the country.

Although the use and analysis of mobile device data are still relatively new, mobility studies have emerged (Okmi et al., 2023).

There are different mechanisms for capturing data from mobile devices. For example, active mobile positioning data involves tracking the location of the mobile phone through a specific query using radio waves. Conversely, passive mobile positioning data (PMD) is automatically recorded in mobile operators' log files and is generated only by user-initiated actions at the nearby cell tower of the respective operator (Ahas et al., 2008; C. Chen et al., 2016). Generally, the operator maintains records of location information, timestamps, mobile origin (indicating the country where the SIM card was registered), user data, and interactions between numbers (persons).

It is possible to estimate travel between regions or communes using data from the mobile device operator, as the mobile device log allows us to determine if a device has changed location, how long it stays in the new place, if it returns to the original location, and if this is the device's frequent location. Additionally, it helps identify if the movement occurs periodically, among other factors. Since a tourist consumer agent is generally someone who already owns a mobile device, the connection between a mobile device and a tourist appears logical.

Monthly data on the flow of domestic tourists, with details at inter-communal and regional levels, derived from mobile devices, would enable the government and local authorities to optimise investment and resource management. This data would allow them to accurately identify high-demand tourist corridors and emerging destinations. It would also facilitate the creation of more effective regional development policies focused on tourism decentralisation and improving road and service infrastructure where visitor flow justifies it.

Based on the above, the research hypothesis is that it is possible to develop a procedure to generate time series of domestic tourist flows from PMD data at the communal and regional levels. For this purpose, the time series must possess typical characteristics of tourism time series and show similarities with data from other sources. Finally, an attempt is made to extract information about the series using forecasting models.

This article continues with a literature review of tourism and mobile device data, followed by a chapter explaining the proposed data generation and validation. Finally, forecasting models are developed, and conclusions are presented.

## 2. Literature review

Numerous studies aim to understand tourist mobility using PMD. Ahas et al. (2008) identify seasonal movement patterns in Estonia, highlighting tourists of various nationalities. Similarly, Xu et al. (2021) determine tourists' countries of origin through mobile devices, assessing and characterising their trips across three South Korean cities over a year. Raun et al. (2020) analyse foreign tourism flows in Estonia by examining multiple data dimensions. Conversely, Saluveer et al. (2020) offer a statistical view of tourist flows, including overnight stay durations. In (N. C. Chen, et al., 2017a), classifiers predict tourists' next destinations in Andorra based on origin and behaviour. N. C. Chen et al. (2017b) focus on estimating tourist movements, vital for transportation planning. Kovacs et al. (2021) leverage social media (Twitter) and mobile network data to gain detailed insights into tourists' behaviour. Their 2023 study analyses six months of data across three Hungarian cities, revealing differences in tourist activity, mobility, and flow. Park et al. (2023) categorise travellers by analysing activity types within their trajectories. Overall, most research identifies tourist types, distribution patterns, and spatial-temporal relationships (Reif & Schmücker, 2020). These works mostly rely on short-term data, with few multi-season studies. For example, Saluveer et al. (2020), using three years of roaming data, produce tourism statistics for Estonia, including types of foreign tourists and overnight trends, correlating data with accommodation information. Additionally, Condeco-Melhorado et al. (2024) analyze COVID-19's impact on domestic tourism in Spain, focusing on travel from Madrid during the summers of 2019–2021.

Data derived from mobile devices serves as a valuable tool for identifying domestic tourism through specific usage characteristics, including the ability to detect whether a device uses a national SIM card. Using one-month Population Movement Data (PMD) in Saudi Arabia, the flow of domestic tourists between regions and to various tourist attractions has been established and compared with data obtained from face-to-face surveys (Alawwad et al., 2016). In (Vanhoof et al., 2017), potential tourism routes within France are proposed, employing an active mobile positioning database spanning 154 consecutive days, analysing duration of stays, and comparing findings with common routes. In (Mamei & Colonna, 2018), a classification method for tourist profiles was developed using mobile device data related to tourism within four regions of Italy over a nine-month period, employing machine learning techniques. Furthermore, in (Zaragozí et al., 2021), a flow

map depicting intra-urban and inter-urban mobility patterns of Spanish tourists visiting Salou from outside the province of Tarragona is presented, utilising mobile data collected over four days alongside demographic information of mobile users. Lastly, in (Olteanu et al., 2011), characteristics of trips, such as the place of origin and duration of overnight stays, are identified.

Obtaining data on domestic tourism travel flows from PMD is complex and faces criticism. The article by Reif et al. argues that PMDs can measure people's mobility in space and time, but they are not suitable for correctly identifying tourists and distinguishing them from non-tourists (Reif & Schmücker, 2020). Similarly, Grassini et al. (Grassini & Dugheri, 2021) use the city of Florence to demonstrate the challenges posed by data processing costs, privacy restrictions, and issues with statistical representativeness, such as foreigners purchasing local chips. Additionally, conducting surveys remains necessary to gather other types of information relevant to tourists. The impact of the COVID-19 pandemic also introduces a different set of difficulties in analysing these data (Condeço-Melhorado et al., 2024).

Several strategies aim to generate time series of tourist flows using big data, including search engine data, web traffic, and social media, among others (Li et al., 2021). Although these techniques show promising results, each has its advantages and disadvantages for time series estimation. For example, with search engine data, Google Trends is easy to use and free but still has some critical quality issues (Cebrián and Domenech, 2023). Since 2004, Google Trends has provided search frequency data for keywords, showing the comparison between the volume of searches for a given keyword and all other keywords over a specific period. It then normalises the search frequency within the range of 0 to 100 and enables filtering by country of origin (De Luca and Rosciano, 2024). In Chile, there are almost no formal records of domestic tourist flows, so the use of these strategies becomes essential. An exception could be data on the entry of tourists into protected reserves, since in these parks, the nationality and the dates of entry and exit of the visiting tourists are recorded. This data has vital validation, given that national tourists receive significant discounts.

