Improved tourism demand forecasting with CIR# model: a case study of disrupted data patterns in Italy

Michele Bufalo and Giuseppe Orlando

Abstract

Purpose – *This study aims to predict overnight stays in Italy at tourist accommodation facilities through a nonlinear, single factor, stochastic model called CIR#. The contribution of this study is twofold: in terms of forecast accuracy and in terms of parsimony (both from the perspective of the data and the complexity of the modeling), especially when a regular pattern in the time series is disrupted. This study shows that the CIR# not only performs better than the considered baseline models but also has a much lower error than other additional models or approaches reported in the literature.*

Design/methodology/approach – *Typically, tourism demand tends to follow regular trends, such as low and high seasons on a quarterly/monthly level and weekends and holidays on a daily level. The data set consists of nights spent in Italy at tourist accommodation establishments as collected on a monthly basis by Eurostat before and during the COVID-19 pandemic breaking regular patterns.*

Findings – *Traditional tourism demand forecasting models may face challenges when massive amounts of search intensity indices are adopted as tourism demand indicators. In addition, given the importance of accurate forecasts, many studies have proposed novel hybrid models or used various combinations of methods. Thus, although there are clear benefits in adopting more complex approaches, the risk is that of dealing with unwieldy models. To demonstrate how this approach can be fruitfully extended to tourism, the accuracy of the CIR# is tested by using standard metrics such as root mean squared errors, mean absolute errors, mean absolute percentage error or average relative mean squared error.*

Research limitations/implications – *The CIR# model is notably simpler than other models found in literature and does not rely on black box techniques such as those used in neural network (NN) or data science-based models. The carried analysis suggests that the CIR# model outperforms other reference predictions in terms of statistical significance of the error.*

Practical implications – *The proposed model stands out for being a viable option to the Holt–Winters (HW) model, particularly when dealing with irregular data.*

Social implications – *The proposed model has demonstrated superiority even when compared to other models in the literature, and it can be especially useful for tourism stakeholders when making decisions in the presence of disruptions in data patterns.*

Originality/value – *The novelty lies in the fact that the proposed model is a valid alternative to the HW, especially when the data are not regular. In addition, compared to many existing models in the literature, the CIR# model is notably simpler and more transparent, avoiding the "black box" nature of NN and data science-based models.*

Keywords *Tourism, ARIMA, Forecasting, EGARCH, COVID-19, SARIMA, CIR#, Holt–Winters, DNNAR*

Paper type *Research paper*

利用*cir#*模型的旅游需求预测改进:意大利中断数据模式的案例研究

摘要

设计/方法/方法: 一般来说，旅游需求往往遵循规律的趋势，例如季度/月的淡季和旺季，以及日常的周末和假期。该数据集包括欧盟统计局在打破常规模式的*2019*冠状病毒病大流行之前和期间每月收集的在意大利旅游住宿设施度过的夜晚。

目的: 本研究旨在通过一个名为*cir#*的非线性单因素随机模型来预测意大利游客住宿设施的过夜住宿情况。这项研究的贡献是双重的:在预测准确性方面和在简洁方面(从数据和建模复杂性的角度来看)，特别是当时间序列

Michele Bufalo is based at Department of Methods and Models for Economics, Territory and Finance, Università degli Studi di Roma La Sapienza, Roma, Italy.
Giuseppe Orlando is based at Department of Mathematics, Università degli Studi di Bari Aldo Moro, Bari, Italy.

JEL classification – C22, C53, L83, Z3

Received 14 April 2023
Revised 10 May 2023
16 June 2023
Accepted 16 June 2023

Since acceptance of this article, the following author has updated their affiliations: Giuseppe Orlando is also at the Department of Economics, HSE University, Saint Petersburg, Russia.

中的规则模式被打乱时。我们表明，*cir#*不仅比考虑的基线模型表现更好，而且比文献中报告的其他模型或方法具有更低的误差。

研究结果：当大量搜索强度指标被作为旅游需求指标时，传统的旅游需求预测模型将面临挑战。此外，鉴于准确预测的重要性，许多研究提出了新的混合模型或使用各种方法的组合。因此，尽管采用更复杂的方法有明显的好处，但风险在于处理难使用的模型。为了证明这种方法能有效地扩展到旅游业，使用*RMSE*、*MAE*、*MAPE*或*AvgReIMSE*等标准指标来测试*cir#*的准确性。

研究局限/启示：*cir#*模型明显比文献中发现的其他模型简单，并且不依赖于黑盒技术，例如在神经网络或基于数据科学的模型中使用的技术。所进行的分析表明，*cir#*模型在误差的统计显著性方面优于其他参考预测。

实际意义：这个模型作为*Holt-Winters*模型的一个拟议模型，特别是在处理不规则数据时。

社会影响：即使与文献中的其他模型相比，所提出的模型也显示出优越性，并且在数据模式中断时对旅游利益相关者做出决策特别有用。

创意/价值：创新之处在于所提出的模型是*Holt-Winters*模型的有效替代方案，特别是当数据不规律时。此外，与文献中的许多现有模型相比，*cir#*模型明显更简单、更透明，避免了神经网络和基于数据科学的模型的"黑箱"性质。

关键词 旅游业，预测，*SARIMA*，霍尔特温特斯，*CIR#*

文章类型 研究型论文

Mejora en la previsión de la demanda turística con el modelo CIR#: un estudio de caso de patrones de datos interrumpidos en Italia

Resumen

Diseño/metodología/enfoque: *Normalmente, la demanda turística tiende a seguir tendencias regulares, como temporadas altas y bajas a nivel trimestral/mensual y fines de semana y festivos a nivel diario. El conjunto de datos consiste en las pernoctaciones en Italia en establecimientos de alojamiento turístico recogidas mensualmente por Eurostat antes y durante la pandemia de COVID-19, rompiendo los patrones regulares.*

Objetivo: *El presente estudio pretende predecir las pernoctaciones en Italia en establecimientos de alojamiento turístico mediante un modelo estocástico no lineal de un solo factor denominado CIR#. La contribución de este estudio es doble: en términos de precisión de la predicción y en términos de parsimonia (tanto desde la perspectiva de los datos como de la complejidad de la modelización), especialmente cuando un patrón regular en la serie temporal se ve interrumpido. Demostramos que el CIR# no sólo aplica mejor que los modelos de referencia considerados, sino que también tiene un error mucho menor que otros modelos o enfoques adicionales de los que se informa en la literatura.*

Resultados: *Los modelos tradicionales de previsión de la demanda turística pueden enfrentarse a desafíos cuando se adoptan cantidades masivas de índices de intensidad de búsqueda como indicadores de la demanda turística. Además, dada la importancia de unas previsiones precisas, muchos estudios han propuesto modelos híbridos novedosos o han utilizado diversas combinaciones de métodos. Así pues, aunque la adopción de enfoques más complejos presenta ventajas evidentes, el riesgo es el de enfrentarse a modelos poco manejables. Para demostrar cómo este enfoque puede extenderse de forma fructífera al turismo, se comprueba la precisión del CIR# utilizando métricas estándar como RMSE, MAE, MAPE o AvgReIMSE.*

Limitaciones/implicaciones de la investigación: *El modelo CIR# es notablemente más sencillo que otros modelos encontrados en la literatura y no se basa en técnicas de caja negra como las utilizadas en los modelos basados en redes neuronales o en la ciencia de datos. El análisis realizado sugiere que el modelo CIR# supera a otras predicciones de referencia en términos de significación estadística del error.*

Implicaciones prácticas: *El modelo propuesto destaca por ser una opción viable al modelo Holt-Winters, sobre todo cuando se trata de datos irregulares.*

Implicaciones sociales: *El modelo propuesto ha demostrado su superioridad incluso cuando se compara con otros modelos de la bibliografía, y puede ser especialmente útil para los agentes del sector turístico a la hora de tomar decisiones cuando se producen alteraciones en los patrones de datos.*

Originalidad/valor: *La novedad radica en que el modelo propuesto es una alternativa válida al Holt-Winters especialmente cuando los datos no son regulares. Además, en comparación con muchos modelos existentes en la literatura, el modelo CIR# es notablemente más sencillo y transparente, evitando la naturaleza de "caja negra" de los modelos basados en redes neuronales y en ciencia de datos.*

Palabras clave *Turismo, Previsión, SARIMA, Holt–Winters, CIR#*

Tipo de papel *Trabajo de investigación*

1. Introduction

Tourism is one of the fastest growing industries in the world, the third largest export category and is crucial for many small developing countries which account for up to 50% of total exports (UNWTO, 2021a, 2021b). Tourist arrivals in 2019 were 1.5 billion, confirming sustained growth for the tenth consecutive year. Tourism as a fair weather sector is heavily dependent on friendly framework conditions such as crime, hospitality and health (Burini, 2020). Frechtling (2012) classified factors affecting tourism into push, pull and resistance. According to Meleddu and Pulina (2016), pull factors pertain to the tourist destination, such as the quality of the natural resources and awareness of ecotourism. For Martins *et al.* (2017), push factors reflect both macroeconomic growth and specific conditions of the source market, such as leisure time, per capita income and relative prices. Conversely, resistance factors are those constraining travel from the source market to the destination such as corruption (Poprawe, 2015; Saha and Yap, 2015) and purchase power (Martins *et al.*, 2017). The motivation to travel can be categorized into intrinsic and extrinsic motivations, each associated with different emotional experiences (Marino and Pariso, 2021).

