High-frequency forecasting from mobile devices' bigdata: an application to tourism destinations' crowdedness

Vicente Ramos
*Department of Applied Economics, University of the Balearic Islands,
Palma de Mallorca, Spain*

Woraphon Yamaka
*Faculty of Economics, Center of Excellence in Econometrics,
Chiang Mai University, Chiang Mai, Thailand*

Bartomeu Alorda
*Department of Physics, University of the Balearic Islands,
Palma de Mallorca, Spain, and*

Songsak Sriboonchitta
*Faculty of Economics, Center of Excellence in Econometrics,
Chiang Mai University, Chiang Mai, Thailand*

Abstract

Purpose – This paper aims to illustrate the potential of high-frequency data for tourism and hospitality analysis, through two research objectives: First, this study describes and test a novel high-frequency forecasting methodology applied on big data characterized by fine-grained time and spatial resolution; Second, this paper elaborates on those estimates' usefulness for visitors and tourism public and private stakeholders, whose decisions are increasingly focusing on short-time horizons.

Design/methodology/approach – This study uses the technical communications between mobile devices and WiFi networks to build a high frequency and precise geolocation of big data. The empirical section compares the forecasting accuracy of several artificial intelligence and time series models.

Findings – The results robustly indicate the long short-term memory networks model superiority, both for in-sample and out-of-sample forecasting. Hence, the proposed methodology provides estimates which are remarkably better than making short-time decision considering the current number of residents and visitors (Naïve I model).

Practical implications – A discussion section exemplifies how high-frequency forecasts can be incorporated into tourism information and management tools to improve visitors' experience and tourism stakeholders' decision-making. Particularly, the paper details its applicability to managing overtourism and Covid-19 mitigating measures.

Originality/value – High-frequency forecast is new in tourism studies and the discussion sheds light on the relevance of this time horizon for dealing with some current tourism challenges. For many tourism-related issues, what to do next is not anymore what to do tomorrow or the next week.

Plain Language Summary – This research initiates high-frequency forecasting in tourism and hospitality studies. Additionally, we detail several examples of how anticipating urban crowdedness requires high-frequency data and can improve visitors' experience and public and private decision-making.

Keywords Big data, Artificial intelligence, Overtourism, COVID-19, Crowdedness, High-frequency forecast

Paper type Research paper

1. Introduction

The expansion of information and communication technologies (ICT) has revolutionized the way in which we produce, consume and enjoy leisure; Tourism activities are no exception to this general trend. As a parallel consequence, the ubiquitous use of technology generates new types of data that were unimaginable only some years ago. Many of these information sources are labeled as big data, characterized by huge amounts of observations and usually higher frequencies (Li *et al.*, 2018). Mariani *et al.* (2018) present a comprehensive review of big data in tourism, identifying relevant contributions and areas for future research.

Many companies are currently storing increasing volumes of data, hoping that it can be useful in the future. However, the characteristics of these data (volume, frequency, etc.) imply that sometimes, they are not fully exploited to generate applicable knowledge. In line with Scott *et al.*'s (2017) suggestions, there is a need for more fluent communication between academia and public and private stakeholders to take full advantage of these relatively new sources of information.

In this context, the main aim of this paper is to illustrate the potential of high frequency (HF) data for tourism and hospitality. This general aim is divided into two related research objectives: First, we describe and test the first HF forecasting methodology applied to tourism-related big data; Second, we elaborate on how these forecasts can be used to extract relevant information for visitors and tourism destination's public and private stakeholders.

The availability of tourism and even general, HF data is still in its infancy. Hence, it is not surprising that there is a scarcity of theoretical and empirical academic research considering this time horizon. More research is needed to understand and model these data challenges and to promote its application to inform tourism decisions and policies. Particularly as many technological innovations (5G networks, the internet of things, travel apps, etc.) are related to real-time or very HF data. This is clearly the case in tourism and hospitality, as decisions are increasingly taken during the visit (Gretzel *et al.*, 2006; Neuhofer *et al.*, 2012; Wang *et al.*, 2012). In this sense, there is an emerging area of research based on HF monitoring (Hardy *et al.*, 2017; Huang *et al.*, 2020; Zheng *et al.*, 2017) which illustrates how this type of data can be useful for tourism and hospitality analysis. However, the logical research development of going from showing what is happening now (monitoring), to anticipate the future (forecasting) constitutes a research gap. As far as the authors are aware, only Xia *et al.* (2009, 2011) and Zheng *et al.* (2017) used a similar conceptual approach, as they attempted to anticipate the next location visited by tourists, with a short-time horizon.

The big data considered in this paper is also original, as it uses the technical communications between mobile devices that move around an urban destination (Palma de

Mallorca, Balearic Islands, Spain) and the WiFi networks deployed through the city. Hence, it falls under the categories of: "device data" following Li *et al.* (2018); and passive positioning data as classified in Shoval and Ahas (2016).

Section 4 presents a detailed description of the data. However, several initial remarks illustrate its significant potential, novel characteristics and replicability to other destinations. First, many destinations, particularly cities, offer free WiFi through a public or a contracted network. Hence, the empirical analysis and managerial applications explained in the current paper are easily replicable, at an affordable cost. Second, the network-device interactions are characterized by precise spatiotemporal information. The communications are done with HF (in general, at least once per minute) and contain rich geolocation information. In other words, the database contains pseudo-anonymized precise information of the presence and movement of people through the covered areas.

Regarding the first research objective, over the past few decades, many approaches (Song *et al.*, 2019) have been used to analyze, characterize and forecast tourism demand. In this paper, we compare the forecasting accuracy of several advanced methods, which might be able to capture the complex statistical characteristics of big data. In particular, we consider four artificial intelligence methods (support vector machine, SVN; artificial neural network, ANN; recurrent neural network, RNN; and long short-term memory networks, LSTM) and two-time series methodologies (state-space ARIMA, SSARIMA; and multiplicative SSARIMA, MSSARIMA).

Considering the second research objective, Section 6 elaborates on how anticipating high-occupancy episodes at precise destinations' locations has a myriad of applications for tourism and hospitality, from visitor's on-site decisions to advanced crowdedness' management systems. This last topic is particularly relevant to deal with two challenges recently faced by many tourism destinations: overtourism and social distancing measures associated with Covid-19. Regarding overtourism, even if the current pandemic crisis has reduced its importance in 2020 and maybe for some additional years, probably many popular destinations will face similar problems in the near future. Overtourism cannot be studied as a general phenomenon happening constantly at a tourism destination. In fact, many of the nuances related to it are associated with overcrowding episodes at specific locations and moments in time. This consideration also applies to social distancing measures related to Covid-19 mitigation. In these two cases and in any other application, crowdedness' analysis should include precise time and location information. In other words, they should be studied with frequencies higher than one day or longer horizons.

The rest of the paper is structured as follows: Section 2 presents the literature review. Section 3 provides a basic overview of the methods used in this study. Those readers interested in modeling can find further details in the Appendix. Section 4 details the origin of big data and its main characteristics. Section 5 describes the in-sample and out-of-sample accuracy measures. Section 6 explains several tourism applications of HF forecasting and highlights the paper's practical and theoretical implications. Finally, Section 7 summarizes the conclusions and research limitations.

2. Literature review
This section starts revising briefly some advances in generic tourism demand. A comprehensive literature review can be found at Song *et al.* (2019). Afterward, we concentrate on the specific issue of HF forecasting.

This paper uses time series and artificial intelligence models; Together with econometrics, those are the main three families of quantitative tourism demand forecasting (Song and Li, 2008). Time-series econometric has been extensively used in the literature

(Chu, 1998; Kulendran and Witt, 2003; Loganathan and Ibrahim, 2010; Thushara *et al.*, 2019). In this vast literature, the most widely used methods are the autoregressive (AR) and moving average (MA) (Geurts and Ibrahim, 1975; Witt and Witt, 1991), the AR integrated MA (ARIMA) (Lim and McAleer, 2002) and the seasonal ARIMA (Lim and McAleer, 2000). Recently, SSARIMA (Hyndman *et al.*, 2008) and MSSARIMA (Svetunkov and Boylan, 2020) were proposed as advanced forecasting alternatives.

Some authors indicated that time series models may yield unsatisfactory forecasting results, particularly when non-linearity and noise exist in the data. In fact, machine learning (ML) approaches are attracting increasing interest for predicting tourism demand. Methodologies such as ANN and SVM are often recommended in the literature. Pouyanfar *et al.* (2018) suggested that these techniques were more appropriate to capture non-linear relationships. Bedi and Toshniwal (2019) claim that these models outperform various forecasting models commonly used so far. Chen and Wang (2007) compare SVM, backpropagation neural network (BPNN) and ARIMA to forecast tourists' arrivals to China. They conclude that SVM outperforms the neural networks and ARIMA models in terms of normalized mean square error (NMSE) and mean absolute percentage error (MAPE). Aslanargun *et al.* (2007) also compared ARIMA with ANN and suggested a higher performance of ANN. Li and Cao (2018) introduced RNN and LSTM for forecasting tourism flows. They found that LSTM methods perform better than ARIMA and BPNN.

Hence, many previous studies found that ML provides good forecasting performance. However, these methodologies also have some limitations. First, they require large data for algorithm training. Additionally, they are time-consuming and require advanced computational resources to produce reliable forecasts. Furthermore, some previous papers (Claveria and Torra, 2014) indicated that ANN does not outperform ARIMA, especially for short-time horizons. A final source of criticism (Xu *et al.*, 2016) is that although ML is computationally accurate and exhibits satisfactory forecasting performance, it is basically a "black-box" that lacks explanatory capacity. In this sense, they cannot be used to provide causal interpretations.