There are limited studies and publications on domestic tourism in Chile. The difficulty and expense of systematically estimating domestic tourism flows and overnight stays mean these are often conducted in a targeted or sporadic manner, typically during comprehensive analyses of the country. Outbound and inbound tourism are easier to examine thanks to data from customs and airports. Overall, the articles suggest that time series forecasts of tourism using SARIMA models have performed very well (Colther and Arriagada-Millaman, 2021), (Mondaca-Marino et al., 2021). In conclusion, as technology advances and data collection methods become more sophisticated, the potential for deploying mobile device data to study domestic tourism will only increase. By addressing privacy concerns and data analysis challenges, it will be possible to make a significant contribution to the sustainable development of Chile's tourism industry.

### 3. Data and methods

The data shows a set of PMD records from one of Chile's major telecommunications providers, which, at the time of data collection, held 30% of the country's mobile device traffic, according to Chile's telecommunications regulator. Mobile network operators (MNOs) collect PMD for network management and billing purposes. To do this, they record, among other parameters, the approximate location of the phone based on the sector where the cellular network antenna it connected to is situated, which is linked to a commune or municipality in Chile. Communes are the smallest administrative divisions in Chile with defined territories. Conversely, a region comprises several communes, and Chile has 16 regions.

A method and set of guidelines are proposed to identify domestic tourism trips using PMD data. This includes, as standard, data integration, cleaning, and extraction processes. Data are queried from 1 January 2015 to 31 March 2019. The data are limited in that they associate each mobile device with a single tourist. The rules are listed below in their order of appearance.

- The data include PDM of mobile devices with and without a plan that have been in use for more than six months, based on the month extracted. This way, data is collected from the national population or residents of the country.
- A person's flow or trip is determined by whether there are location changes between communes for the mobile.
- For each flow, the commune of origin is identified by the most frequently recorded location of the mobile phone during the six months prior to data extraction. Since the user's declared location in the system might be outdated, and for phones without a plan, this information may be unknown (Reif & Schmücker, 2020).
- It is regarded as a day with an overnight stay if the mobile has not returned to the starting point before 00:00 and after 06:00 hrs. (Grassini & Dugheri, 2021).
- Flows where the mobile is away from its commune of origin for three or fewer overnight stays are excluded. This applies if they include working days during the week. This removes flows for business or shopping trips.
- Weekend flows without contiguous holidays are removed. This excludes non-tourist transfer trips.
- For each resulting flow, the destination commune is identified as the farthest location the mobile has visited before returning to the commune of origin (using a table of distances between communes). No flows are recorded for intermediate stops.
- The flows obtained are grouped by month and year. If a flow starts in one month and ends in another, it is assigned to the month in which it began.
- Repetitive flows are identified and removed, excluding trips for work, studies, second homes, or similar purposes. This involves determining which flow visits the same destination community three or more times annually.
- The mobile must return to its commune of origin within a maximum of 60 days to be regarded as a tourist flow.
- A validation by distance travel was not performed (Reif & Schmücker, 2020), as it was considered irrelevant whether it complies with the other rules.

The above rules were applied to the PMD data, resulting in a database (DB\_VF) for the entire country with the following structure: commune of origin, region of origin, commune of destination, region of destination, year, month, and number of trips. The big data is processed on the mobile operator's servers.

The information produced contains biases inherent to its technological origins. Some of the main biases associated with this type of data include the following:

- The data is a sample of the population.
- Antenna and communication systems failure, which prevents them from picking up signals during lapses of time.
- There are also issues with communications, power, signal blockages, and areas without coverage. This causes some communes not to submit data or for data to be recorded on different dates or in other communes. This kind of problem is complex to resolve because identifying the faults is difficult.

The data obtained is linearly scaled to manage it at the country level. When aggregated, the data is much smaller than the data published by SERNATUR on domestic tourism by area.

Therefore, correction factors are applied to the data in the DB\_VF database. These factors are based on the only available study of domestic tourism by SERNATUR in the same period, called "National cross-sectional

study” (Subsecretaría de Turismo de Chile, 2018). The data were collected through face-to-face surveys conducted between April 2017 and March 2018 and supplemented with information from the country’s 2017 Census (<http://resultados.censo2017.cl/>).

The calculation is a factor per region and involves dividing the total trips within the region from April 2017 to March 2018, as reported by the national cross-sectional survey, by the domestic tourism trips obtained from the DB\_VF base for the same period in the region. This constant factor enables the generation of the base BD\_VF2 from the base BD\_VF, thus allowing the storage of estimated numbers at the national level for domestic tourism trips over the 51 months of the study.

In summary, the representativeness of the data relies on PMD data from one of Chile’s major operators. This data covers 30% of the country’s mobile devices on telephone plans that are over six months old and are exclusively offered to residents. Moreover, correction factors and linear scaling—based on traditional tourism studies and censuses—are applied to enhance national and regional representativeness, indicating an effort to capture the full scope of domestic tourism. However, due to concerns about data confidentiality, the demographic composition of the 30% sample could not be verified for representativeness.

Nevertheless, it is considered acceptable based on the operator’s cross-sectional coverage of social and generational groups (offering a mid-range rate and service provision). Regarding potential biases, the representativeness mentioned above, failures in the communications infrastructure (such as antennas and coverage), and the assumption that each mobile device equates to one tourist stand out, as these can inflate or deflate the actual figures. Other biases arise from the criteria used to define a tourist trip and from the exclusion of trips that could be considered tourist trips but are mistakenly categorised as work, study, or second-home trips. Additionally, the requirement to return the mobile phone to its original location within 60 days at the latest excludes longer domestic tourism trips. These biases are not always straightforward to correct and demand careful interpretation of the results.