Italy, with 64.5 million international tourist arrivals in 2019, was the fifth most visited destination in the world (see UNWTO, 2020). Tourism is a highly volatile industry depending on seasonality, social trends, connectivity and infrastructure. This volatility might be further exacerbated by both internal and external shocks. Regarding the latter, the industry was badly hit by the COVID-19 pandemic to the point that the World Tourism Organization (UNWTO) expected a decline of over 70% in 2020, back to levels of 30 years ago (UNWTO, 2020). Still in 2021, according to the latest UNWTO data, international tourist arrivals are expected to remain 70%–75% below 2019 levels (UNWTO, 2021a, 2021b).

Despite current difficulties, the tourism industry has experienced phenomenal growth over the last 30 years generating demand in modeling and forecasting. In fact, "accurate forecasts are critical for destinations where the decision-makers try to capitalize on developments in the tourism market and/or to balance their local ecological and social carrying capacities" Song *et al.* (2019). Song and Witt (2012) provided an account of econometric modeling methodologies, and Song *et al.* (2019) reviewed 211 key papers published between 1968 and 2018 with the intent to explain how the methods of forecasting tourism demand have evolved over time. The authors, therefore, concluded that there is no single method that works well for all situations and that the evolution of forecasting methods is still ongoing. That is why, for benchmarking, we decided to take six different baseline models, from time series to econometric and artificial intelligence models.

As mentioned by Law *et al.* (2019), "traditional tourism demand forecasting models may face challenges when massive amounts of search intensity indices are adopted as tourism demand indicators." In addition, given the importance of accurate forecasts, many studies have proposed novel hybrid models or used various combinations of methods. Thus, although there are clear benefits in adopting more complex approaches, the risk is that of dealing with unwieldy models. Among them, we mention data requirements and related issues (e.g. availability, cleansing and validation), estimation issues and model risk. In addition, typically, tourism demand tends to follow regular trends, such as low and high seasons on a quarterly/monthly level and weekends and holidays on a daily level (Hu *et al.*, 2021). This has been disrupted by the COVID-19 pandemic (Xie *et al.*, 2020). In fact, unexpected excessive tourist demand and consumption place significant strain on resources and infrastructure. This strain can pose challenges for local authorities in terms of investment management and workforce recruitment, ultimately disrupting the stability of community economics. As a result, precise forecasting of tourist demand across different resources can greatly benefit researchers, industry workers and local authorities responsible for decision-making (Yao and Cao, 2020).

As a solution to the above-mentioned issues, in the present work, we propose a single factor, parsimonious stochastic model as a practical means of tourism forecasting. The suggested

approach is robust to the breaking of regular patterns which is one of the main contributions to the current literature. Over the years, Orlando *et al.* (2018, 2019a, 2019b, 2020) and Orlando and Bufalo (2021a) have developed the so-called CIR# model for forecasting time series in finance. The model, while parsimonious in terms of data and complexity, has proven robust when it comes to modeling cluster volatility, regime changes and spikes in time series that can typically be observed during turbulent periods such as the COVID-19 pandemic (Yonar *et al.*, 2020). To demonstrate how this approach can be fruitfully extended to tourism, the accuracy of the CIR# is tested by using standard metrics such as root-mean-squared errors (RMSE), mean absolute errors (MAE), mean absolute percentage error (MAPE) or average relative mean squared error (AvgReIMSE). We show that the CIR# not only performs better than the considered baseline models but also has a much lower error than other additional models or approaches reported in the literature such as those by Gunter and Önder (2015), Kourentzes and Athanasopoulos (2019), Di Fonzo and Girolimetto (2022) and Wu *et al.* (2022). This holds true for the entire data set, as well as during the COVID-19 pandemic.

The main novelty of the proposed approach is its effectiveness as an alternative to the Holt–Winters (HW) model and artificial intelligence techniques, particularly in cases where data irregularity is a concern. Compared to many advanced existing models in the literature, the CIR# model is notably simpler and more transparent, avoiding the black box nature of neural network (NN) and data science-based models. In summary, we demonstrate that the CIR# model effectively handles the challenges posed by tourism time series when they exhibit characteristics such as positive kurtosis, nonnormality, autocorrelation and heteroscedasticity. This suitability for future purposes makes it valuable for tourism stakeholders who require reliable forecasts, particularly during periods of volatility in data patterns (Xie *et al.*, 2020). The timing of the research is significant due to the increased need for accurate tourism demand forecasting, especially in the context of tourism disruptions caused by sudden travel restrictions, changes in consumer behavior and economic crises.

The remainder of the article is organized as follows. Section 2 provides a sketch of the existing literature. Section 3 presents the data set as well as its statistical characteristics. Section 4 is divided into three parts: proposed model, baseline models and measures of accuracy. Section 5 reports the obtained out-of-sample results. Section 6 provides a summary on the applicability of models and discusses results and the implications of the research. The last section concludes.

2. Literature review

As mentioned, Song *et al.* (2019) provided a comprehensive review of studies published on tourism demand forecasting. Essentially, forecasting methods fall into three categories: time series, regression and artificial intelligence models.

2.1 Time series models

Time series models such as the autoregressive integrated moving average (ARIMA) model have been used by Park *et al.* (2017) to internet search data from Google Trends to provide short-term forecasts for the inflow of Japanese tourists to South Korea. Generalized autoregressive conditional heteroscedastic (GARCH) model can be used in combination to forecast variances. Chan *et al.* (2005) applied ARIMA-EGARCH (exponential generalized autoregressive conditional heteroscedasticity) model from Japan, New Zealand, UK and USA to Australia confirming interdependency in the conditional variances between the four leading countries. Similarly, Coshall (2009) used an ARIMA-GARCH model for the UK demand for international tourism showing that volatility is relevant for tourism demand and that effects are asymmetric. The exponentially weighted moving average (EWMA) predicts future values based on appropriately weighted past observations, giving more importance to recent observations (Holt, 2004). Exponential smoothing methods have been developed by

Holt (2004) and Winters (1960). The HW approach is an extension of EWMA incorporating linear trend and seasonality into the exponential smoothing (Rosy and Ponnusamy, 2017). Guojun and Ningning (2021) suggested exponential smoothing and ARIMA predict the number of tourists and the total amount of tourism consumption in Guangxi. Abdulmajeed *et al.* (2020), to forecast COVID-19 cases in Nigeria, devised an ARIMA model and a HW exponential smoothing model combined with a GARCH.

2.2 Regression models

Gunter and Önder (2015) compared the predictive accuracy of various uni- and multivariate models in forecasting international city tourism demand for Paris from its five most important foreign source markets (Germany, Italy, Japan, UK and USA). Specifically, they used seven different forecast models, i.e. error correction-autoregressive distributed lag (EC-ADLM), classical and Bayesian vector autoregression (BVAR), time-varying parameter (TVP), autoregressive moving average (ARMA) and exponential smoothing (ETS), as well as the naïve-1 model. However, the outcome is that there is no clear indication of the best model across countries and forecast horizons. Tratar and Strmčnik (2016), while studying heat load, found that multiple regression is the best for daily and weekly short-term forecasting, whereas HW method suits better for monthly and longer horizons forecasting.

2.3 Artificial intelligence models

For artificial intelligence approaches, a popular choice is machine learning based on NNs. Silva *et al.* (2019) claimed that hybrid singular spectrum analysis (SSA) combined with a neural network autoregression model (DNNAR) outperforms ARIMA forecasts in most cases. This is because "it is noise and not seasonality which hinders neural network autoregression (NNAR) forecasting capabilities." In this sense, SSA can be used for removing noise from data so that the NNAR model can be used on "smooth" data.