The above paragraphs show that time series and artificial intelligence have been extensively used. However, previous studies focused basically on low-frequency data (yearly, quarterly, monthly and exceptionally, daily). Two main reasons may explain the dominance of these frequencies: the lack of data at higher frequencies and the need to answer relevant research questions within the lower frequencies' domain. Nonetheless, as it has been described in the introduction, the ICT generalization produces new databases characterized by much higher frequencies (close to real-time data) which provide innovative knowledge for tourism and hospitality.

HF prediction is new in tourism research. However, it has been increasingly used in other research fields as transport, financial markets or environmental studies. Kumar *et al.* (2013) applied ANN to forecast short-term traffic flows using 5 and 15 min data. Aqib *et al.* (2019) used deep learning methods to forecast traffic data on freeways with a 5 min frequency. They conclude that HF data produces near-real-time forecasts which are useful for dynamic road traffic management. Xing *et al.* (2019) developed asymmetric extreme learning machine cluster model to predict traffic congestion with 10 min intervals. In the financial field, Degiannakis and Filis (2018) supported HF forecasting, emphasizing that financial markets' intraday information should not be ignored. In this line, Lachiheb and Gouider (2018) successfully found that deep learning neural networks could forecast 5 min returns at the Tunisian stock markets. Chen *et al.* (2020) and Shintate and Pichl (2019) considered 5 min frequencies for studying cryptocurrencies. Both papers indicate that ML algorithms have higher efficiency than classical statistical methods. Finally, HF data were also applied by

Khosravi *et al.* (2018) in environmental research. They proposed three models of ML algorithms (multilayer feed-forward neural network, support vector regression and adaptive neuro-fuzzy inference system) to predict wind speed, wind direction and output power of a wind turbine measured at HF intervals. They found that support vector regression provided the best forecasts.

3. Methodology
This paper applies several methodologies and compares its forecasting accuracy. The choice of models is based on the data's characteristics. Despite the popularity of classical ARIMA-type approaches, these models display a poor performance when forecasting large, complex and nonlinear data (Rundo *et al.*, 2019). Thus, more advanced time series models such as state-space-type models (SARIMA and MSARIMA) and deep learning models (ANN, RNN and LSTM) are more appropriate to forecast HF data. These models have similar advantages as they account for nonlinear dynamic behaviors. In fact, state-space (Dordonnat *et al.*, 2008; Elghafghuf *et al.*, 2018) and deep learning (Kuo and Huang, 2018; Paoli *et al.*, 2011) have been used before for HF forecasting. This section presents the basic methodologies' characteristics, while further mathematical and technical details can be found in the Appendix.

3.1 Support vector machine
The support vector machine (SVM) is a popular and robust artificial intelligence method, based on learning algorithms, which has been widely used (Cortes and Vapnik, 1995; Smola and Scholkopf, 2004). This is a flexible forecasting technique that provides accurate tourism forecasts (Chen and Wang, 2007). Its main advantages are that structural complexity is controlled in the optimization and it uses convex quadratic programming that leads to a globally optimal solution.

3.2 Artificial neural network
ANN is a deep learning method that captures various forms of non-linear relationships. There are several types of ANN, but the most popular is the multilayer perceptron. The model is based on self-learning and self-adapting protocols, which make it suitable for finding the relationship between a set of inputs and an output (Zhang *et al.*, 1998). It consists of three components (a multi-layer, weights and neurons) that connect the inputs and the output. In terms of the estimation procedure, backpropagation is usually used to find the optimal weights. ANN is an iterative and recursive method to find the weights that minimize the loss function.

Researchers need to define the number of layers in the ANN's structure. Following previous literature (Seyyedsalehi and Seyyedsalehi, 2015; Yin *et al.*, 2017) we considered an ANN's structure with five-layers; each contains j neurons connected among them and with all the neurons in its neighbor layers.

3.3 Recurrent neural network
RNN is an ANN's extension introduced to deal with historical data dependencies. This network records the information from the past and uses the current output to predict the next output. RNN has repetitive loops that combine previous and new input information to predict current and future outputs (Mandic and Chambers, 2001).

Bandara *et al.* (2017) mentioned that although RNN is successful in solving short-term dependencies, they cannot handle long-term dependencies. This limitation is due to what is called the "vanishing gradient problem": as time advances, the gradient becomes smaller. As

a consequence, it is more difficult to train the algorithm in the presence of long-term effects. LSTM was suggested to solve this limitation.

3.4 Long short-term memory network

This deep learning model is an extension of RNN that produces accurate forecasts (Li and Cao, 2018; Zhang *et al.*, 2019). LSTM effectively solves vanishing gradients by using memory cells (Hochreiter and Schmidhuber, 1997). This model can process both single data points and entire sequences of data. The architecture of LSTM consists of input, forgotten and output gates.

3.5 State-space autoregressive integrated moving average and multiplicative state-space autoregressive integrated moving average

The *SSARIMA* was introduced by Harvey and Phillips (1979). Afterward, Hyndman *et al.* (2008) extended the model and suggested the recursive estimation for obtaining the model's parameters. By using the state space approach, it is possible to find the appropriate ARIMA's order without hypotheses testing. Instead, this approach performs model selection based on the information criteria (Svetunkov and Boylan, 2020). In its application to big data analysis selecting ARIMA's order is quite slow and troublesome. Thus, SSARIMA reduces the computation time in the estimation process.

The main advantages of SSARIMA are: First, it generates predictions on observation $t = 0$, which increases the estimation's degrees of freedom (Hyndman *et al.*, 2008). It should be acknowledged that this general advantage is not relevant in its application to big data, as there are many observations. Second, it is based on time-varying parameters, so it provides additional flexibility. Third, SSARIMA is estimated by maximum likelihood, thus it is easy to apply information criteria for selecting the best model specification without hypotheses testing. It should be noted that a 5 min' frequency involves producing forecasts for a large data set. Hence, the automatized order selection, instead of the manual analysis of conventional ARIMA, is relevant for big data applications (Ramos *et al.*, 2015).

Recently, Svetunkov and Boylan (2020), indicated that SSARIMA involves multiple estimation steps, which implies a high computational cost. Thus, they proposed MSSARIMA, which decreases the transition matrix dimension by skipping zero polynomials. As a result, its estimation is remarkably faster than SSARIMA on HF data. That implies a relevant decrease in computational costs, which deserves consideration in this type of analyzes.

3.6 Forecasting accuracy

Conventionally, the forecasting accuracy is evaluated by analyzing two loss functions, namely, root mean squares error (RMSE) and mean absolute error (MAE). Sometimes, these two loss functions provide different ranking complicating model selection (Ma *et al.*, 2019). Consequently, this paper also conducted the model confidence set (MCS) of Hansen *et al.* (2011) to analyze the robustness of the forecasting results. This procedure consists of a sequence of statistical tests that identifies the "superior set models" (SSM). Specifically, MCS evaluates if the forecasting performance of a candidate model is significantly different from a reference model. If the performance is not significantly different, it will be included in the SSM.

In this study, RMSE and MAE are used as loss functions for the MCS tests, which are implemented through a 1,000 replications' bootstrap. The performance of all candidate models is evaluated and if the baseline one is not rejected, a new candidate is then compared. The model with the highest p-value is selected as the best forecasting option.

4. Big data source: origin, characteristics and statistical description
4.1 Origin of the data
Data is elaborated from the communications between the public WiFi network of an urban tourism destination, Palma de Mallorca and the mobile devices in its coverage area. The network, Palma's SmartWiFi (P-SW), provides free internet services in most city areas. All the system is managed by a single firm, WIONGO.

Palma is the capital of Mallorca Island and its international airport's location. With more than 11.8 million visitors in 2019 (86% international) Mallorca is one of the main tourism destinations in the Mediterranean. Figure 1 displays the location of the island (panel a) and the city (panel b). Palma has around 50,000 tourism beds (15 of the Island's total), so it is a destination itself and it is also a frequent one-day excursion for those visitors staying in different areas. Palma is an appropriate destination to prove the usefulness of HF crowdedness forecasts as it was among the first territories in which overtourism arose as a social debate labeled as tourismophobia (Milano *et al.*, 2019). Regarding the impact of Covid-19, its insularity implies high air connectivity dependence; hence it has been severely impacted by the pandemic. Like any other destination, the short-run tourism evolution will
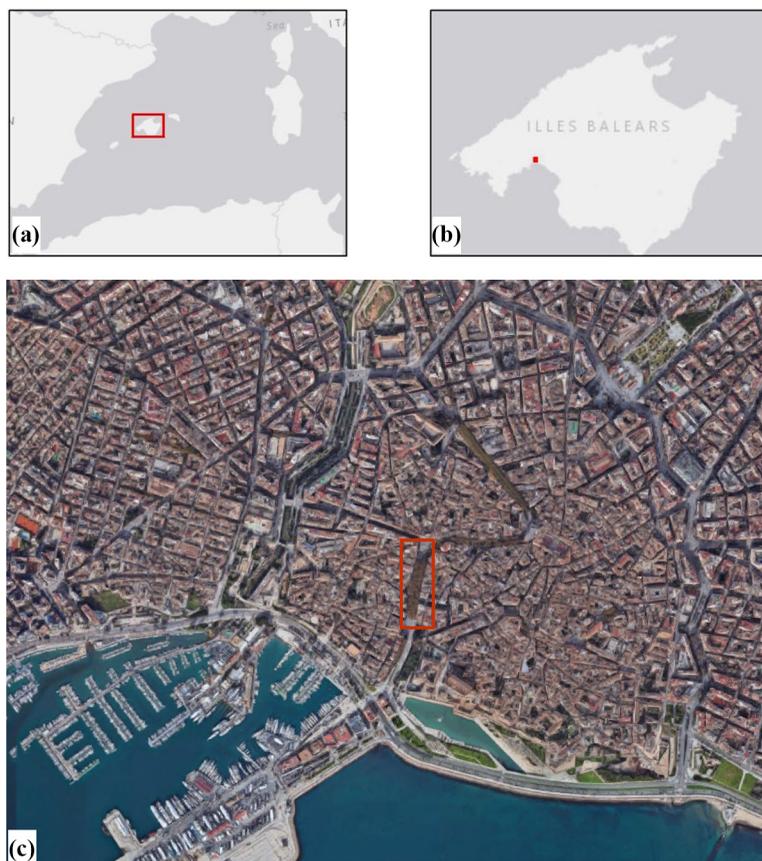
be very dependent on its safe destination image and social distancing is among the top Covid-19 safety protocols.