## 4. Data validation

The large volume of information necessitates delimiting the study. This enables us to concentrate on a specific time series and enhance data validation. We focus on data regarding inbound domestic tourism to Valparaíso, as it is the most relevant region for the study, being the one that records the highest flow of inbound domestic tourism. The data sourced from the BD\_VF2 database are shown in Table 1, under the two columns “Domestic trips from RM” and “Domestic trips from other regions,” which will be explained later. The total of both columns provides the “Total domestic trips to Valparaíso.” These figures are compiled monthly. The total inbound trips over the study period amount to 21,826,896.

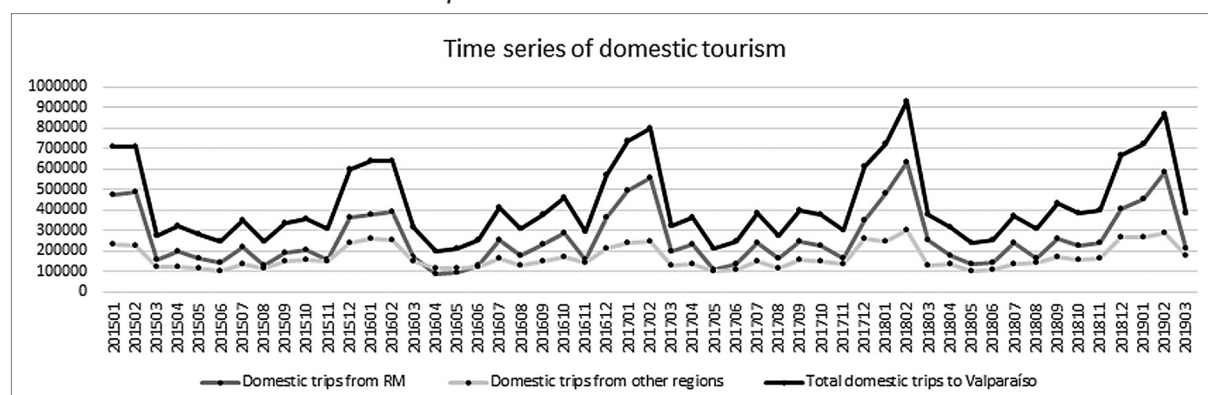
**Table 1**  
*Domestic tourism trips to Valparaíso*

Date	Domestic trips from RM	Domestic trips from other regions	Google Trends	SNASPE unit	Date	Domestic trips from RM	Domestic trips from other regions	Google Trends	SNASPE unit
201501	473,465	233,222	79	18,993	201702	554,765	245,716	85	14,658
201502	486,559	222,016	73	19,152	201703	196,046	125,127	51	10,078
201503	154,991	120,019	39	7,306	201704	231,190	132,370	40	11,855
201504	198,979	121,974	34	9,308	201705	110,123	101,313	43	8,892
201505	166,086	116,187	33	10,687	201706	136,958	105,768	41	8,755
201506	139,701	103,765	32	5,074	201707	240,344	146,527	50	12,171
201507	216,845	134,259	41	7,079	201708	161,742	115,075	42	10,671
201508	126,118	117,229	40	3,891	201709	246,374	154,785	47	12,957

**Table 1 (continued)**

201509	188,330	146,419	40	9,154	201710	226,613	149,046	46	17,821
201510	201,871	155,120	43	14,345	201711	165,956	135,453	48	11,213
201511	156,973	152,531	42	13,865	201712	352,365	258,853	60	9,555
201512	361,639	238,290	49	11,679	201801	478,697	243,659	85	14,354
201601	379,057	260,564	70	20,827	201802	634,507	298,420	96	18,252
201602	390,359	250,445	72	21,116	201803	249,825	130,223	51	10,381
201603	169,530	148,893	43	13,240	201804	179,168	134,425	44	8,729
201604	84,445	115,422	35	7,431	201805	138,643	102,776	44	10,792
201605	95,182	115,357	38	10,247	201806	142,144	110,509	39	4,928
201606	130,487	119,977	36	7,271	201807	235,821	137,284	51	8,107
201607	250,498	160,719	51	12,038	201808	161,596	143,200	43	7,098
201608	176,706	128,308	36	10,527	201809	258,582	170,625	50	12,234
201609	228,797	150,371	43	15,332	201810	226,465	159,332	52	13,484
201610	288,513	170,268	48	20,110	201811	238,006	162,451	53	11,783
201611	153,773	143,647	43	11,982	201812	405,061	264,884	57	8,318
201612	361,813	209,933	51	11,145	201901	452,575	267,538	87	12,940
201701	497,544	238,936	77	11,987	201902	583,205	285,816	91	18,485
					201903	208,158	178,660	54	8,832