To tackle the problem of insufficient interpretability in tourism demand forecasting, Wu *et al.* (2022) proposed the use of a temporal fusion transformer (TFT) model, optimized using an adaptive differential evolution algorithm (ADE). TFT is a newly developed attention-based deep learning model that offers both high-performance prediction and time-dynamic interpretable analysis.

2.4 Alternative approaches

As the "no free lunch" theorem holds (see, e.g. Wolpert, 1996; Wolpert and Macready, 1997), one has to recognize that there is no single method outperforming the others on all scenarios in terms of accuracy and that all methods have their own limitations. For example, Law *et al.* (2019) stated that "time-series and econometrics models rely on the stability of historical patterns and economic structure, while artificial intelligence models are dependent on the quality and size of available training data." Thus, the choice between different models depends on the type and quality of data having in mind the trade-off between complexity and improved modeling accuracy. Along these lines, Orlando *et al.* (2018, 2019a, 2019b, 2020) and Orlando and Bufalo (2021b) have suggested an effective parsimonious nonlinear stochastic model called CIR# for modeling time series in presence of skewed distributions, jumps and cluster volatility. This, through an appropriate partitioning, shifting and calibration of the time series, has proved effective in terms of directionality of data and forecasting error even in case of disruption of data patterns.

3. Data set

The data set consists of nights spent in Italy at tourist accommodation establishments as collected on a monthly basis by Eurostat (2021).

Figure 1 shows, for the Eurostat code NACE_R2, the total nights reported by hotels, holiday and other short-stay accommodations, camping grounds, recreational vehicle parks and trailer parks. The data exhibits a strong dependence on seasonality and displays an increasing trend. However, the outbreak of the COVID-19 pandemic caused a significant disruption in the pattern, as seen from record 260 onwards. Subsequently, there was a partial recovery in the data. This observation is further supported by Figure 2, which displays the seasonally adjusted data.

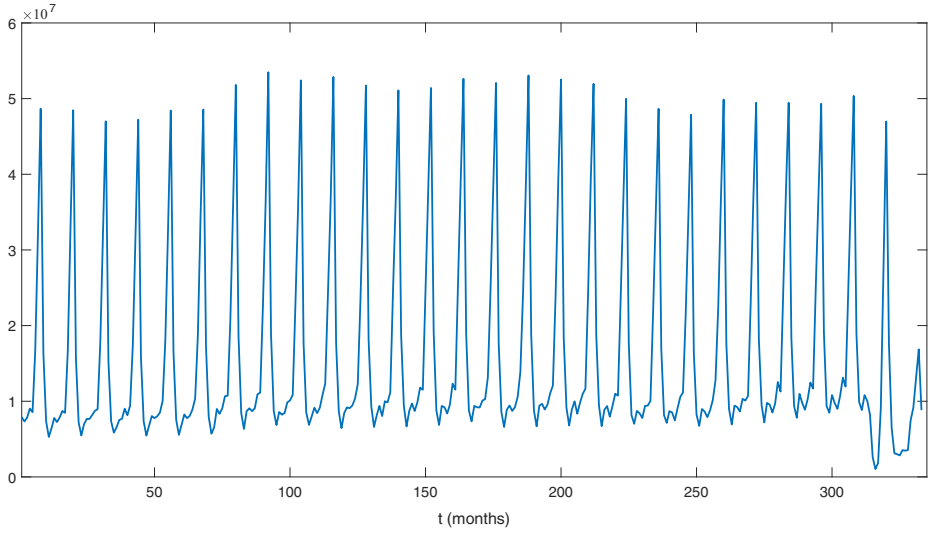
3.1 Statistical characteristics of data

Table 1 displays positive kurtosis and non-normality. The Kolmogorov–Smirnov test (normal distribution) and the Ljung–Box Q and Engle (ARCH) tests were conducted to evaluate the nights spent in Italy time series. The rejection decision, along with the corresponding *p*-value, was obtained from these tests. Table 2 confirms that data are not normally distributed, exhibit autocorrelation and heteroscedasticity. When a time series exhibits bias, positive kurtosis, nonnormality, autocorrelation and heteroscedasticity, it can be challenging to model and predict using traditional statistical models like linear regression. In such situations, models specifically designed to handle these characteristics, such as the CIR# model, would be more appropriate for fitting the data and producing accurate predictions. The analysis of Tables 2 and 1 shows that the time series of nights spent in Italy demonstrates these characteristics, making the CIR# model a suitable choice for accurately modeling and predicting the data.

4. Methodology

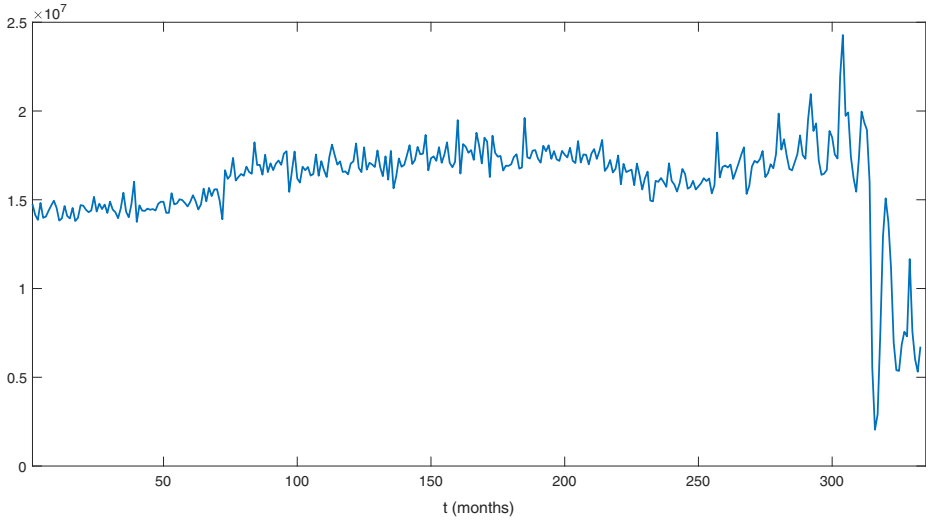
As mentioned, we aim at predicting with a parsimonious model overnight stays in Italy at tourist accommodation facilities when a regular pattern in the time series is disrupted. To this end, we have selected a nonlinear, single factor, stochastic model called CIR#, and we compare it to others ranging from simple EWMA to HW; from

Figure 1 Nights spent in Italy at tourist accommodation from January 1994 to September 2021



Note: Total number of observations = 332 (monthly data)
Source: Eurostat (2021)

Figure 2 Seasonally adjusted nights spent in Italy at tourist accommodation from January 1994 to September 2021



Source: Filter $S_{3\times5}$, see MathWorks (2021)

Table 1 Rejection decision *h* and *p*-value from the Kolmogorov–Smirnov (KS) test (normal distribution), the Ljung–Box Q (LBQ) test and the Engle (ARCH) test, carried out for the nights spent in Italy

Test	*KS*	*LBQ*	*ARCH*
h	1	1	1
p-value	1.0222×10^{-21}	0	1.0914×10^{-9}

Table 2 First four central moments for the changes in nights spent in Italy

Moments			
μ_1	μ_2	μ_3	μ_4
1.6117×10^7	1.3077	1.6235	4.4519

integrated ARIMA combined with EGARCH models, to seasonal integrated autoregressive moving average (SARIMA) and DNNAR models (see Appendix). As already mentioned, multiple regression is found to be effective for short-term forecasting on a daily and weekly basis, while the HW method is more suitable for monthly data and longer-term forecasting (Tratar and Strmčnik, 2016). Therefore, we have not included multiple regression from our set of baseline models. Finally, in Section 5, we also offer a comparison with other models used in the literature, including EC-ADLM, VAR, BVAR, TVP models and TFT model, optimized using an ADE as explored by Gunter and Önder (2015) and Wu *et al.* (2022), respectively.

4.1 The proposed model: the CIR#

The CIR process is a stochastic differential equation describing the evolution of the stochastic variable $V(t)$ as introduced by Cox *et al.* (1985).

$$dV(t) = k(\theta - V(t))dt + \sigma\sqrt{V(t)}\,dW(t), \tag{1}$$

where $V(0) = V_0 > 0$ is the initial condition and $W(t)$ a Wiener process. The CIR process is said to be a single-factor, time-homogeneous model, because σ, k and θ are time-independent, and $V(t)$ is the only variable. The CIR model has been devised for fixed income pricing, but in the context of this work, the stochastic variable is tourist accommodations described in Section 3.