In terms of data generation, any mobile device with its WiFi adapter enabled is constantly screening the available networks. The interaction between the device and a network generates a record of request start and finish timestamp (with EPOCH coding by the second), unique device's Media Access Control (MAC) and precise location. Alessandrini *et al.* (2017) concluded that WiFi data provides high spatial resolution, appropriate for urban mobility analysis. Regarding privacy protection and ethical concerns, the technical data does only report the number of devices in different locations. Hence, it does not include any type of personal information. In this sense, this data follows the European Regulation (2016/ 679; EU, 2016) that indicates: "The principles of data protection should, therefore, not apply to anonymous information, namely, information which does not relate to an identified or identifiable natural person." The MAC addresses are not classified as a personal information because they only identify a particular WIFI transceiver unrelated to the user's identity.

WiFi data has not been applied before in tourism research. The information can be classified as passive position data (Shoval and Ahas, 2016). As such, it presents some similarities with that provided by private mobile operators which have been successfully used in tourism research (Ahas *et al.*, 2007, 2008; Raun *et al.*, 2016). However, public WiFi data presents some distinctive features that might be appealing for tourism destinations such as lower cost; all information is collected with a homogeneous style and the geolocation data is very precise (note that usually private telecommunications companies provide broad locations).

Each data set might be more appropriate for some research objectives. Data from mobile operators or internet-based applications (Apple, Google, Baidu, etc.) provide useful insight for analyzing movements through a wider geographic context. WiFi network data has advantages for specific locations such a: city area, tourism attraction, beach area and boardwalk.

4.2 Big data characteristics, zoning and devices' characterization

This paper uses data collected during 2019's high tourism season (July to September). Mallorca has a remarkable seasonality; those three months account for 47% of annual arrivals (INE, 2020). Hence, that is the appropriate period for considering issues related to the destination's crowdedness. The database gathers a daily average of 3.7 million observations, from more than 214,000 unique devices.

Several decisions have been made to limit the research: First, we aggregated the real-time data into 5 min intervals, which captures the dynamic city use. Second, the paper presents the analysis of only one specific location. To do the zoning a homogeneous grid was drawn on the city's map. Taking into account its tourism relevance and the optimal WiFi coverage, we choose to present the analysis for the area corresponding to Paseo del Borne, indicated in Figure 1 (panel c). This is the city's heart and a usual pedestrian route with numerous shops, restaurants, bars and historical buildings. Third, the forecast exercise is based on the observations between 8 a.m. and 8 p.m. The intraday pattern presented in Figure 2 indicates that pedestrians concentrate during that period.

Additionally, the observations were classified as city residents or visitors. Hence, we can model their different urban space use. Any crowdedness' analysis should not focus only on tourists or residents, as both groups share the city. However, different policies and tools might be appropriate for each group. The classification of observations as residents or visitors is based on its presence over the three months' period. A recursive query has been programed to label as visitors those observations which are only captured in a period equal or inferior to 10 days. The other observations are considered residents.

15 July 2019
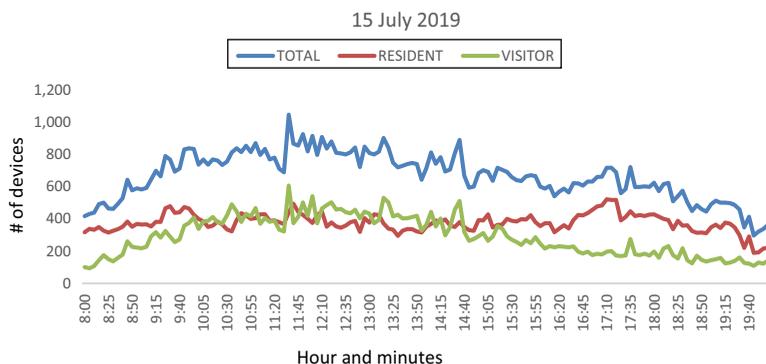
Figure 2.
Intraday behavior
(15 of July)

With all the above considerations, the final sample consists of 13,248 observations counting the number of residents and visitors in the selected location. Figure 2 plots the intraday (15th of July) pattern of the total number of devices and its disaggregation as residents or visitors. The graph illustrates the different behavior of the two groups: Residents have a more constant presence throughout the day; while visitors tend to start gathering later (after 10 a.m.), they outnumber the residents during the central part of the day and their presence is slightly decreasing after 3 p.m. We found a relatively stable pattern so that there is intraday time dependence. Finally, it is interesting that the city use peaks are usually determined by a visitors' increase. This finding supports the use of specifically managing visitors' flows to avoid overcrowding episodes in a given time and space.

Table 1 presents the main descriptive statistics; the mean is 319.5 for residents and 209.2 for visitors. Hence, there is a relevant proportion (65%) of visitors. That is an expected result as the area corresponds to the commercial streets of an urban tourism destination during its high season. Given data's HF, the standard deviations are quite high, 87.3 for residents and 72.3 for visitors. Both series exhibit high kurtosis (greater than 3), therefore these data's distribution tails are heavier than in a normal distribution. The last row of Table 1 displays the augmented Dickey–Fuller's unit root test conducted to evaluate data stationarity. The results reject the null hypothesis of a unit root at a 1% significance level, implying that the series is stationary and can be modeled without further transformations.

Statistics	Residents	Visitors
Mean	319.50	209.22
Median	324	1,962
Maximum	1,205	7,602
Minimum	46	132
Std. dev.	72.30	87.30
Skewness	0.28	0.96
Kurtosis	8.448	4.69
Observations	13,248	13,248
ADF	−16.54***	−13.99***

Note: *** indicates 1% significance level

Table 1.
Data description

5. Models' forecasting accuracy

The next two subsections present forecasting accuracy based on the two-loss functions and MCS. The results are presented for both, in-sample and out-of-sample exercises. The conventional former analysis (Section 5.1) can be used to compare our results with previous literature. The out-of-sample analysis (Section 5.2) has been specifically designing to illustrate how the current approach provides HF forecasts useful to address tourism destination's challenges related to spatiotemporal crowdedness.

5.1 In-sample statistical performance

This first empirical analysis evaluates the in-sample forecasts to compare the overall performance of the six proposed models. Hence, this section uses MAE, RMSE and MCS to evaluate each model's ability to represent the real data. The in-sample statistical performance for the two groups (residents and visitors) are provided in Table 2. In terms of interpretation, a lower RMSE and MAE and a greater, p-value imply that the model provides more accurate forecasts.

Table 2 results' consistently indicate that in the current database, LSTM has an outstanding performance for forecasting in-sample residents and visitors. In other words, it is undoubtedly statistically superior to all other models. The loss function measures (MAE and RMSE) present lower values for LSTM. Additionally, the MCS rejects all other methodologies. The big data complexity hinders providing a strong explanation for the above results. However, LSTM's capacity to adapt to changing situations while preserving its long-term memory seems to provide appropriate characteristics to this model.

5.2 Out-of-sample forecasting accuracy

This paper proposes to use the above methodology to provide a recursive HF estimation. Hence, the out-of-sample analysis can be done for any time span, provided that there was sufficient data for model specification and algorithm training. Out of the 13,248 observations, Table 2 indicates the three specific periods of 1 h that have been used to present the models' forecasting accuracy. The selection criterion has been to choose: one period per month, in a moment of high urban crowdedness, but that was not an outlier.

Model	MAE	MCS	RMSE	MCS
Residents				
SSARIMA	27.2096	0.0	36.9795	0.0
MSARIMA	25.6556	0.0	35.1861	0.0
ANN	26.1371	0.0	35.7731	0.0
LSTM	*7.1775*	*1.0*	*11.9080*	*1.0*
RNN	26.4149	0.0	36.5010	0.0
SVM	26.0751	0.0	35.7283	0.0
Visitors				
SSARIMA	26.8257	0.0	41.8671	0.0
MSSARIMA	25.3089	0.0	38.9997	0.0
ANN	25.8263	0.0	39.9966	0.0
LSTM	*6.9909*	*1.0*	*12.6752*	*1.0*
RNN	26.5301	0.0	40.5582	0.0
SVM	26.2127	0.0	40.3041	0.0

Note: The italic numbers indicate the lowest error rate (MAE and RMSE)

Table 2.
In-sample forecasting accuracy

Hence, these three episodes correspond to a situation when the number of devices is around its statistical maximum (the third quartile, plus half of the interquartile difference, Q3-Q1).