**Figure 3**  
Time series of domestic tourism income to Valparaíso

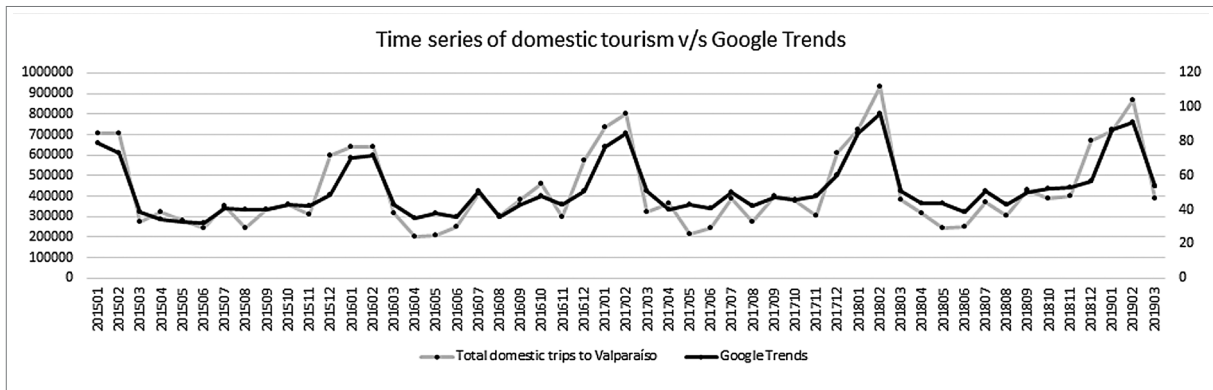


The “Total domestic trips to Valparaíso” data series cannot be precisely validated, as no local source offers accurate data on inbound tourism to Valparaíso throughout the entire study period. However, examining the graph of the series in Figure 3 reveals seasonality during the summer months, which include December, January, February, and March, when tourist flows increase, aligning with expectations. A similar rise is also observed during the National Holidays in September, the students' winter holidays in July, and at Easter, although to a lesser degree. The most notable drop occurs in May, when holiday travel is relatively scarce. The typical Z-Score and IQR methods do not identify any outliers.

When searching for any correlation between the data and another source, these are compared with information from Google Trends and data on tourist arrivals to the protected reserves of Valparaíso (SNASPE unit). The “Google Trends” column in Table 1 shows the downloaded data. The information is presented graphically in Figure 4. The data from Google Trends corresponds to a query for the proportion of searches for one or more terms on the Internet. To gather the data, a search was performed for the most popular tourist destinations in this region, namely: Valparaíso, Viña del Mar, Reñaca, El Quisco, El Tabo, Cartagena, Algarrobo, Los Andes and San Antonio. Additionally, the filter “by travel reasons” was applied, and the searches were conducted by users within Chile. This indicates the interest in travelling to the Valparaíso Region. The query

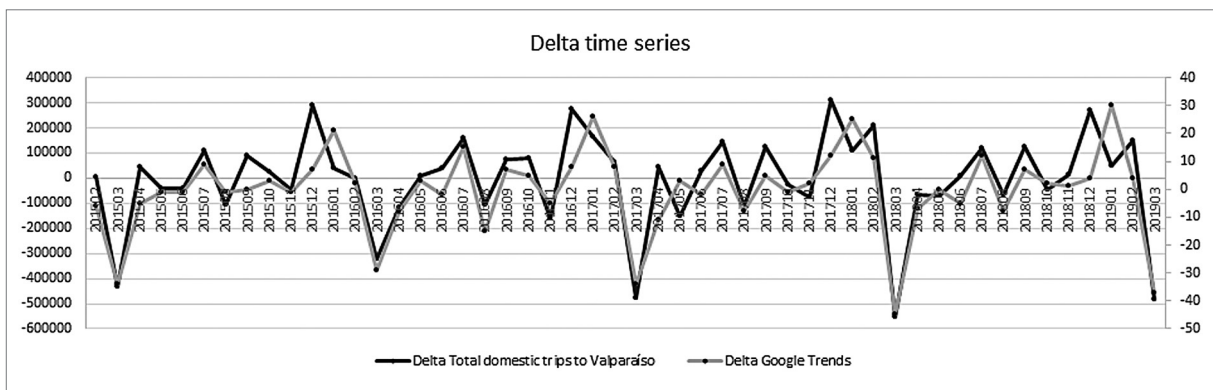
can be replicated with this link <https://trends.google.es/trends/explore?cat=67&date=all&geo=CL&q=valpara%20raiso%20Bvi%20C3%B1a%20del%20mar%20Bre%20C3%B1aca%20Bel%20quisco%20Bel%20tabo%20Bcartagen%20Balgarrobo%20Blos%20andes%20Bsan%20antonio&hl=es>

**Figure 4**  
Time series of domestic tourism income to Valparaíso v/s Google Trends



A correlation and a causal relationship are observed between the two series in their first difference ( $\Delta$ ). Figure 4 shows a strong similarity between both time series, in terms of stationarity and trend, although the Google Trends series is smaller in magnitude, comparatively. When analysing the first difference of both series using the *tseries* statistical package for  $R^1$ , it is observed that both series are stationary ( $p$ -value  $< 0.05$ ) for the Dickey-Fuller test, as shown in Table 2. To analyse causality, the Granger Causality test is used. The *lmtest* library of R is used with a lag equal to 2. The result is an F-test statistic equal to 4.7176 and a  $p$ -value equal to 0.0141, which allows us to reject the null hypothesis and indicate that there is a causal relationship. Figure 5 shows visually that the delta tourist flow values anticipate the results of the Google Trend data, so it is estimated that the causal relationship is in that sense and that there is a correlation between the time series.

**Figure 5**  
Time series for Delta Total domestic trips to Valparaíso v/s Delta Google Trends



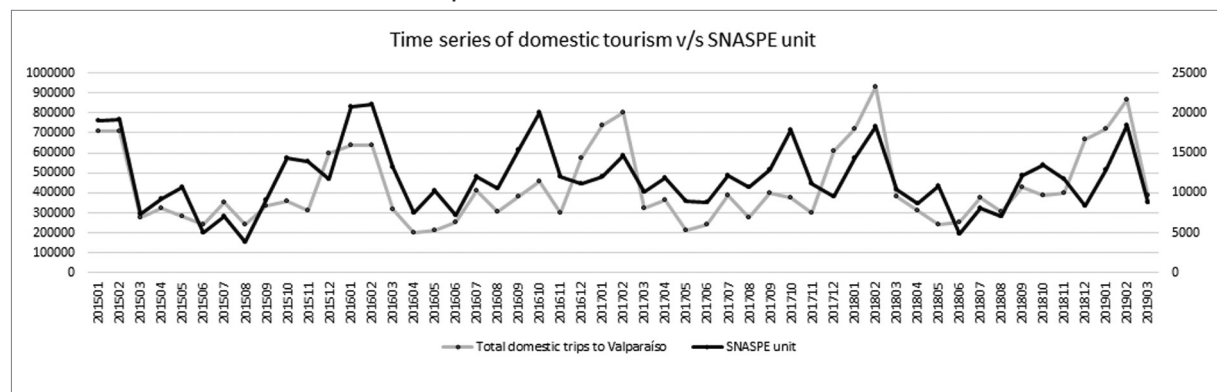
<sup>1</sup> <https://www.r-project.org/>