The proposed CIR# preserves the structure of the original CIR model by Cox *et al.* (1985) and involves partitioning the market data into subsamples to capture statistically significant changes in variance and jumps. The second step involves fitting an "optimal" ARIMA model to each subsample of market data and applying Johnson's transformation (Johnson, 1949) to the standardized residuals to ensure they resemble Gaussian white noise. This is necessary because empirical excess returns of financial assets often have more kurtosis and positive serial correlation (Orlando and Bufalo, 2021b). For more details, a reader can refer to Orlando *et al.* (2019a, 2019b), Orlando and Bufalo (2021b, 2023).

4.2 Baseline models

Below is the description of some basic models to compare their performance with those of the CIR# model. Notice that, throughout this section, $(v_h)_{h\in[1,n]}$ are the observations related to the explanatory variable $V(t)$ over n periods.

4.2.1 Autoregressive integrated moving average exponential generalized autoregressive conditional heteroscedasticity model. An ARIMA(p, i, q) model is described by:

$$\left(1 - \sum_{j=1}^{p} \phi_j L^j\right)(1 - L^i)v_h = \left(1 + \sum_{j=1}^{q} \theta_j L^j\right)\varepsilon_h \qquad \left((p, i, q) \in \mathbb{N}_*^3\right), \tag{2}$$

where L is the lag operator, ϕ_j and θ_j indicate the parameters of the autoregressive part and of the moving average part of the model, respectively, and ε are white noise terms.

The EGARCH(a, b) can be expressed as:

$$\ln \sigma_h^2 = \omega + \sum_{j=1}^{1} \gamma_j \ln \sigma_{h-j}^2 + \sum_{j=1}^{b} \delta_j \left[\nu\tilde{\varepsilon}_{j-h} + \xi\left(|\tilde{\varepsilon}_{j-h}| - \mathbb{E}\left[|\tilde{\varepsilon}_{j-h}|\right]\right)\right] \qquad \left((a, b) \in \mathbb{N}_*^2\right), \tag{3}$$

where the variable $\tilde{\varepsilon}$ follows a generalized Gaussian distribution, σ^2 is the conditional variance process and ω, γ, δ, ν and ξ are real parameters. The ARIMA combined with EGARCH is suitable for time series data with heteroscedasticity, volatility clustering and conditional skewness or kurtosis. The ARIMA component models the autocorrelation and trend, while the EGARCH component models the volatility and asymmetry in the error terms. This combined model can provide more accurate forecasts and account for the nonlinear patterns and dynamics present in financial time series data.

4.2.2 Seasonal integrated autoregressive moving average model. The ARIMA model is an extension of the classical ARIMA, which is capable of modeling a wide range of seasonal data. SARIMA models are capable of capturing complex patterns in the data and can be used to forecast future values with a certain level of accuracy. However, they may not be suitable for nonstationary data or data with irregular patterns.

4.2.3 Exponentially weighted moving average model. The EWMA (see e.g. Perry, 2010) is a weighting scheme to simulate future values averaging on a historical data set. Holt (2004) suggested that the EWMA address trends and seasonality in forecasts. The EWMA is a means of smoothing out random fluctuations with some desirable properties such as:

- decrease in weight attributed to older data;
- ease of computation; and
- minimum data requirement.

4.2.4 Holt–Winters model. The HW model is a generalization of the EWMA model in presence of seasonal data (Chatfield, 1978). There are two versions of such a method, the so-called additive version and the multiplicative one. The additive method is best suited when seasonal variations are roughly constant across the time series, as observed in our data set (see Subsection 3). The multiplicative method is preferable when seasonal variations change proportionally to the level of the time series. The method is useful when the data has a strong seasonal pattern that repeats over time. It can also handle trend and level changes in the time series. However, the method assumes that the time series is stationary, which means that its statistical properties do not change over time. It may not be suitable for nonstationary data with trends and seasonality changes.

4.2.5 Neural network autoregression model. A NN is a network of neurons structured in layers where the predictions (outputs) form the upper layer and the predictors (inputs) form the lower layer. Often there are also some layers of neurons between predictors and predictions and which are called hidden layers. This configuration is called a multilayer feed-forward network.

As explained in Silva *et al.* (2019), the NNAR model has to be implemented jointly with the SSA to denoise data. SSA is a nonparametric estimation method where the covariance matrix is decomposed into a spectrum of eigenvalues. The resulting (denoised) time series is then used as the input for a NNAR model. This hybrid approach is called DNNAR, and in the remainder of this work, the "optimal" DNNAR model is denoted by DNNAR*. The details of the performed SSA are available in the Appendix. The DNNAR model is particularly useful for time series with high volatility and nonlinearity, where traditional statistical models may struggle to provide accurate forecasts.

4.3 Measures of accuracy

Here, we list the forecast accuracy measures used to compare the results between the considered models:

- The RMSE measures the closeness between the observed data and their predictions.
- The MAPE is another measure of the prediction accuracy of a forecasting model.
- The AvgRelMSE compares two different forecasts.

The AvgRelMSE is symmetric to over and underforecasting (Davydenko and Fildes, 2013) and is used in the literature to test different scenarios (Kourentzes and Athanasopoulos, 2019). As a reference, a given model A is more accurate than the selected benchmark B when the AvgRelMSE is smaller than 1. In particular, the considered model is better than the benchmark by (1-AvgRelMSE)100%. A variant of the AvgRelMSE is the AvgRelRMSE where the RMSE is taken instead of the MSE.

5. Results

5.1 Out-of-sample results on the full data set

In this section, we show the out-of-sample performance of the proposed model against the results obtained with the five baseline models listed in Section 4.2. Concerning the CIR# model, to predict the future value, first, we calibrate the parameters (k, θ, σ, p, i, q) through a rolling window W of length $M = 12$ so that $W = \{v_h, \ldots, v_{h+M-1}(h \geq 1)$. Then, the forecasted values $v^F_{h+M+s}\,(s \geq 0)$, are determined with the procedure explained in

Orlando *et al.* (2019b). Similarly, for the baseline models, we first calibrate the parameters on the same window and then we compute the next value (i.e. the forecast for the next month). In this regard, Figure 3 shows the out-of-sample forecasts versus actual data, while Figure 4 displays the error. As can be seen:

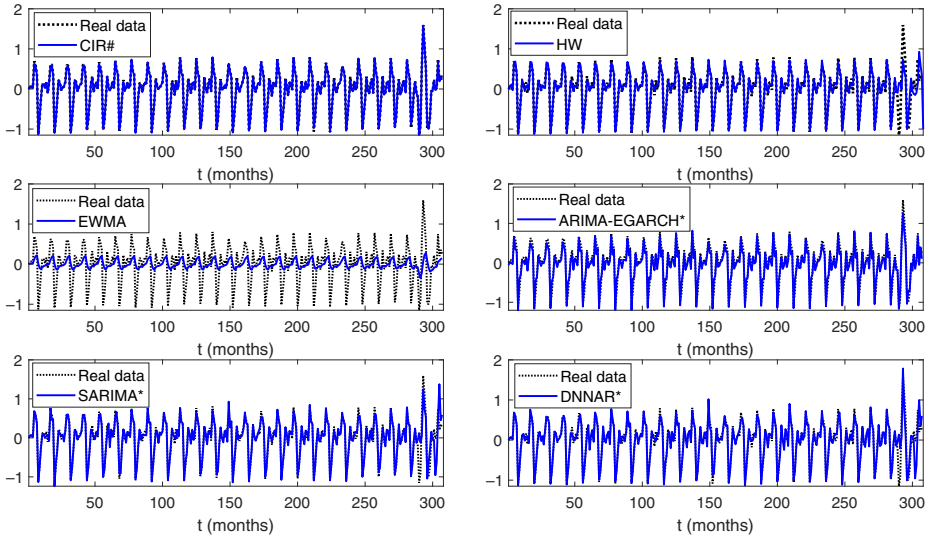
- the EWMA is not a suitable choice for this type of data;
- the SARIMA, HW and DNNAR models are quite good until the COVID-19 pandemic where they perform poorly; and
- CIR# and ARIMA-EGARCH seem to perform better even during the pandemic.

In particular, there are two jumps from record 260 onward that are captured by CIR# and ARIMA-EGARCH models.