The models' parameters are optimized on a training set, indicated in the second column of Table 3; afterward, the forecasting period (third column) is used to test each model's forecast against reality.

A detailed analysis of the models' accuracy is presented in Table 4 for residents and Table 5 for visitors. The two tables have a similar structure; they present the MAE, RMSE and MCS for each model and period. Note that we also included a basic Naïve 1 (No change) model, as it is frequently used as a benchmark (Athanasopoulos *et al.*, 2011). Moreover, it indicates what information would have been used to make future decisions based on monitoring, instead of HF forecasting.

Additionally, the results are presented for two forecast horizons: $t + 6$, which corresponds to a half-an-hour; and $t + 12$, which is a 1-h forecast. A rolling estimation

Period of study	Training period	Forecasting period
Period 1	1 July 2019, 8.00 am to 15 July 2019, 10.45 am	15 July 2019, 10.50 am.–11.50 am
Period 2	1 August 2019, 8.00 am to 27 August 2019, 9.40 am	27 August 2019, 9.45 am.–10.45 am
Period 3	1 September 2019, 8.00 am to 21 September 2019, 10.40 am	21 September 2019, 10.45 am–11.45 am

Table 3.
Training and
forecasting periods

Model	6-steps ahead				12-steps ahead			
	MAE	MCS	RMSE	MCS	MAE	MCS	RMSE	MCS
Period 1								
SSARIMA	36.6931	0.0	39.1526	0.0	46.9385	0.0	50.1991	0.0
MSARIMA	40.9271	0.0	43.1191	0.0	50.8621	0.0	53.3158	0.0
ANN	34.4374	0.0	38.1727	0.0	45.5748	0.0	49.0298	0.0
LSTM	*25.1175*	*1.0*	*30.6745*	*1.0*	*29.0328*	*1.0*	*33.7029*	*1.0*
RNN	34.9685	0.0	38.9538	0.0	45.6026	0.0	49.9126	0.0
SVM	38.5527	0.0	42.5519	0.0	49.4569	0.0	53.1682	0.0
Naïve 1	42.7589	0.0	45.3650	0.0	50.8461	0.0	54.0716	0.0
Period 2								
SSARIMA	16.5303	0.31	18.2601	0.25	17.0349	0.0	18.8242	0.0
MSARIMA	19.5120	0.0	20.5429	0.0	19.4928	0.0	22.7098	0.0
ANN	27.0616	0.0	28.9998	0.0	29.6235	0.0	32.6826	0.0
LSTM	*16.4383*	*1.0*	*19.6039*	*1.0*	*13.8928*	*1.0*	*15.3219*	*1.0*
RNN	32.2864	0.0	34.5893	0.0	29.7150	0.0	32.7571	0.0
SVM	22.8643	0.0	24.9171	0.0	22.9737	0.0	24.9730	0.0
Naïve 1	32.2121	0.0	33.1662	0.0	30.9231	0.0	34.8500	0.0
Period 3								
SSARIMA	26.7798	0.0	31.6663	0.0	21.8793	0.0	26.0592	0.0
MSARIMA	24.1895	0.0	28.2725	0.0	22.7425	0.0	26.3276	0.0
ANN	28.4134	0.0	33.8211	0.0	25.1311	0.0	28.1644	0.0
LSTM	*22.7517*	*1.0*	*24.9344*	*1.0*	*13.4012*	*1.0*	*14.9947*	*1.0*
RNN	23.6151	0.25	25.3158	0.31	18.4009	0.0	20.9979	0.0
SVM	30.1544	0.0	35.1258	0.0	27.1183	0.0	29.9809	0.0
Naïve 1	31.7545	0.0	36.1185	0.0	30.3846	0.0	32.5221	0.0

Note: The italic numbers indicate the lowest error rate (MAE and RMSE)

Table 4.
Out-of-sample
forecasting accuracy
for residents

Model	6-step ahead				12-step ahead			
	MAE	MCS	RMSE	MCS	MAE	MCS	RMSE	MCS
Period 1								
SSARIMA	41.4288	0.0	44.2381	0.0	42.5185	0.0	51.8779	0.0
MSARIMA	47.3881	0.0	49.9024	0.0	46.6675	0.0	57.1757	0.0
ANN	46.1676	0.0	49.2756	0.0	54.8511	0.0	60.9110	0.0
LSTM	*28.9648*	*1.0*	*30.4676*	*1.0*	*32.1464*	*1.0*	*45.2101*	*1.0*
RNN	45.1481	0.0	46.4471	0.0	46.1706	0.0	56.7902	0.0
SVM	48.8050	0.0	55.8872	0.0	58.6210	0.0	64.5037	0.0
Naïve 1	50.3311	0.0	53.8176	0.0	58.6815	0.0	64.8541	0.0
Period 2								
SSARIMA	35.8164	0.0	39.5107	0.0	31.0932	0.0	35.9545	0.0
MSARIMA	21.1724	0.0	26.0561	0.0	27.9434	0.0	30.3806	0.0
ANN	42.0653	0.0	43.9968	0.0	44.2593	0.0	46.8779	0.0
LSTM	*20.0821*	*1.0*	*22.7662*	*1.0*	*19.4792*	*1.0*	*20.9253*	*1.0*
RNN	40.4964	0.0	43.9783	0.0	42.0421	0.0	45.4903	0.0
SVM	28.3697	0.0	29.9665	0.0	29.5785	0.0	34.9865	0.0
Naïve 1	42.3125	0.0	44.1540	0.0	48.9230	0.0	51.3325	0.0
Period 3								
SSARIMA	31.9379	0.0	38.7095	0.0	32.0336	0.0	37.1365	0.0
MSARIMA	32.9646	0.0	39.3103	0.0	32.1794	0.0	39.8404	0.0
ANN	38.0234	0.0	44.4781	0.0	39.5558	0.0	45.8675	0.0
LSTM	*28.7461*	*1.0*	*30.0618*	*1.0*	*26.1132*	*1.0*	*27.5015*	*1.0*
RNN	30.0229	0.0	33.1055	0.0	30.3003	0.0	35.7012	0.0
SVM	50.2901	0.0	55.6145	0.0	47.6601	0.0	52.6841	0.0
Naïve 1	53.1315	0.0	56.3621	0.0	48.1125	0.0	53.1002	0.0

Note: The italic numbers indicate the lowest error rate (MAE and RMSE)

Table 5.
Out-of-sample
forecasting accuracy
for visitors

scheme was adopted to maintain a constant sample size over the out-of-sample forecasts. In this sense, for each forecast ($t + j$; j: 1,..., 12) the initial observations of the training periods are deleted and the corresponding ($t + j$)-1 forecasts are added.

Interestingly, in all cases RMSE, MAE and MCS provide a uniform result: a remarkable LSTM's forecasting superiority with the current database. Note that this model presents the best performance at the two relevant forecasting horizons, 6- and 12-step ahead, in both samples (residents and visitors) and for all periods. In all cases, LSTM has the lowest RMSE and MAE, indicating that the deviation between the predicted and the actual number of individuals is minimized as compared with all the competing models. This is corroborated by MCS. The LSTM p-values are equal to one, while for the other five models their values are below the 0.10 threshold. It means that these models would be removed in the MCS inspection process, and thus the LSTM would be the only survivor model.

From a methodological point of view, this robust result indicates that LSTM is an appropriate approach to forecast long HF sequences, at least for the current database. Note that it is a recurrent neural network, so the algorithm learns from previous episodes. This result corroborates the findings of Zhang *et al.* (2019), which also found a stronger performance of LSTM for long dependencies.

6. Discussion
Beyond the technical analysis provided in the previous paragraphs, this discussion section aims at First, illustrating how the HF forecasting approach proposed in this paper has

relevant practical applications for tourism and second, discussing several theoretical implications.

6.1 Practical implications

Let start emphasizing what would be provided to destinations' visitors, private and public tourism stakeholders: the above methodology implements a program that sequentially (for example, every 30 min or 1 h), generates an HF forecast (5 min frequency) of the number of residents and visitors in an area of interest (company's surroundings, tourism attraction, transport network, commercial street, city's pedestrian mobility bottleneck, etc.).

To visualize these forecasts, Figures 3 and 4 represent the out-of-sample forecasts and the actual number of residents (Figure 3) and visitors (Figure 4) observed after. We used the same periods described in Table 3.

Both figures display a 1-h (with 5 min frequency) out-of-sample models' forecast (what would have been provided to stakeholders) and the real number observed afterward. These real observations are the red lines, while the other series correspond to the six models' forecast, plus the Naïve 1 (horizontal green line). As found in Tables 4 and 5, the figures unanimously indicate the superiority of LSTM (orange). This model clearly displays a remarkable capacity to anticipate visitors' and residents' numbers.

It is particularly relevant to compare LSTM (orange, the chosen HF forecasting model), Naïve 1 (horizontal green line, which implies making future decisions based on the current situation) and the observed number (red, what subsequently happened in reality). Clearly, the proposed forecast is much closer to reality than a simple no-change estimation. Hence, by using this paper's approach, visitors and tourism stakeholders make decisions based on a robust crowdedness forecast, which is remarkably more accurate than using only current information.