**Table 2**  
*Dickey-Fuller test for differentiated series*

Serie	Test statistic	p-value
Δ Total domestic trips to Valparaíso	-14.877	8.937e-19
Δ Google trends	-11.809	1.629e-24
Δ SNASPE unit	-5.898	2.808e-04
Δ Domestic trips from RM	-3.594	0.0058
Δ Domestic trips from other regions	-11.921	5.027e-22

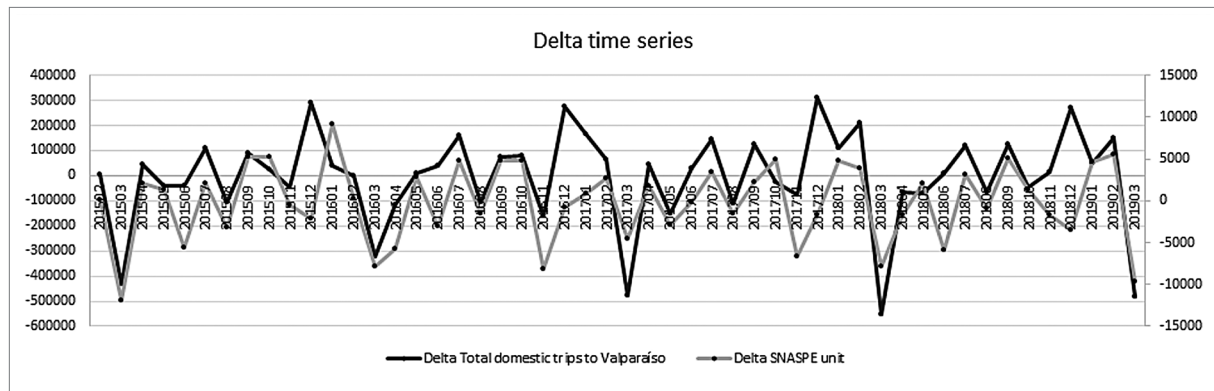
Data on tourist arrivals to the protected reserves of Valparaíso are also used to validate the time series. These areas have different entrance fees for foreigners and nationals and therefore record the number of visits by domestic tourists. In addition, the number of visits by Chilean tourists entering the country since 2012 is recorded and published monthly by the Reporte SNASPE (SERNATUR, 2024b). The protected reserves belonging to the SNASPE unit of Valparaíso include Isla Cachagua, Archipiélago Juan Fernandez, La Campana, Rapa Nui, El Yali, Lago Peñuelas, Río Blanco, Jardín Botánico, and Laguna El Peral. Among the disadvantages of using this information is the fact that the same tourist can repeat or visit several of these reserves on the same trip. Additionally, several of them require means of access, which could better segment wealthier tourists. The SNASPE unit column in Table 1 presents this information.

**Figure 6**  
*Time series of Domestic Tourism income to Valparaíso v/s SNASPE unit*



A correlation and a causal relationship are observed between the two series in their first difference. Figure 6 shows some similarities between the two-time series in terms of seasonality, although not in the trend. It also highlights that the peak of the 2017 series is different in the SNASPE unit series, which is in October 2016, and it does not happen in the summer months as expected. The explanation lies in La Campana National Park, which doubled its visits in October 2016 thanks to improvements in its infrastructure, after a sharp decline following forest fires in the summer. When analysing the first difference of both series using the tseries statistical package for R, it is observed that both series are stationary ( $p$ -value  $< 0.05$ ) for the Dickey-Fuller test, as shown in Table 2. In the causality analysis using the Granger Causality test, with a lag of 3, the results yield an F-value of 4.5371 and a  $p$ -value of 0.0079. This allows us to reject the null hypothesis and indicate that there is a causal relationship. Figure 7 shows the time-series delta. Obtaining a higher lag than in the previous experiment only suggests that the causal effect of one variable on another is manifested with a certain time lag. That is, changes in one variable do not have an immediate impact on the other, but their effects are observed after a greater number of periods.

**Figure 7**  
Time series for Delta domestic trips v/s Delta SNASPE unit



It is concluded that the Google Trends and SNASPE unit time series confirm an expected behaviour for the Total Domestic Trips time series, both seasonal and the differential between months, but not the trend, which is more consistent with the Google Trends series.

## 5. Analysis and forecasting

The study analysed the time series “Total domestic trips to Valparaíso”, decomposing it into two series: “Domestic trips from RM” and “Domestic trips from other regions”. Both series were modelled using four forecasting models, with the intention of understanding their predictive capacity and discovering patterns in the data.

The original time series was split into two because the behaviour of tourist flows from regions outside the RM is relatively similar; it was decided to create a time series for these regions and another for the MR. Both are presented in Table 1. Thus, the column “Domestic trips from RM” presents the flow of trips whose origin is the communes of the RM and whose destination is Valparaíso. The column “Domestic trips from other regions” corresponds to the sum of the time series of the other regions (communes) plus the flow of domestic tourism trips between communes of the Valparaíso Region (which is comparatively small). The three-time series is presented in Figure 3.