To better appreciate the performance of the CIR# versus the baseline models, Table 3 reports the results of the measures listed in Section 4.3. In agreement with the above graphical analysis, the CIR# appears to be the best model in terms of all measures (MAE, RMSE and MAPE). Concerning the literature, we refer to Gunter and Önder (2015) who used multiple benchmarks such as EC-ADLM, VAR, Bayesian VAR, TVP models (multivariate or econometric models), ARMA and the ETS models (univariate or time-series models). In their tests, Gunter and Önder (2015) reported that the minimum prediction error (RMSE and MAE) of the considered models over a month horizon is about 7% and that the average is 10% (i.e. at least 3.5 times higher). Other models, like the TFT that uses an ADE, achieve a best MAPE of 3.02% (Wu *et al.*, 2022), which is twice as high as that of CIR#.

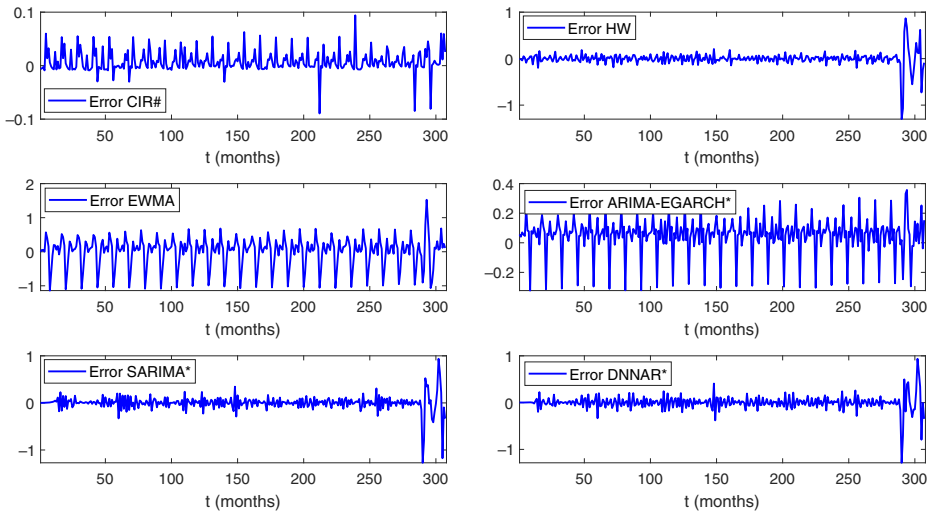
Finally, Table 4 shows the high accuracy in terms of AvgRelMSE (as well as the AvgRelRMSE) of the CIR# model with respect to the selected baseline models. For reference, Kourentzes and Athanasopoulos (2019), when comparing the forecasts for Australian tourism against the

Figure 3 From the top left to the right, first row: (a) real data vs CIR# and (b) real data vs HW model; from the top left to the right, second row: (a) real data vs EWMA and (b) real data vs ARIMA-EGARCH*; from the top left to the right, third row: (a) real data vs SARIMA* and (b) real data vs DNNAR*



Notes: Out-of-sample results over monthly data (changes). Note that there are two jumps from record 260 onwards that are captured by CIR# and ARIMA-EGARCH models

Figure 4 From the top left to the right, first row: (a) error produced by CIR# and (b) error produced by HW model; from the top left to the right, second row: (a) forecasting error with the EWMA and (b) forecasting error with the ARIMA-EGARCH*; from the top left to the right, third row: (a) forecasting error with the SARIMA* and (b) forecasting error with the DNNAR*



Note: Out-of-sample results over monthly data (changes)

Table 3 Forecasting accuracy across models

Statistics	*CIR# (%)*	*HW (%)*	*EWMA (%)*	*ARIMA-EGARCH* (%)*	*SARIMA* (%)*	*DNNAR* (%)*
MAE	1.32	36.84	31.47	10.25	9.21	9.49
RMSE	2.12	15.56	43.90	13.44	16.74	16.20
MAPE	1.18	7.14	8.38	12.07	10.93	11.18

Note: Out-of-sample results over monthly data (changes)

Table 4 Accuracy of the CIR# in terms of AvgRelMSE and AvgRelRMSE with respect to the baseline models

Statistics	*Baseline models*				
	HW	*EWMA*	*ARIMA-EGARCH**	*SARIMA**	*DNNAR**
AvgRelMSE	0.0681	0.0021	0.0217	0.0467	0.0407
(1-AvgRelMSE)100%	93.18%	99.79%	97.82%	95.33%	95.92%
AvgRelRMSE	0.2610	0.0453	0.1473	0.2161	0.2018
(1-AvgRelRMSE)100%	73.90%	95.47%	85.27%	78.39%	79.82%

ETS and the ARIMA model under different scenarios, obtained a minimum value of 0.916. In our case, the AvgRelMSE corresponding to the EWMA and the ARIMA-EGARCH are 0.0217 and 0.0467, respectively. Similarly, Di Fonzo and Girolimetto (2022), when performing a test on the accuracy of the proposed forecast combination-based approach, found an AvgRelMSE equal to 0.9618 (i.e. much higher than any value reported in Table 4).

5.2 Out-of-sample results amid COVID-19

In the previous section, we have shown the out-of-sample performance on the full data set of the proposed model against the results obtained with the five baseline models and

reported prediction errors in other studies of the literature. In this section, we focus on the part of the data set where the data pattern was disrupted because of COVID-19 pandemic. Figure 5 presents the out-of-sample predictions compared to the actual data, whereas Figure 6 illustrates the corresponding errors. As previously seen in Section 5.1 during the COVID-19 pandemic, it can be observed that:

- The EWMA model is not appropriate for this type of data.
- The SARIMA, HW and DNNAR models perform well until the data pattern is significantly disrupted.
- The CIR# and ARIMA-EGARCH models demonstrate superior performance in any condition, with the former outperforming the latter.

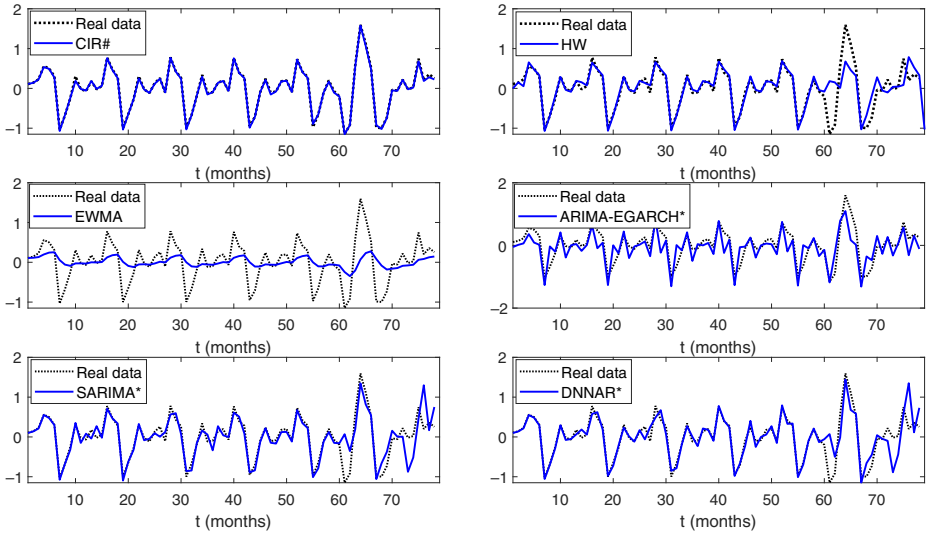
Table 5 presents the results of the measures listed in Section 4.3 to provide a better understanding of the performance of the CIR# model compared to the baseline models. As confirmed by the graphical analysis discussed earlier, the CIR# model outperforms all other models in terms of all measures (MAE, RMSE and MAPE). It is worth noting that these results are superior to those reported in other studies such as Gunter and Önder (2015) and Wu *et al.* (2022).

Finally, Table 6 shows the high accuracy in terms of AvgRelMSE (as well as the AvgRelRMSE) of the CIR# model with respect to the selected baseline models which, once again, are better than those found in the mentioned literature (Kourentzes and Athanasopoulos, 2019; Di Fonzo and Girolimetto, 2022).