These recursive forecasts can be integrated into any tourism management or information tool. Some examples include Tourism recommender systems that integrate crowdedness' forecasts; Online heat-maps showing pedestrian occupancy; Or advance alert systems that trigger appropriate policy responses when the forecasted number of users is above a certain threshold. Additionally, once the forecasts are shared, there are innumerable applications for visitors and public and private decision-making such as real-time visitors' decision on where to go avoiding crowds; In-site marketing activated at high occupancy's episodes; Traffic flow measures to improve urban mobility, as changing traffic lights' duration; Dynamic pricing on tourism attractions considering crowdedness; or public resource's allocation, as police or tourism information staff.

Moreover, forecasting cities' crowdedness is crucial for managing two challenges of many mature tourism destinations: guaranteeing social distancing related to Covid-19; And, with a longer-term perspective, dealing with overtourism. Regarding the former, the Covid-19 pandemic has brought to light social (or physical) distancing, as the main cost-efficient measure to deal with the pandemic. Many destinations are currently striving to be considered "safe" destinations. In this sense, implementing high-occupancy alert systems like those described above is a useful tool for destinations. With a more long-run perspective, it is likely that overtourism debates will emerge again. This paper's methodology provides a tool to anticipate episodes of overcrowding and implement mitigating policies.

6.2 Theoretical implications

The paper's main contributions, initiating HF forecasting in tourism and elaborating on its applications, are essentially empirical and practical. However, several theoretical implications are presented in the following paragraphs.

1990
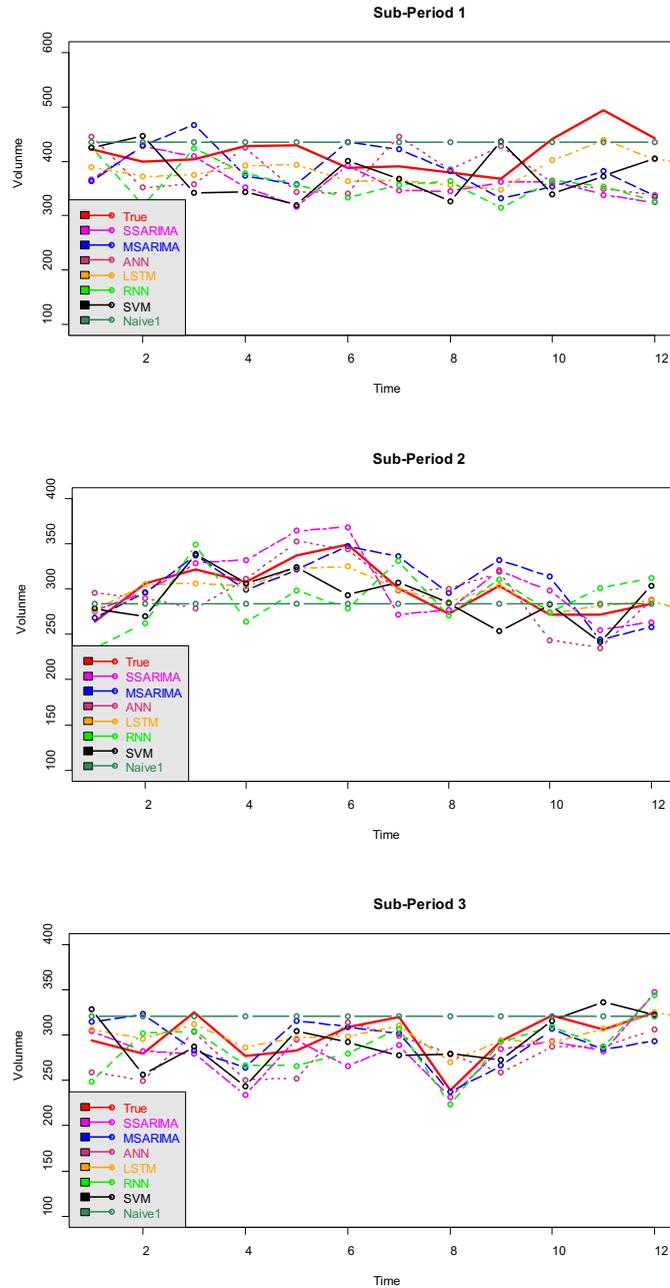
Figure 3.
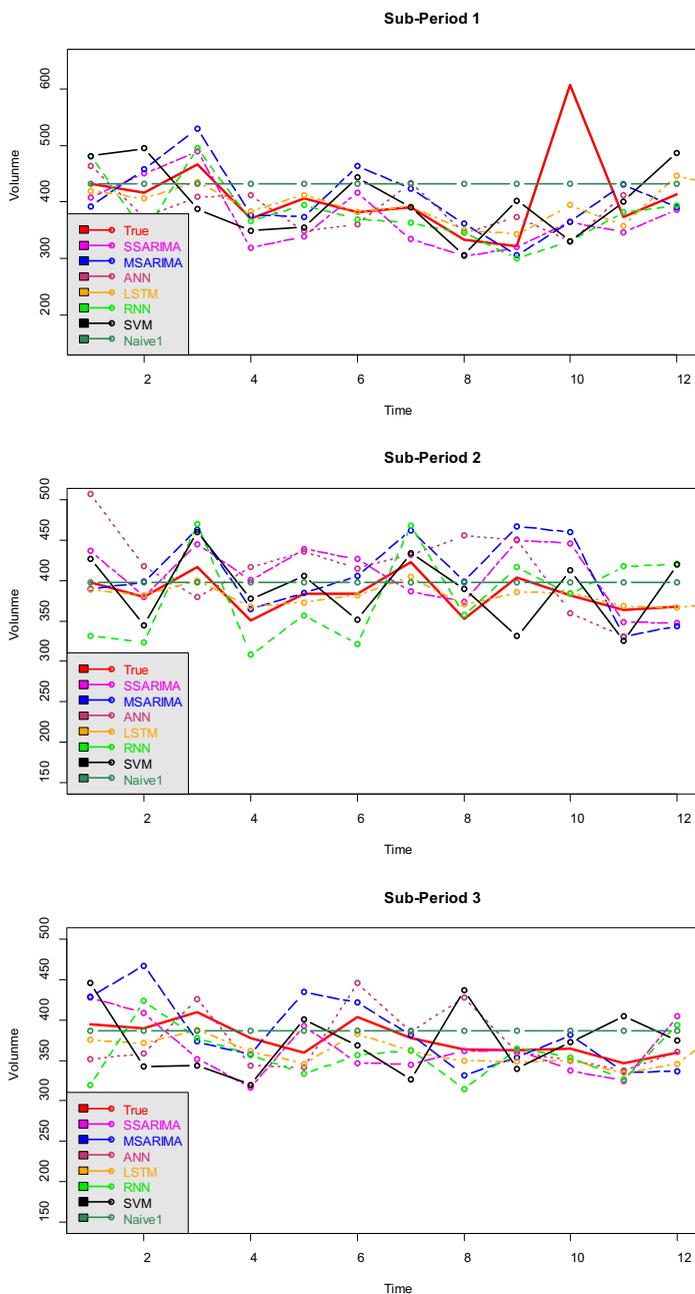Residents: real data
and forecasts

Figure 4.
Visitors: real data and
forecasts

The paper emphasizes that most aspects of human behavior are dynamic in nature (Järv *et al.*, 2018) and change in space and time. The relevance of combining these two dimensions was already highlighted in classical behavioral geography. Time geography (Hägerstraand, 1970) incorporated time budget allocation to study the space use. Since then, data capture techniques evolved from basic surveys and route diaries to portable devices and finally, tracking technologies. This process led to richer data sets that pave new research avenues. Tourism studies evolved from mainly focusing on the spatial analysis (Lew and McKercher, 2006; Xia *et al.*, 2009) to incorporate a detailed time dimension (Shoval and Ahas, 2016; Shoval and Isaacson, 2007). As data record methods were providing higher frequencies, researchers were able to tackle more topics, as intradiurnal activities' analysis (Birenboim *et al.*, 2013; Grinberger and Shoval, 2019). The currently available real-time or very HF data, as the big data used in this paper, allows new conceptualizations with short-time horizons.

Additionally, in most previous tourism studies considering spatiotemporal analyzes, like the ones mentioned above, the subjects of interest are the tourists and their behavior as they move throughout the destination. That is clearly the focus of time allocation as explained in time geography and of the above-cited papers which monitor and describe tourists' behavior. Differently, in this paper, the subject of interest is the destination and the dynamic use of its space. Any urban location has a given endowment of available space. When people crowds, this scarce resource is exhausted and time-dependent congestion problems appear. Hence, our conceptual approach is closer to the smart city paradigm (Kitchin, 2014) and its application to tourism as smart destinations (Buhalis and Amaranggana, 2013; Xiang *et al.*, 2015).

Finally, the proven forecasting accuracy of our models supports the spatiotemporal regularity of human mobility, which has been previously identified (Birenboim *et al.*, 2013; Song *et al.*, 2010; Zheng *et al.*, 2017). The individual regularity can be extended to understand aggregate regularities in the use of urban spaces.

7. Concluding remarks

During the last 20 years, ICT has invaded most aspects of our life. As a side effect, an overwhelming amount of information is now recorded in what is usually labeled as big data. On some occasions, these data sets provide extremely high frequencies (close to real-time data) and precise spatial resolutions, which were inconceivable only some years ago.

With the above considerations in mind, this paper highlights the relevance and applicability of tourism big data with HF. First, we faced the forecasting's challenges; and then, we discussed the usefulness of those estimates for visitors and tourism public and private stakeholders, whose decisions are increasingly focusing on short-time horizons.