A Linear Regression with Seasonal Component (MLR) model is created for each of the two-time series that make up the “Total domestic trips to Valparaíso” series. A SARIMA model is also created, due to its good results in predicting tourism time series. Additional comparisons are made with two forecasting models whose implementation is simpler, since neither requires a stationary series for their application and both work quite well with time series with annual seasonality. The models are the exponential smoothing model called Holt-Winters (Holt, 2004) and the Machine Learning model known as LSTM neural networks (Hochreiter & Schmidhuber, 1997; Muangprathub et al., 2025).

To create the prediction models, the data is divided into two parts: the first 48 months for training and the last 3 months, during the high season, for testing. The test months would be 2019/01, 2019/02 and 2019/03. To evaluate the different techniques with the training data, the following measures of accuracy are calculated: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and MAPE (Mean Absolute Percentage Error)

MLR models are implemented using the timeseries Forecasting 1.0.27 package of the WEKA<sup>2</sup> tool using the following command: “LinearRegression -S 0 -R 1.0E-8 -additional-stats -num-decimal-places 8”. The results of the experiment are presented in Table 3.

<sup>2</sup> <https://www.cs.waikato.ac.nz/ml/weka/>

**Table 3**  
*Results from MLR*

Serie	Model	RMSE	MAE	R <sup>2</sup>	MAPE	Residual Med	Ljung-Box	Shapiro Wilk
Domestic trips from RM	0.21723216 * T <sub>-1</sub> + 0.86271263 * T <sub>-12</sub> + -11299.45735346	57,384	45,026	0.64	20.42%	2,611	0.2501	0.8688
Domestic trips from other regions	0.14981631 * T <sub>-1</sub> + -0.24913705 * T <sub>-3</sub> + -0.2402964 * T <sub>-6</sub> + -0.2261302 * T <sub>-9</sub> + 0.71511747 * T <sub>-12</sub> + 140258.31315529	19,049	15,544	0.75	9.95%	2,595	0.8657	0.6493

Table 3 displays the results of the MLR models for the period from January 2016 to December 2018 for each time series. The R<sup>2</sup> values indicate that both models do not fully account for the behaviour of the series. Model diagnostics confirmed acceptable performance, with no significant autocorrelation (Ljung-Box test) and normality verified by the Shapiro-Wilk test. All model coefficients were statistically significant (p-value < 0.05), although some heteroscedasticity was observed due to outliers. When examining the MRL models, we observe effects for lags-1 and-12, with the latter showing high significance in its coefficient. Additionally, the MRL model for the “Domestic trips from other regions” series requires consideration of lags-3, -6, and-9. These lags in tourism throughout the year correspond to summer holidays and long holidays such as Easter, winter holidays, and national celebrations. Notably, this series accounts for 39% of domestic tourists entering the region during the studied period, while the “Domestic trips from RM” series accounts for 61% of tourists entering the region. The results of the SARIMA model are summarised in Table 4. It was previously verified that both series are stationary at their first differences, with an ADF test p-value < 0.01, see Table 2. To obtain the optimal parameters for the SARIMA model, the auto-ARIMA function from the forecast library of the R software was used. In Table 4, the indicators associated with the residuals are calculated over the period from January 2016 to December 2018. The models were statistically validated, with all estimated parameters deemed significant. Diagnostic tests confirmed the models' reliability, as evidenced by the Ljung-Box Q-test p-values and the Shapiro-Wilk test verifying the normality of the residuals. Furthermore, the Dickey-Fuller test confirmed that the residuals of both models are white noise. However, the models exhibited varying error levels: while the “Domestic trips from RM” model performed poorly (MAPE approx. 16.60%), the “Domestic trips from other regions” model achieved a MAPE below 9.93%, making it useful for series understanding despite its high error measures.

**Table 4**  
*Results from SARIMA*

Serie	Model	RMSE	MAE	MAPE	Residual Med	Ljung-Box	Shapiro Wilk
Domestic trips from RM	ARIMA(1,0,0)(0,1,1)[12]	47,288	35,812	16.60	9,370	0.5569	0.7953
Domestic trips from other regions	ARIMA(1,1,0)(0,1,1)[12]	17,859	14,421	9.20	-682.1	0.4632	0.3095

The SARIMA model for both series considers the lag 1 and the annual differencing of order 1 to be relevant, as well as an annual seasonal moving average error component of order 1. Additionally, the “Domestic trips from other regions” series considers the monthly differentiation of order 1 as part of the model.

The forecast for the 3-month peak season for the models is presented in Table 5.

Table 5 illustrates the strong performance of the SARIMA model. A notable difference is seen in the “Domestic trips from RM” time series, with a MAPE of 4.1%. A similar forecast is observed for SARIMA in

the “Domestic trips from other regions” time series, with a MAPE of less than or equal to 6.0%. However, the SARIMA model is not entirely dependable due to the errors it exhibited during training. An exception is the “Domestic Trips from Other Regions” time series, where Holt-Winters outperforms by one-tenth of a percentage point.

**Table 5**  
*Forecasting*

Serie	Period	Real	MAPE MRL	MAPE SARIMA	MAPE Holt-Winters	MAPE LSTM
Domestic trips from RM	201901	452,575	27.8%	4.1%	10.0%	16.2%
	201902	583,205				
	201903	208,158				
Domestic trips from other regions	201901	267,538	9.1%	6.0%	5.9%	14.5%
	201902	285,816				
	201903	178,660				

It is concluded that both MLR and SARIMA consider lag 1 and 12 as important, and that the SARIMA model provides a better representation of the time series for forecasting purposes.

### 5.1. Ethical and privacy considerations

Using passive data from mobile phone records raises significant ethical concerns about individual privacy and the responsible handling of personal data. This study uses data anonymisation to guarantee irreversible de-identification, mitigating the risk of re-identification (de Montjoye et al., 2013; Zang & Bolot, 2011).