5.3 Model selection and distribution of forecast errors

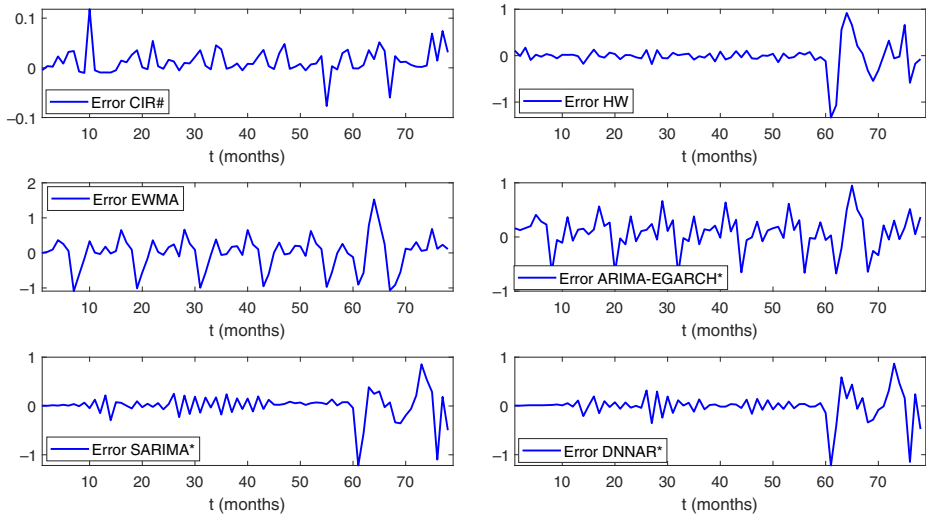
The analysis presented above indicates that the CIR# model consistently has a small error compared to other reference predictions. However, it is important to determine whether this

Figure 5 From the top left to the right, first row: (a) real data vs CIR# and (b) real data vs HW model; from the top left to the right, second row: (a) real data vs EWMA and (b) real data vs ARIMA-EGARCH*; from the top left to the right, third row: (a) real data vs SARIMA* and (b) real data vs DNNAR*



Note: Out-of-sample results over monthly data (changes) during the COVID-19 period

Figure 6 From the top left to the right, first row: (a) error produced by CIR# and (b) error produced by HW model; from the top left to the right, second row: (a) forecasting error with the EWMA and (b) forecasting error with the ARIMA-EGARCH*; from the top left to the right, third row: (a) forecasting error with the SARIMA* and (b) forecasting error with the DNNAR*



Note: Out-of-sample results over monthly data (changes) during the COVID-19 period

Table 5 Forecasting accuracy across models

Statistics	*CIR# (%)*	*HW (%)*	*EWMA (%)*	*ARIMA-EGARCH* (%)*	*SARIMA* (%)*	*DNNAR* (%)*
MAE	1.88	40.24	33.49	25.47	16.50	15.65
RMSE	2.85	28.44	47.68	33.43	27.35	2.56
MAPE	1.69	5.47	7.05	13.86	14.08	12.15

Note: Out-of-sample results over monthly data (changes) during the COVID-19 period

Table 6 Accuracy of the CIR# in terms of AvgRelMSE and AvgRelRMSE with respect to the baseline models, during the COVID-19 period

Statistics	*Baseline models*				
	HW	*EWMA*	*ARIMA-EGARCH**	*SARIMA**	*DNNAR**
AvgRelMSE	0.0432	0.0029	0.0057	0.0239	0.0280
(1-AvgRelMSE)100%	95.67%	99.70%	99.42%	97.60%	97.19%
AvgRelRMSE	0.2080	0.0539	0.0757	0.1548	0.1674
(1-AvgRelRMSE)100%	79.19%	94.60%	92.42%	84.51%	83.25%

difference is statistically significant or if it could be attributed to the specific sample of data used. Table 7 demonstrates that the differences between the CIR# forecasts and the reference predictions are statistically significant. Furthermore, Table 8 shows that, except for the SARIMA model, all the models produce homoscedastic errors indicating that the variance of errors is constant across all levels of the independent variable. This allows for more accurate statistical modeling and estimation of parameters, as well as valid inference and hypothesis testing. However, it is worth noting that the forecasting errors of all the models are not normally distributed.

Table 7 *p*-Value of the Diebold–Mariano (DM) test for assessing the different nature for the CIR# forecasts vs other benchmark predictions

			Baseline models		
Test	*HW*	*EWMA*	*ARIMA-EGARCH**	*SARIMA**	*DNNAR**
DM *p*-value	0.0020	6.5277×10^{-21}	1.5928×10^{-23}	5.8997×10^{-4}	2.0884×10^{-4}

Table 8 *p*-Value of the heteroscedasticity test (white) and KS test (normal distribution) of the prediction errors

				Models		
Test	*CIR#*	*HW*	*EWMA*	*ARIMA-EGARCH**	*SARIMA**	*DNNAR**
White *p*-value	0.1841	0.2524	0.2066	0.0921	4.5325×10^{-4}	0.6716
KS *p*-value	2.5537×10^{-9}	4.7863×10^{-10}	1.0881×10^{-9}	1.3574×10^{-9}	8.7524×10^{-7}	1.3455×10^{-5}

6. Discussion and implications of the research

Analysis of data shows that the nights spent in Italy time series have a positive kurtosis and exhibit nonnormality. When a time series presents bias, positive kurtosis, nonnormality, autocorrelation and heteroscedasticity, it can be challenging to model and predict using conventional statistical models like linear regression. In such cases, models specifically developed to handle these characteristics, such as the CIR# model, would be more appropriate for fitting the data. Tables 1 and 2 confirm that the nights spent in Italy time series exhibit these characteristics, reinforcing the need for models that can handle such challenges. Therefore, it is not a case that results in Section 5 prove that the CIR# model performs better than the baseline models in terms of all measures. Furthermore, the carried analysis suggests that the CIR# model outperforms other reference predictions in terms of error the statistical significance of this difference. The proposed model has demonstrated superiority even when compared to other models in the literature and can be especially useful for tourism stakeholders in making decisions when there are disruptions in data patterns. Although all models, except for SARIMA, produce homoscedastic errors, it is worth noting that the forecasting errors of all models are not normally distributed. This assumption is important for statistical modeling, and violating it can lead to biased estimates, invalid inferences and inaccurate predictions. Such a result is not unexpected because, by design, the CIR# model is built in a way that standardized residuals resemble Gaussian white noise (Orlando *et al.*, 2019a, 2019b).

In summary, the CIR# model offers a valuable addition to the body of knowledge regarding tourism demand forecasting, and its use could lead to more accurate predictions and better-informed decisions for stakeholders in the tourism industry. In fact, accurate forecasting of tourist demand is invaluable for researchers, industry workers and decision-makers as it helps mitigate the challenges posed by excessive unexpected demand during peak periods and the underutilization of capacity during low-demand periods.

7. Conclusion

In this article, we have explained the relevance of tourism both in economic and social terms. Then, we have presented the case of Italy and we have walked through relevant literature for forecasting tourism demand. As mentioned by Song *et al.* (2019), the evolution of forecasting methods is still ongoing. All methods have their limitations, and there is no single method outperforming the others (Law *et al.*, 2019). Moreover, because there is no free lunch (Wolpert, 1996; Wolpert and Macready, 1997), there is a trade-off between complexity and accuracy. Ever-increasing sophisticated models may suffer from model risk such as incorrect specification, wrong implementation, lack of sufficient data and calibration errors. The CIR#

model is parsimonious as it requires a single time series, and its forecasting power is tested against numerous benchmarks common in the literature. The result is that, for the case of Italy at the time of the COVID-19 pandemic, the proposed stochastic approach compares well against all the other considered baseline models providing a forecast error reduced by 70%. The same applies when comparing the obtained results to those available in the literature (e.g. see Gunter and Önder, 2015; Kourentzes and Athanasopoulos, 2019; Di Fonzo and Girolimetto, 2022; Wu *et al.*, 2022).

The proposed model stands out for being a viable option to the HW model, particularly when dealing with irregular data. In addition, the CIR# model is notably simpler than other advanced models found in literature and does not rely on black box techniques such as those used in NN or data science-based models. In this sense, we have presented a simple but not simplistic model that may be added to the range of tools for predicting tourism volumes. Due to the weight of tourism in Italy, this has important implications in terms of development policies and regulations. Next research will explore the suitability of the CIR# model to forecasting tourism time series in other countries.

The limitations of the proposed approach arise when the underlying data of the tourism time series do not exhibit significant disruptions. In such cases, the CIR# model may not provide any significant advantages over conventional models such as the HW. Therefore, when interpreting the results, it is crucial to consider the specific characteristics of the data, the context of the research and the timing of the analysis. These factors can influence the suitability and performance of the hypothesized models, offering insights into the limitations and applicability of the findings.