We used big data (3.7 million daily observations) obtained from the technical communications between a public WiFi network and the mobile devices in its coverage area. In this sense, the collecting protocol described in the paper is easily replicable in other tourism destinations at a reasonable cost. The main data characteristics are its HF (by the second) and its precise geolocation.

Regarding HF forecasting, four artificial intelligence methods and two-time series models are compared to select the most accurate alternative for the current database. The paper provides both, conventional in-sample forecasting analysis and several out-of-sample exercises. The results robustly indicate that LSTM performs better considering loss-function analyzes (MAE, RMSE) and MCS tests. This paper uses a single context-specific database, so the model evaluation results, even if consistent with previous literature (Zhang *et al.*, 2019), might be different for other destinations. Nevertheless, the objective of the paper is the description of an HF forecasting methodology, which is seminal in tourism studies and not the specific model

selection. It is strongly recommended that subsequent applications to other tourism destinations or data sets, perform models' evaluation as suggested in the paper.

Beyond the technical analysis of the models, the discussion section's first part focuses on practical applications for tourism and hospitality. We detailed how HF forecasting can be used to improve both: on-site visitors' experience and tourism public and private decision-making in short-term horizons.

Additionally, we provided theoretical implications derived from the research conceptual approach. As with any emerging study area, the theoretical body is still immature. We believe that real-time data challenges are already here and will be increasingly attracting tourism scholars. This process will contribute to creating the required body of knowledge on HF analysis.

7.1 Limitations and future research

The current approach has some limitations that should be addressed by future research. First, it is basically an empirical exercise based on robust forecasting techniques. In this sense, it does not provide a theoretical foundation analysis. The methodologies used in the paper have been consistently proven superior for forecasting. However, they are not appropriate to identify causal relations or to perform policy evaluations (Song *et al.*, 2019; Song and Li, 2008). In this sense, future research should elaborate on the theoretical implications of HF forecasts in general and specifically in tourism and hospitality. Additionally, the data used in this paper is context-specific, as it has been captured at a single destination. The extension of HF forecasting to other tourism destinations and research objectives is needed to face current and future questions, which increasingly involve short-time horizons.

References

Ahas, R., Aasa, A., Roose, A., Mark, Ü. and Silm, S. (2008), "Evaluating passive mobile positioning data for tourism surveys: an Estonian case study", *Tourism Management*, Vol. 29 No. 3, pp. 469-486.

Ahas, R., Aasa, A., Silm, S. and Tiru, M. (2007), "Mobile positioning data in tourism studies and monitoring: case study in Tartu, Estonia", *ENTER*, pp. 119-128.

Alessandrini, A., Gioia, C., Sermi, F., Sofos, I., Tarchi, D. and Vespe, M. (2017), "WiFi positioning and big data to monitor flows of people on a wide scale", *European Navigation Conference (ENC). IEEE*, pp. 322-328.

Aqib, M., Mehmood, R., Alzahrani, A., Katib, I., Albeshri, A. and Altowaijri, S.M. (2019), "Smarter traffic prediction using big data, in-memory computing, deep learning and GPUs", *Sensors*, Vol. 19 No. 9, p. 2206.

Aslanargun, A., Mammadov, M., Yazici, B. and Yolacan, S. (2007), "Comparison of ARIMA, neural networks and hybrid models in time series: tourist arrival forecasting", *Journal of Statistical Computation and Simulation*, Vol. 77 No. 1, pp. 29-53.

Athanasopoulos, G., Hyndman, R.J., Song, H. and Wu, D.C. (2011), "The tourism forecasting competition", *International Journal of Forecasting*, Vol. 27 No. 3, pp. 822-844.

Bandara, K., Bergmeir, C. and Smyl, S. (2017), "Forecasting across time series databases using long short-term memory networks on groups of similar series", Vol. 8, pp. 805-815.

Bedi, J. and Toshniwal, D. (2019), "Deep learning framework to forecast electricity demand", *Applied Energy*, Vol. 238, pp. 1312-1326.

Birenboim, A., Anton-Clavé, S., Russo, A.P. and Shoval, N. (2013), "Temporal activity patterns of theme park visitors", *Tourism Geographies*, Vol. 15 No. 4, pp. 601-619.

Buhalis, D. and Amaranggana, A. (2013), "Smart tourism destinations", *Information and Communication Technologies in Tourism 2014*, Springer International Publishing, Cham, Vol. 31, pp. 553-564.

Chen, Z., Li, C. and Sun, W. (2020), "Bitcoin price prediction using machine learning: an approach to sample dimension engineering", *Journal of Computational and Applied Mathematics*, Vol. 365, p. 112395.

Chen, K.Y. and Wang, C.H. (2007), "Support vector regression with genetic algorithms in forecasting tourism demand", *Tourism Management*, Vol. 28 No. 1, pp. 215-226.

Chu, F.L. (1998), "Forecasting tourism demand in Asian-Pacific countries", *Annals of Tourism Research*, Vol. 25 No. 3, pp. 597-615.

Claveria, O. and Torra, S. (2014), "Forecasting tourism demand to Catalonia: neural networks vs. time series models", *Economic Modelling*, Vol. 36, pp. 220-228.

Cortes, C. and Vapnik, V. (1995), "Support-vector networks", *Machine Learning*, Vol. 20 No. 3, pp. 273-297.

Degiannakis, S. and Filis, G. (2018), "Forecasting oil prices: high-frequency financial data are indeed useful", *Energy Economics*, Vol. 76, pp. 388-402.

Dordonnat, V., Koopman, S.J., Ooms, M., Dessertaine, A. and Collet, J. (2008), "An hourly periodic state space model for modelling French national electricity load", *International Journal of Forecasting*, Vol. 24 No. 4, pp. 566-587.

Elghafghuf, A., Vanderstichel, R., St-Hilaire, S. and Stryhn, H. (2018), "Using state-space models to predict the abundance of juvenile and adult sea lice on Atlantic salmon", *Epidemics*, Vol. 24, pp. 76-87.

EU (2016), On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/E, European Union, 2016/679.

Geurts, M.D. and Ibrahim, I. (1975), "Comparing the Box-Jenkins approach with the exponentially smoothed forecasting model application to Hawaii tourists", *Journal of Marketing Research*, Vol. 12 No. 2, pp. 182-188.

Gretzel, U., Fesenmaier, D. and O'Leary, J. (2006), "The transformation of consumer behavior", in Buhalis, D. and Costa, C. (Eds), *Tourism Business Frontier*, Elsevier, Oxford.

Grinberger, A.Y. and Shoval, N. (2019), "Spatiotemporal contingencies in tourists' intradiurnal mobility patterns", *Journal of Travel Research*, Vol. 58 No. 3, pp. 512-530.

Hägerstraand, T. (1970), "What about people in regional science?", *Papers in Regional Science*, Vol. 24 No. 1, pp. 7-21.

Hansen, P.R., Lunde, A. and Nason, J.M. (2011), "The model confidence set", *Econometrica*, Vol. 79 No. 2, pp. 453-497.

Hardy, A., Hyslop, S., Booth, K., Robards, B., Aryal, J., Gretzel, U. and Eccleston, R. (2017), "Tracking tourists' travel with smartphone-based GPS technology: a methodological discussion", *Information Technology and Tourism*, Vol. 17 No. 3, pp. 255-274.

Harvey, A.C. and Phillips, G.D. (1979), "Maximum likelihood estimation of regression models with autoregressive-moving average disturbances", *Biometrika*, Vol. 66 No. 1, pp. 49-58.

Hochreiter, S. and Schmidhuber, J. (1997), "Long short-term memory", *Neural Computation*, Vol. 9 No. 8, pp. 1735-1780.

Huang, X., Li, M., Zhang, J., Zhang, L., Zhang, H. and Yan, S. (2020), "Tourists' spatial-temporal behavior patterns in theme parks: a case study of ocean park Hong Kong", *Journal of Destination Marketing and Management*, Vol. 15.

Hyndman, R., Koehler, A.B., Ord, J.K. and Snyder, R.D. (2008), *Forecasting with Exponential Smoothing: The State Space Approach*, Springer Science and Business Media.

INE (2020), "Tourist movement on borders survey (FRONTUR)", *Instituto nacional de Estadística*, available at: www.ine.es/en/index.htm

Khosravi, A., Koury, R.N.N., Machado, L. and Pabon, J.J.G. (2018), "Prediction of wind speed and wind direction using artificial neural network, support vector regression and adaptive neuro-fuzzy inference system", *Sustainable Energy Technologies and Assessments*, Vol. 25, pp. 146-160.

Kitchin, R. (2014), "The real-time city? Big data and smart urbanism", *GeoJournal*, Vol. 79 No. 1, pp. 1-14.

Kulendran, N. and Witt, S.F. (2003), "Forecasting the demand for international business tourism", *Journal of Travel Research*, Vol. 41 No. 3, pp. 265-271.

Kumar, K., Parida, M. and Katiyar, V.K. (2013), "Short term traffic flow prediction for a non urban highway using artificial neural network", *Procedia – Social and Behavioral Sciences*, Vol. 104 No. 2, pp. 755-764.

Kuo, P.H. and Huang, C.J. (2018), "A high precision artificial neural networks model for short-term energy load forecasting", *Energies*, Vol. 11 No. 1, p. 213.