In the Chilean legal framework, strict adherence to Law 19.628 on the Protection of Private Life is essential (Biblioteca del Congreso Nacional de Chile [BCN], 1999), as it governs the processing of personal data — including anonymised data, the information that can be stored and its duration — if indirect identification remains feasible. It is also recommended to align research practices with international regulatory frameworks such as the European Union's General Data Protection Regulation (GDPR) (European Union, 2016), which establishes principles like data minimisation, purpose limitation, and researcher accountability. A privacy-by-design approach should be implemented, integrating data protection into the project's methodological structure (Cavoukian, 2012), including privacy impact assessments (PIAs), independent ethical audits, and algorithmic transparency. In summary, this research adopts robust ethical safeguards that not only comply with current legal standards but also foster public trust, protect digital rights, and enhance the scientific credibility of the study.

## 6. Conclusions

This study demonstrates that mobile phone data (PDM) is a valuable tool for analysing and modelling domestic tourism flows in Chile. The data collected enables the construction of time series at the communal and monthly levels, offering greater geographic and temporal variability compared to traditional approaches.

To evaluate the data, the study focused on the Valparaíso Region, which records the highest number of domestic tourist visits. The data confirmed significant correlations with the percentage of visits to the region's tourist centres, as obtained from Google Trends, as well as with data on visits to protected reserves in the Valparaíso Region. According to the Granger Causality test, there was even evidence of causality. However, the quality of mobile phone data can be affected by factors such as device usage patterns, uneven network coverage, and potential biases in representativeness. Data privacy regulations also limit access to detailed user information, which may influence analyses.

To gather information about the time series, forecasting models are constructed. The SARIMA model, especially for the time series tracking tourist flows between communes in the Valparaíso Region, showed the best performance, with minimal errors in the forecast for high season ( $MAPE \leq 6.0\%$ ). This indicates that this method is appropriate for predicting short-term trends. MLR and SARIMA models consider lags 1 and 12 as significant.

Within the limitations of this work, it is observed that while mobile phone data offers a rich source of information, challenges remain in ensuring data quality and accuracy. Factors like device usage patterns, network coverage, and data privacy regulations can influence the reliability of the data.

As a suggestion for future work, it is recommended to collaborate with mobile network operators to obtain more detailed data, including information on user demographics and behaviour. Additionally, it is advised to use surveys and other methods to gather baseline data and validate time series, as well as to incorporate external factors such as economic indicators, weather data, and social media sentiment into the prediction models.

---

### Declaration of Competing Interests

The authors declare that they have no known competing interests that could have appeared to influence the work reported in this paper.

## References

- Ahas, R., Aasa, A., Roose, A., Mark, Ü., & 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>
- Alawwad, Z., Alkowiileet, W., Alsubaiee, S., Alfaris, A., Alansari, M., & Ragab, A. M. (2016). Analyzing call detail records to support tourism statistics in Saudi Arabia: An exploratory study. *14th Global Forum on Tourism Statistics*, 23–25.
- Biblioteca del Congreso Nacional de Chile. (1999). *Ley 19.628: Sobre protección de la vida privada*. [Law 19.628: On the protection of private life]. <https://www.bcn.cl/leychile/navegar?idNorma=141599>
- Cavoukian, A. (2012). Privacy by design: Origins, meaning, and prospects for assuring privacy and trust in the information era. In G. Yee (Ed.), *Privacy Protection Measures and Technologies in Business Organizations: Aspects and Standards* (pp. 170–208). IGI Global Scientific Publishing. <https://doi.org/10.4018/978-1-61350-501-4.ch007>
- Cebrián, E., & Domenech, J. (2023). Is Google Trends a quality data source? *Applied Economics Letters*, 30(6), 811–815. <https://doi.org/10.1080/13504851.2021.2023088>
- Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, 68, 285–299. <https://doi.org/10.1016/j.trc.2016.04.005>
- Chen, N. C., Xie, W., Welsch, R. E., Larson, K., & Xie, J. (2017a). Comprehensive predictions of tourists' next visit location based on call detail records using machine learning and deep learning methods. In *Proceeding - 2017 IEEE International Congress on Big Data (BigData Congress)* (pp. 1–6). IEEE. <https://doi.org/10.1109/BigDataCongress.2017.10>
- Chen, N. C., Xie, J., Tinn, P., Alonso, L., Nagakura, T., & Larson, K. (2017b). Data mining tourism patterns: call detail records as complementary tools for urban decision making. In P. Janssen, P. Loh, A. Raonic, & M. A. Schnabel (Eds.), *Proceedings of the 22nd International Conference on Computer-Aided Architectural Design Research in Asia (CAADRIA 2017)* (pp. 685–694). CAADRIA.
- Colther, C., & Arriagada-Millaman, A. (2021). Pronóstico de la demanda turística de Chile basado en modelos lineales y no lineales estacionales. *PASOS: Revista de Turismo y Patrimonio Cultural*, 19(2), 323–336. <https://doi.org/10.25145/j.pasos.2021.19.021>
- Condeço-Melhorado, A., García-Palomares, J. C., & Gutiérrez, J. (2024). The uneven impact of the COVID-19 pandemic on domestic tourist flows: What does mobile phone data tell us? *Journal of Hospitality and Tourism Technology*, 15(3), 345–362. <https://doi.org/10.1108/JHTT-04-2023-0103>