References

Abdulmajeed, K., Adeleke, M. and Popoola, L. (2020), "Online forecasting of COVID-19 cases in Nigeria using limited data", *Data in Brief*, Vol. 30, p. 105683.

Burini, F. (2020), *Tourism Facing a Pandemic: From Crisis to Recovery*, Università degli studi di Bergamo, Bergamo.

Chan, F., Lim, C. and McAleer, M. (2005), "Modelling multivariate international tourism demand and volatility", *Tourism Management*, Vol. 26 No. 3, pp. 459-471.

Chatfield, C. (1978), "The Holt-Winters forecasting procedure", *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol. 27 No. 3, pp. 264-279.

Coshall, J.T. (2009), "Combining volatility and smoothing forecasts of UK demand for international tourism", *Tourism Management*, Vol. 30 No. 4, pp. 495-511.

Cox, J.C., Ingersoll, J.E. and Ross, S.A. (1985), "A theory of the term structure of interest rates", *Econometrica*, Vol. 53 No. 2, pp. 385-407.

Davydenko, A. and Fildes, R. (2013), "Measuring forecasting accuracy: the case of judgmental adjustments to SKU-level demand forecasts", *International Journal of Forecasting*, Vol. 29 No. 3, pp. 510-522.

Di Fonzo, T. and Girolimetto, D. (2022), "Forecast combination-based forecast reconciliation: insights and extensions", *International Journal of Forecasting*, Vol. 1.

Eurostat (2021), "Nights spent at tourist accommodation establishments – monthly data", available at: https://ec.europa.eu/eurostat/en/web/products-datasets/-/TOUR_OCC_NIM (accessed 20 December 2021).

Frechtling, D. (2012), *Forecasting Tourism Demand*, Routledge.

Gunter, U. and Önder, I. (2015), "Forecasting international city tourism demand for Paris: accuracy of uni- and multivariate models employing monthly data", *Tourism Management*, Vol. 46, pp. 123-135.

Guojun, H. and Ningning, L. (2021), "Prediction of ecotourism population based on exponential smoothing and ARIMA mixed model", *International Conference on Intelligent Computing, Automation and Systems (ICICAS)*, IEEE, pp. 39-42.

Holt, C.C. (2004), "Forecasting seasonals and trends by exponentially weighted moving averages", *International Journal of Forecasting*, Vol. 20 No. 1, pp. 5-10.

Hu, M., Qiu, R.T.R., Wu, D.C. and Song, H. (2021), "Hierarchical pattern recognition for tourism demand forecasting", *Tourism Management*, Vol. 84, p. 104263.

Johnson, N.L. (1949), "Systems of frequency curves generated by methods of translation", *Biometrika*, Vol. 36 Nos 1/2, pp. 149-176.

Kourentzes, N. and Athanasopoulos, G. (2019), "Cross-temporal coherent forecasts for Australian tourism", *Annals of Tourism Research*, Vol. 75, pp. 393-409.

Law, R., Li, G., Fong, D.K.C. and Han, X. (2019), "Tourism demand forecasting: a deep learning approach", *Annals of Tourism Research*, Vol. 75, pp. 410-423.

Marino, A. and Pariso, P. (2021), "E-tourism: how ICTs help the local tourist district drive economic vitality. The case of Campania, Italy", *International Journal of Innovation and Technology Management*, Vol. 18 No. 3, p. 2150009.

Martins, L.F., Gan, Y. and Ferreira-Lopes, A. (2017), "An empirical analysis of the influence of macroeconomic determinants on world tourism demand", *Tourism Management*, Vol. 61, pp. 248-260.

MathWorks (2021), "Seasonal adjustment using S(n,m) seasonal filters", available at: https://it.mathworks.com/help/econ/seasonal-adjustment-using-snxd7m-seasonal-filters.html (accessed 10 October 2022).

Meleddu, M. and Pulina, M. (2016), "Evaluation of individuals' intention to pay a premium price for ecotourism: an exploratory study", *Journal of Behavioral and Experimental Economics*, Vol. 65, pp. 67-78.

Orlando, G. and Bufalo, M. (2021a), "Interest rates forecasting: between Hull and White and the CIR#—how to make a single-factor model work", *Journal of Forecasting*, Vol. 40 No. 8, pp. 1566-1580.

Orlando, G. and Bufalo, M. (2021b), "Empirical evidences on the interconnectedness between sampling and asset returns' distributions", *Risks*, Vol. 9 No. 5, p. 88, doi: 10.3390/risks9050088.

Orlando, G. and Bufalo, M. (2023), "Time series forecasting with the CIR# model: from hectic markets sentiments to regular seasonal tourism. 1", Vol. 29 No. 4, pp. 1216-1238, doi: 10.3846/tede.2023.19294.

Orlando, G., Mininni, R.M. and Bufalo, M. (2018), "A new approach to CIR short-term rates modelling", in Mili, M., Medina Samaniego, R. and Filippo, D.P. (Eds), *New Methods in Fixed Income Modeling – Fixed Income Modeling*, Springer International, New York, NY, pp. 35-44.

Orlando, G., Mininni, R.M. and Bufalo, M. (2019a), "Interest rates calibration with a CIR model", *The Journal of Risk Finance*, Vol. 20 No. 4, pp. 370-387.

Orlando, G., Mininni, R.M. and Bufalo, M. (2019b), "A new approach to forecast market interest rates through the CIR model", *Studies in Economics and Finance*, Vol. 37 No. 2, pp. 267-292.

Orlando, G., Mininni, R.M. and Bufalo, M. (2020), "Forecasting interest rates through Vasicek and CIR models: a partitioning approach", *Journal of Forecasting*, Vol. 39 No. 4, pp. 569-579.

Park, S., Lee, J. and Song, W. (2017), "Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data", *Journal of Travel & Tourism Marketing*, Vol. 34 No. 3, pp. 357-368.

Perry, M.B. (2010), "The exponentially weighted moving average", *Wiley Encyclopedia of Operations Research and Management Science*, Wiley, New York, NY.

Poprawe, M. (2015), "A panel data analysis of the effect of corruption on tourism", *Applied Economics*, Vol. 47 No. 23, pp. 2399-2412.

Rosy, C.P. and Ponnusamy, R. (2017), "Evaluating and forecasting room demand in tourist spot using Holt-Winters method", *International Journal of Computer Applications*, Vol. 975, p. 8887.

Saha, S. and Yap, G. (2015), "Corruption and tourism: an empirical investigation in a non-linear framework", *International Journal of Tourism Research*, Vol. 17 No. 3, pp. 272-281.

Silva, E.S., Hassani, H., Heravi, S. and Huang, X. (2019), "Forecasting tourism demand with denoised neural networks", *Annals of Tourism Research*, Vol. 74, pp. 134-154.

Song, H. and Witt, S.F. (2012), *Tourism Demand Modelling and Forecasting*, Routledge, New York, NY.

Song, H., Qiu, R.T. and Park, J. (2019), "A review of research on tourism demand forecasting: launching the annals of tourism research curated collection on tourism demand forecasting", *Annals of Tourism Research*, Vol. 75, pp. 338-362.

Tratar, L.F. and Strmčnik, E. (2016), "The comparison of Holt–Winters method and multiple regression method: a case study", *Energy*, Vol. 109, pp. 266-276.

UNWTO (2020), "UNWTO world tourism barometer and statistical annex, December 2020", *UNWTO World Tourism Barometer (English Version)*, Vol. 18 No. 7, pp. 1-36.

UNWTO (2021a), "UNWTO world tourism barometer and statistical annex, November 2021", *UNWTO World Tourism Barometer (English Version)*, Vol. 19 No. 6, pp. 1-36.

UNWTO (2021b), "2020: a year in review", available at: www.unwto.org/covid-19-and-tourism-2020 (accessed 10 October 2021).

Winters, P.R. (1960), "Forecasting sales by exponentially weighted moving averages", *Management Science*, Vol. 6 No. 3, pp. 324-342.

Wolpert, D.H. (1996), "The lack of a priori distinctions between learning algorithms", *Neural Computation*, Vol. 8 No. 7, pp. 1341-1390.

Wolpert, D.H. and Macready, W.G. (1997), "No free lunch theorems for optimization", *IEEE Transactions on Evolutionary Computation*, Vol. 1 No. 1, pp. 67-82.

Wu, B., Wang, L. and Zeng, Y.-R. (2022), "Interpretable tourism demand forecasting with temporal fusion transformers amid COVID-19", *Applied Intelligence*, Vol. 53 No. 11, pp. 1-22.