Lachiheb, O. and Gouider, M.S. (2018), "A hierarchical deep neural network design for stock returns prediction", *Procedia Computer Science*, Vol. 126, pp. 264-272.

Lew, A. and McKercher, B. (2006), "Modeling tourist movements: a local destination analysis", *Annals of Tourism Research*, Vol. 33 No. 2, pp. 403-423.

Li, Y. and Cao, H. (2018), "Prediction for tourism flow based on LSTM neural network", *Procedia Computer Science*, Vol. 129, pp. 277-283.

Lim, C. and McAleer, M. (2000), "A seasonal analysis of Asian tourist arrivals to Australia", *Applied Economics*, Vol. 32 No. 4, pp. 499-509.

Lim, C. and McAleer, M. (2002), "Time series forecasts of international travel demand for Australia", *Tourism Management*, Vol. 23 No. 4, pp. 389-396.

Li, J., Xu, L., Tang, L., Wang, S. and Li, L. (2018), "Big data in tourism research: a literature review", *Tourism Management*, Vol. 68, pp. 301-323.

Loganathan, N. and Ibrahim, Y. (2010), "Forecasting international tourism demand in Malaysia using Box Jenkins Sarima application", *South Asian Journal of Tourism and Heritage*, Vol. 3 No. 2, pp. 50-60.

Ma, R., Zhou, C., Cai, H. and Deng, C. (2019), "The forecasting power of EPU for crude oil return volatility", *Energy Reports*, Vol. 5, pp. 866-873.

Mandic, D.P. and Chambers, J. (2001), *Recurrent Neural Networks for Prediction: learning Algorithms, Architectures and Stability*, John Wiley and Sons, Inc.

Mariani, M., Baggio, R., Fuchs, M. and Höepken, W. (2018), "Business intelligence and big data in hospitality and tourism: a systematic literature review", *International Journal of Contemporary Hospitality Management*, Vol. 30 No. 12, pp. 3514-3554.

Milano, C., Novelli, M. and Cheer, J.M. (2019), "Overtourism and tourismphobia: a journey through four decades of tourism development, planning and local concerns", *Tourism Planning and Development*, Vol. 16 No. 4, pp. 353-357.

Neuhofer, B., Buhalis, D. and Ladkin, A. (2012), "Conceptualising technology enhanced destination experiences", *Journal of Destination Marketing and Management*, Vol. 1 Nos 1/2, pp. 36-46.

Paoli, C., Notton, G., Nivet, M.L., Padovani, M. and Savelli, J.L. (2011), "A neural network model forecasting for prediction of hourly ozone concentration in Corsica", *2011 10th International Conference on Environment and Electrical Engineering*, IEEE, pp. 1-4.

Poornima, S. and Pushpalatha, M. (2019), "Prediction of rainfall using intensified LSTM based recurrent neural network with weighted linear units", *Atmosphere*, Vol. 10 No. 11, p. 668.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M.P., Shyu, M.L., Chen, S.C. and Iyengar, S.S. (2018), "A survey on deep learning: algorithms, techniques, and applications", *ACM Computing Surveys*, Vol. 51 No. 5, p. 92.

Ramos, P., Santos, N. and Rebelo, R. (2015), "Performance of state space and ARIMA models for consumer retail sales forecasting", *Robotics and Computer-Integrated Manufacturing*, Vol. 34, pp. 151-163.

Raun, J., Ahas, R. and Tiru, M. (2016), "Measuring tourism destinations using mobile tracking data", *Tourism Managemen*, Vol. 57, pp. 202-212.

Rundo, F., Trenta, F., di Stallo, A.L. and Battiato, S. (2019), "Machine learning for quantitative finance applications: a survey", *Applied Sciences*, Vol. 9 No. 24, p. 5574.

Scott, N., Van Niekerk, M. and De Martino, M. (2017), *Knowledge Transfer to and within Tourism: Academic, Industry and Government Bridges*, Emerald Publishing, Bingley, Vol. 8.

Seyyedsalehi, S.Z. and Seyyedsalehi, S.A. (2015), "A fast and efficient pre-training method based on layer-by-layer maximum discrimination for deep neural networks", *Neurocomputing*, Vol. 168, pp. 669-680.

Shintate, T. and Pichl, L. (2019), "Trend prediction classification for high frequency bitcoin time series with deep learning", *Journal of Risk and Financial Management*, Vol. 12 No. 1, p. 17.

Shoval, N. and Ahas, R. (2016), "The use of tracking technologies in tourism research: the first decade", *Tourism Geographies*, Vol. 18 No. 5, pp. 587-606.

Shoval, N. and Isaacson, M. (2007), "Tracking tourists in the digital age", *Annals of Tourism Research*, Vol. 34 No. 1, pp. 141-159.

Smola, A.J. and Scholkopf, B. (2004), "A tutorial on support vector regression", *Statistics and Computing*, Vol. 14 No. 3, pp. 199-222.

Song, H. and Li, G. (2008), "Tourism demand modelling and forecasting – a review of recent research", *Tourism Management*, Vol. 29 No. 2, pp. 203-220.

Song, H., Qiu, R.T.R. and Park, J. (2019), "A review of research on tourism demand forecasting", *Annals of Tourism Research*, Vol. 75 No. September 2018, pp. 338-362.

Song, C., Qu, Z., Blumm, N. and Barabási, A.L. (2010), "Limits of predictability in human mobility", *Science*, Vol. 327 No. 5968, pp. 1018-1021.

Svetunkov, I. and Boylan, J.E. (2020), "State-space ARIMA for supply-chain forecasting", *International Journal of Production Research*, Vol. 58 No. 3, pp. 818-827.

Thushara, S.C., Su, J.J. and Bandara, J.S. (2019), "Forecasting international tourist arrivals in formulating tourism strategies and planning: the case of Sri Lanka", *Cogent Economics and Finance*, Vol. 7 No. 1, p. 1699884.

Wang, D., Park, S. and Fesenmaier, D. (2012), "The role of smartphones in mediating the touristic experience", *Journal of Travel Research*, Vol. 51 No. 4, pp. 371-387.

Witt, S.F. and Witt, C.A. (1991), "Tourism forecasting: error magnitude, direction of change error and trend change error", *Journal of Travel Research*, Vol. 30 No. 2, pp. 26-33.

Xiang, Z., Tussyadiah, I. and Buhalis, D. (2015), "Smart destinations: foundations, analytics, and applications", *Journal of Destination Marketing and Management*, Vol. 4 No. 3, pp. 143-144.

Xia, J.C., Zeephongsekul, P. and Arrowsmith, C. (2009), "Modelling spatio-temporal movement of tourists using finite Markov chains", *Mathematics and Computers in Simulation*, Vol. 79 No. 5, pp. 1544-1553.

Xia, J.C., Zeephongsekul, P. and Packer, D. (2011), "Spatial and temporal modelling of tourist movements using Semi-Markov processes", *Tourism Management*, Vol. 32 No. 4, pp. 844-851.

Xing, Y., Ban, X., Liu, X. and Shen, Q. (2019), "Large-scale traffic congestion prediction based on the symmetric extreme learning machine cluster fast learning method", *Symmetry*, Vol. 11 No. 6, p. 730.

Xu, X., Law, R., Chen, W. and Tang, L. (2016), "Forecasting tourism demand by extracting fuzzy Takagi–Sugeno rules from trained SVMs", *CAAI Transactions on Intelligence Technology*, Vol. 1 No. 1, pp. 30-42.

Yin, T., Zhong, G., Zhang, J., He, S. and Ran, B. (2017), "A prediction model of bus arrival time at stops with multi-routes", *Transportation Research Procedia*, Vol. 25, pp. 4623-4636.

Zhang, G., Patuwo, B.E. and Hu, M.Y. (1998), "Forecasting with artificial neural networks: the state of the art", *International Journal of Forecasting*, Vol. 14 No. 1, pp. 35-62.

Zhang, B., Pu, Y., Wang, Y. and Li, J. (2019), "Forecasting hotel accommodation demand based on LSTM model incorporating internet search index", *Sustainability*, Vol. 11 No. 17, p. 4708.

Zheng, W., Huang, X. and Li, Y. (2017), "Understanding the tourist mobility using GPS: where is the next place?", *Tourism Management*, Vol. 59, pp. 267-280.

Further reading

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*, Vol. 46, pp. 454-464.

Appendix. SVM

This model includes a training data set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ with n inputs; y is the dependent variable explained by the inputs $x = \{y_{t-1}, \ldots, y_{t-p}\}$, which are lagged values of y. SVM model develops a mapping $f(x)$: $\Re^n \to \Re$ to fit input data into the so-called high dimensional feature space $\Re$:

$$f(x, W) = W' \cdot \phi(x) + b, \tag{1}$$

where W is the weight parameters and ϕ is a nonlinear transformation function. This procedure transforms the non-linear input space into a high-dimensional linear feature space. In the computation aspect, we can determine the unknown W by minimizing the sum of the loss function $loss()$ and a complexity term $\frac{1}{2}\|w\|^2$; thus, a convex optimization problem can be written as:

$$\frac{1}{2}\|W\|^2 + loss\left(\sum_{i=1}^{n}\left(\gamma_i + \gamma_i^*\right)\right), \tag{2}$$

subject to the following additional constraints:

$$y_i - f(x, W) \leq \varepsilon + \gamma_i^*, \tag{3}$$

$$f(x, W) - y_i \leq \varepsilon + \gamma_i, \tag{4}$$

$$\gamma_i, \gamma_i^* \geq 0. \tag{5}$$

where γ_i, γ_i^* are the slack variables used to cope with optimization problem's infeasible constraints. ε is the tolerated error. By using the Lagrange approach, equations (2)-(5) can be expressed as:

$$\mathcal{L} = \frac{1}{2}\|W\|^2 + loss\left(\sum_{i=1}^{n}\left(\gamma_n + \gamma_n^*\right)\right) - \sum_{i=1}^{n}\lambda_i\left(-\varepsilon - \gamma_i^* + y_i - f(x, W)\right)$$

$$- \sum_{i=1}^{n}\lambda_i^*\left(-\varepsilon - \gamma_i - y_i + f(x, W)\right), \tag{6}$$

where $\mathcal{L}$ is the Lagrangian function and $\lambda_i^*, \lambda_i > 0$ are its multipliers.