- De Luca, G., & Rosciano, M. (2024). Google Trends data and transfer function models to predict tourism demand in Italy. *Journal of Tourism Futures*. <https://doi.org/10.1108/JTF-01-2023-0018>
- de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3(1), Article 1376. <https://doi.org/10.1038/srep01376>
- European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation)*. <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>
- Grassini, L., & Dugheri, G. (2021). Mobile phone data and tourism statistics: A broken promise? *National Accounting Review*, 3(1), 50-68. <https://doi.org/10.3934/NAR.2021002>
- Hall, C. M. (2005). Reconsidering the geography of tourism and contemporary mobility. *Geographical Research*, 43(2), 125-139. <https://doi.org/10.1111/j.1745-5871.2005.00308.x>
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1), 5-10. <https://doi.org/10.1016/j.ijforecast.2003.09.015>
- Kovacs, Z., Vida, G., Elekes, A., & Kovalcsik, T. (2021). Combining social media and mobile positioning data in the analysis of tourist flows: A case study from Szeged, Hungary. *Sustainability*, 13(5), Article 2926. <https://doi.org/10.3390/su13052926>
- Li, X., Law, R., Xie, G., & Wang, S. (2021). Review of tourism forecasting research with internet data. *Tourism Management*, 83, Article 104245. <https://doi.org/10.1016/j.tourman.2020.104245>
- Mamei, M., & Colonna, M. (2018). Analysis of tourist classification from cellular network data. *Journal of Location Based Services*, 12(1), 19-39. <https://doi.org/10.1080/17489725.2018.1463466>
- Mondaca-Marino, C. M., Arriagada Millaman, A., & Piffaut, P. V. (2021). Forecasting tourism demand in Chile: Regional analysis using the seasonal autoregressive model. *El Periplo Sustentable*, (41), 234-254.
- Muangprathub, J., Kaewmanee, P., Sealee, J., Warintarawej, P., & Sae-jie, W., & (2025). Forecasting trends in foreign tourism by machine learning. *Tourism: An International Interdisciplinary Journal*, 73(3), 392-409. <https://doi.org/10.37741/t.73.3.1>
- Organisation for Economic Co-operation and Development. (2024). *OECD tourism trends and policies 2024*. <https://doi.org/10.1787/6b47b985-en>
- Okmi, M., Por, L., Ang, T., & Ku, C. (2023). Mobile phone data: A survey of techniques, features, and applications. *Sensors*, 23(2), Article 908. <https://doi.org/10.3390/s23020908>
- Olteanu, A. M., Trasarti, R., Couronné, T., Giannotti, F., Nanni, M., Smoreda, Z., & Ziemlicki, C. (2011, October). GSM data analysis for tourism application. In *Proceedings of the 7th International Symposium on Spatial Data Quality (ISSDQ)* (pp. 1-6).
- Park, S., Zu, J., Xu, Y., Zhang, F., Liu, Y., & Li, J. (2023). Analyzing travel mobility patterns in city destinations: Implications for destination design. *Tourism Management*, 96, Article 104718. <https://doi.org/10.1016/j.tourman.2022.104718>
- Raun, J., Shoval, N., & Tiru, M. (2020). Gateways for intra-national tourism flows: Measured using two types of tracking technologies. *International Journal Of Tourism Cities*, 6(2), 261-278. <https://doi.org/10.1108/IJTC-08-2019-0123>
- Reif, J., & Schmücker, D. (2020). Exploring new ways of visitor tracking using big data sources: Opportunities and limits of passive mobile data for tourism. *Journal of Destination Marketing and Management*, 18, Article 100481. <https://doi.org/10.1016/j.jdmm.2020.100481>
- Saluveer, E., Raun, J., Tiru, M., Altin, L., Kroon, J., Snitsarenko, T., Aasa, A. & Silm, S. (2020). Methodological framework for producing national tourism statistics from mobile positioning data. *Annals of Tourism Research*, 81, Article 102895. <https://doi.org/10.1016/j.annals.2020.102895>
- Servicio Nacional de Turismo. (2024a). *Servicio Nacional de Turismo*. <https://www.sernatur.cl>
- Servicio Nacional de Turismo. (2024b). *Reporte SNASPE*. <https://datosturismo.sernatur.cl/siet/reportesDinamicoSNASPE>

- Statista. (2024). *Travel & tourism – Chile*. <https://www.statista.com/outlook/mmo/travel-tourism/chile>
- Subsecretaría de Turismo de Chile. (2018). *Turismo interno de Chile* [Infographic]. <https://www.subturismo.gob.cl/wp-content/uploads/2015/10/Infografia-Turismo-Interno-de-Chile.pdf>
- Subsecretaría de Turismo de Chile. (2024). *Subsecretaría de Turismo*. <http://www.subturismo.gob.cl>
- Tang, C. F., Qi, Z., & Trinugroho, I. (2025). Tourism-led growth in Shandong, China. *Tourism: An International Interdisciplinary Journal*, 73(2), 227–239. <https://doi.org/10.37741/t.73.2.3>
- United Nations World Tourism Organization. (2008). International recommendations for tourism statistics. <https://www.unwto.org/standards/on-basic-tourism-statistics-irts-2008>
- Vanhoof, M., Hendrickx, L., Puussaar, A., Verstraeten, G., Ploetz, T., & Smoreda, Z. (2017). Exploring the use of mobile phone data for domestic tourism trip analysis. *Netcom*, 31,(3/4), 335–372. <https://doi.org/10.4000/netcom.2742>
- Xu, Y., Xue, J., Park, S., & Yue, Y. (2021). Towards a multidimensional view of tourist mobility patterns in cities: A mobile phone data perspective. *Computers, Environment and Urban Systems*, 86, Article 101593. <https://doi.org/10.1016/j.compenvurbsys.2020.101593>
- Zang, H., & Bolot, J. (2011). Anonymization of location data does not work. In *Proceedings of the 17th annual international conference on Mobile computing and networking* (pp. 145–156). Association for Computing Machinery. <https://doi.org/10.1145/2030613.2030630>
- Zaragozí, B., Trilles, S., & Gutiérrez, A. (2021). Passive mobile data for studying seasonal tourism mobilities: An application in a mediterranean coastal destination. *ISPRS International Journal of Geo-Information*, 10(2), Article 98. <https://doi.org/10.3390/ijgi10020098>

Submitted: August 14, 2024

Revised: November 17, 2025

Accepted: January 13, 2026