ANN

The input hidden layer is modeled by summing the path of input weights times the input data and adding the bias node (b). Thus, these functions can be expressed as:

$$H_j^I = g\left(x' w_j^I + b_j^I\right), \tag{7}$$

where H_j^I are the j-th hidden neuron's input and the vector $w^I = \{w_1^I, \ldots, w_j^I\}$ captures the weights that show the strength of the path between the hidden and input layers. $g(\cdot)$ is the hidden layer logistic activation function. b_j^I is the bias term of the input layer. H_j^I is transformed into output y, through the output layer's activation function. Thus, the model can be written as:

$$\hat{y}_t = f\left(H_t^I w_j^o + b_j^o\right), \tag{8}$$

where $\hat{y}_t$ is the estimated variable; the vector $w_j^o = \{w_1^o, \ldots, w_j^o\}$ represents the output weights connecting hidden and output layers; b_j^o is the bias term of the output layer; Finally, the output layer activation function is $f(\cdot)$.

The ML protocol typically uses backpropagation to adjust the path weights and the node biases to minimizing the difference between observed and estimated output that match the data. The loss function can be formulated as follows:

$$Loss\left(w^I, w^o\right) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \hat{y}_t | w^I, w^o\right)^2 \tag{9}$$

Thus, the optimal weights' vector $\hat{\mathbf{w}} = \{\hat{w}^I, \hat{w}^o\}$ is obtained as:

$$\hat{\mathbf{w}} = \underset{w^I, w^o}{\operatorname{argmin}} = loss\left(w^I, w^o\right) \tag{10}$$

RNN

Figure A1 shows an example of the architecture of the RNN.

The current state h_t at timestamp t is computed considering the contemporary input x_t and the previously hidden state h_{t-1}. Mathematically it can be expressed as:

$$h_t = f_\theta(Ux_t + Wh_{t-1}) \tag{11}$$

$$o_t = f_\alpha(Vh_t) \tag{12}$$

where o_t denotes the output at the timestamp t; f_θ and f_α are sigmoid activation functions; and U, W_s and V are the weight parameters. The sigmoid function is appropriate for non-linear data with high-value variability and it is a frequently used ANN activation function (Claveria and Torra, 2014; Poornima and Pushpalatha, 2019).
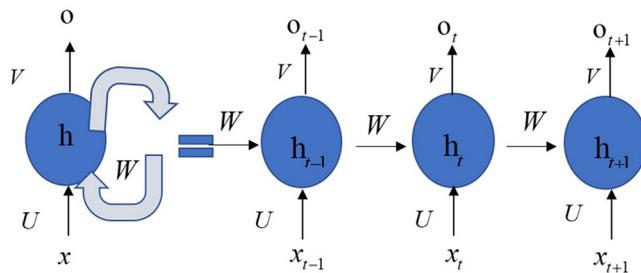
Figure A1.
Example of Recurrent
Neural Network's
Architecture

LSTM

Figure 2 illustrates the architecture of the hidden layer in the LSTM network (Figure A2).

c_t denotes the memory cell at timestamp t that is used to replace the RNN's hidden layer neurons. At each timestamp, a few layers are used to regulate the information along with the sequences and thereby capturing any long-term dependencies. To do so, the hidden layer h_t is updated considering the information from: input x_t, hidden layer h_{t-1} and three gates (input, I_t; forget, f_t; and output, o_t). These gates are represented as:

$$f_t = sigmoid\left(W_f \cdot [x_t \cdot h_{t-1}] + b_f\right), \tag{13}$$

where $sigmoid(\cdot)$ is the activation function; $[x_t h_{t-1}]$ is the vector of input x_t and hidden layer h_{t-1}; while W_f and b_f represent, respectively, the weights' matrix and the bias of the forget gate. Note that this forgets gate determines the number of cell states from previous time c_{t-1}, that are reserved for the cell state c_t. A zero weight would be assigned when the information is eliminated and one if the information is preserved.

The input gate determines how much of the current input x_t is reserved into the cell state c_t. Thus, the input gate becomes:

$$I_t = sigmoid(W_I \cdot [h_{t-1} \cdot x_t] + b_I), \tag{14}$$

where W_I and b_I are also the weights' matrix and bias of this gate. Then, the cell state can be updated at timestamps t as:

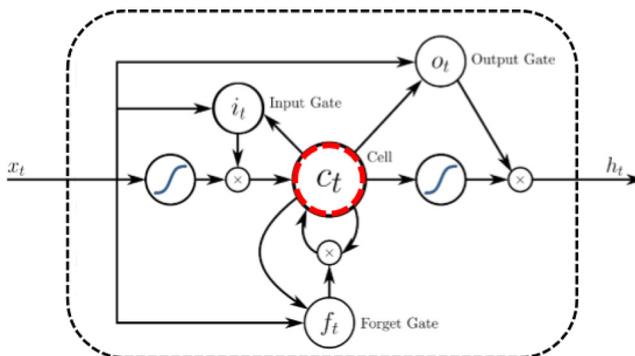$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1} \cdot x_t] + b_c) \tag{15}$$

$$c_t = f_t \cdot c_{t-1} \cdot I_t \cdot \tilde{c}_t, \tag{16}$$

Similarly, W_c and b_c correspond to the weights' matrix and bias in the cell state. $\tanh(\cdot)$ is an activation hyperbolic tangent function used to rescale the logistic sigmoid. Finally, the output value of the cell is defined as:

$$o_t = sigmoid(W_o \cdot [h_{t-1} \cdot x_t] + b_o), \tag{17}$$

$$h_t = o_t \cdot \tanh(c_t), \tag{18}$$

where W_o is the weights' matrix and b_o is the bias.

Figure A2.
Architecture of
LSTM network

SSARIMA

The mathematical representation of the proposed SSARIMA is:

$$
\begin{aligned}
y_t &= v_{t-1} + \varepsilon_t, \\
v_{j,t} &= \phi_j v_{j,t-1} + v_{j+1,t-1} + v_{K+1,t-1} + \left(\phi_j + \eta_j\right)\varepsilon_t, \quad \text{for } j = 1, \\
v_{j,t} &= \phi_j v_{j,t-1} + v_{j+1,t-1} + \left(\phi_j + \eta_j\right)\varepsilon_t, \qquad\qquad \text{for } 1 < j \le K, \\
v_{K+1,t} &= v_{K+1,t-1,}
\end{aligned}
\tag{19}
$$

where $v_{j,t}$ is the j-th component of AR and MA terms and $v_{K+1,0} = \alpha_0$ equation (22) then can be written as:

$$
\begin{aligned}
y_t &= \boldsymbol{\omega}' \mathbf{v}_{t-1} + \varepsilon_t, \\
\mathbf{v}_t &= \mathbf{F}\mathbf{v}_{t-1} + \mathbf{G}\varepsilon_t,
\end{aligned}
\tag{20}
$$

where $\boldsymbol{\omega}$ is the measurement vector, $\mathbf{F}$ is a transition matrix and $\mathbf{G}$ is the persistence vector. These three components can be expanded as:

$$
\mathbf{F} = \begin{pmatrix} \phi_1 & 1 & 0 & \cdots & 0 & 1 \\ \phi_2 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \phi_K & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, \boldsymbol{\omega} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \mathbf{G} = \begin{pmatrix} \phi_1 + \eta_1 \\ \phi_2 + \eta_2 \\ \vdots \\ \phi_K + \eta_k \\ 0 \end{pmatrix}.,
\tag{21}
$$

MCS

The EPA hypothesis for a given set of models M can be formulated as:

$$
\begin{aligned}
H_{0,M} &: \mathrm{E}\left(d_{ij}\right) = 0 \quad \text{for all} \quad i,j = 1, \ldots, m \\
H_{a,M} &: \mathrm{E}\left(d_{ij}\right) \ne 0 \quad \text{for all} \quad i,j = 1, \ldots, m,
\end{aligned}
\tag{22}
$$

Different criteria can be used to establish EPA. In this study, we use the "range" statistic:

$$
T_R = \max_{i,j \in M} \frac{\left|\overline{d}_{ij}\right|}{\sqrt{\widehat{\mathrm{var}}\left(\overline{d}_{ij}\right)}},
\tag{23}
$$

where $\overline{d}_{ij} = m^{-1} \sum_{t=1}^{m} d_{ij,t}$ is the mean loss differential between each pair of forecasting models i and j; $d_{ij,t}$ captures the sample loss between these models at time t; and $\widehat{\mathrm{var}}\left(\overline{d}_{ij}\right)$ is the bootstrapped variance estimate of this difference.

Corresponding author
Vicente Ramos can be contacted at: vicente.ramos@uib.es