Xie, G., Qian, Y. and Wang, S. (2020), "A decomposition-ensemble approach for tourism forecasting", *Annals of Tourism Research*, Vol. 81, p. 102891.

Yao, Y. and Cao, Y. (2020), "A neural network enhanced hidden Markov model for tourism demand forecasting", *Applied Soft Computing*, Vol. 94, p. 106465.

Yonar, H., Yonar, A., Tekindal, M.A. and Tekindal, M. (2020), "Modeling and forecasting for the number of cases of the COVID-19 pandemic with the curve estimation models, the box-Jenkins and exponential smoothing methods", *Eurasian Journal of Medicine and Oncology*, Vol. 4 No. 2, pp. 160-165.

Appendix. Singular spectrum analysis

In this appendix, we show the details of the singular spectrum analysis procedure, applied to our time series to denoise data as suggested by Silva *et al.* (2019):

- First of all, we have to normalize our time series, so we introduce the series $z_h = \frac{v_h - \hat{\mu}}{\hat{\sigma}}$, where $\hat{\mu}$, $\hat{\sigma}$ are the sample mean and sample standard deviation of the observations $(v_h)_{h\in[1,n]}$, respectively.
- Then, we compute the covariance matrix C (see Figure A1). To do this, we determine C by the following scalar product:

$$C = \frac{Y \cdot Y}{n - s + 1},$$

where the matrix Y is the time-delayed embedding of $(z_h)_{h\in[1,n]}$, i.e.:

$$Y_{h,k} = z_{h+s-1} \qquad (h \in [1, n - s + 1],\ k \in [1, s]).$$

Notice that, even though the estimated C matrix does not have a Toeplitz structure (with nonsymmetric or antisymmetric eigenvectors), it at least guarantees that C is positive semidefinite:

- Figure A2 shows the eigenvalues and eigenvectors of matrix C.
- Figure A3 shows the principal components as obtained by the scalar product between Y, the time-delayed embedding of (zh)h∈[1,n], and the eigenvectors P.
- Figure A4 shows the reconstructed components, as obtained by inverting the projecting PC = Y s P.
- The upper panel of Figure A5 shows the original time series $(vh)_{h\in[1,n]}$ as obtained by the sum of all reconstructed components. The lower panel of Figure A5 displays the denoised time series obtained with the first pair of reconstructed components.

Figure A1 Covariance matrix

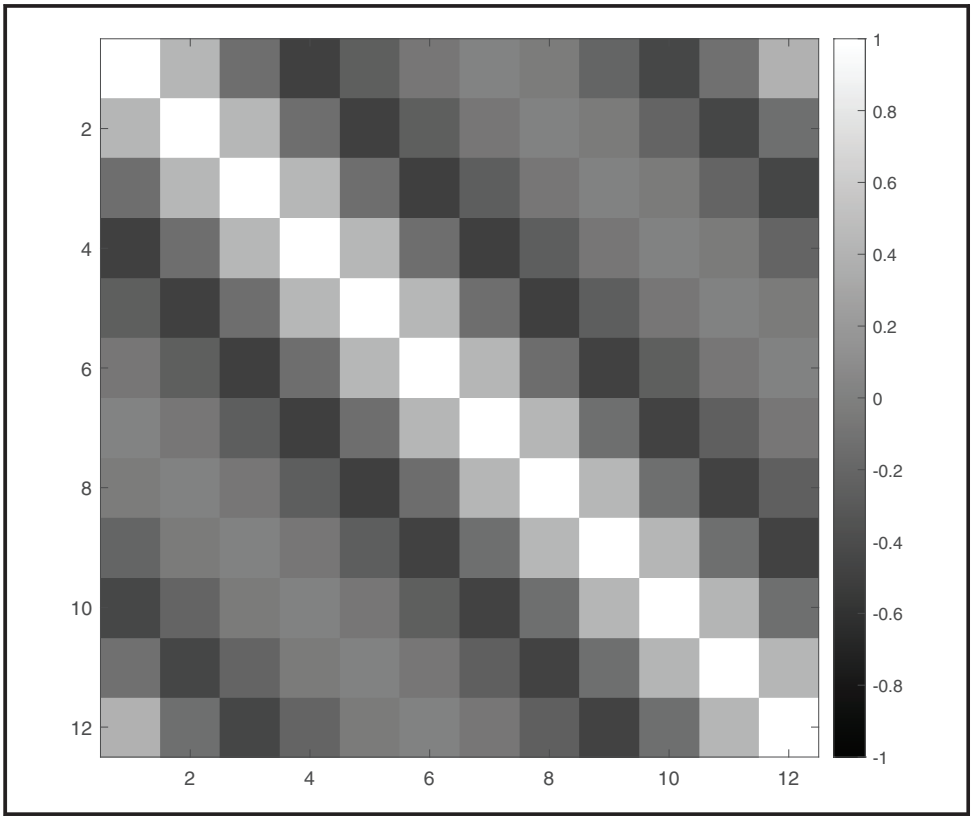


Figure A2 First plot: eigenvalues; second plot: first and second eigenvectors; and third plot: third and fourth eigenvectors

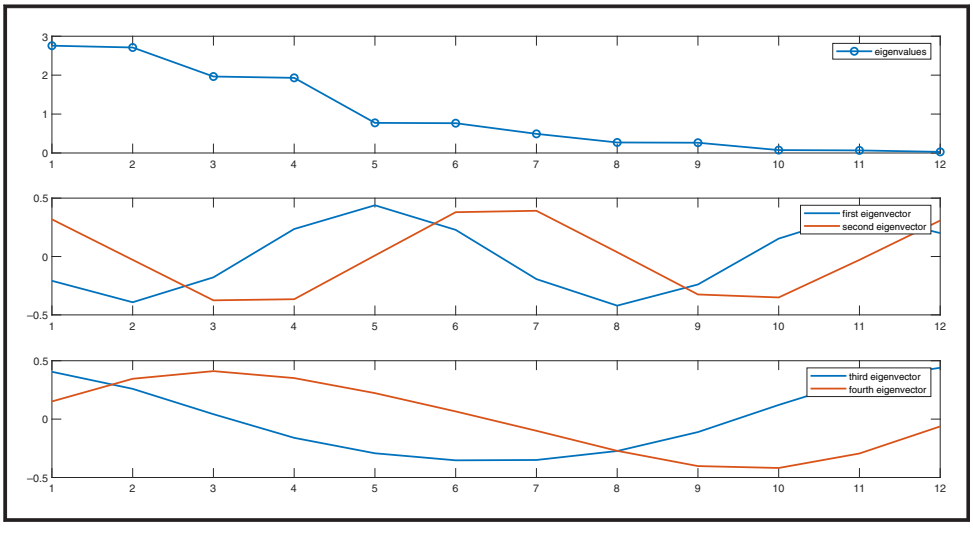


Figure A3 PCs components

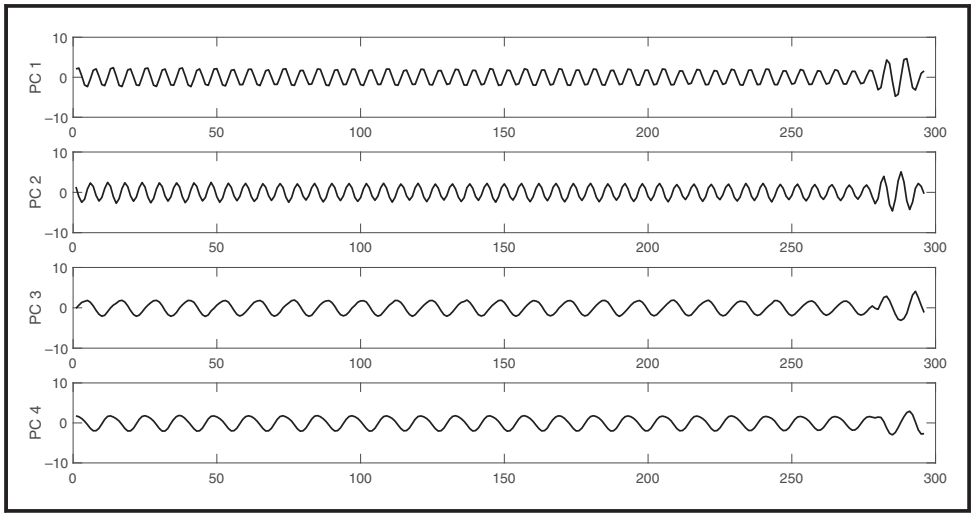


Figure A4 Reconstructed PCs components

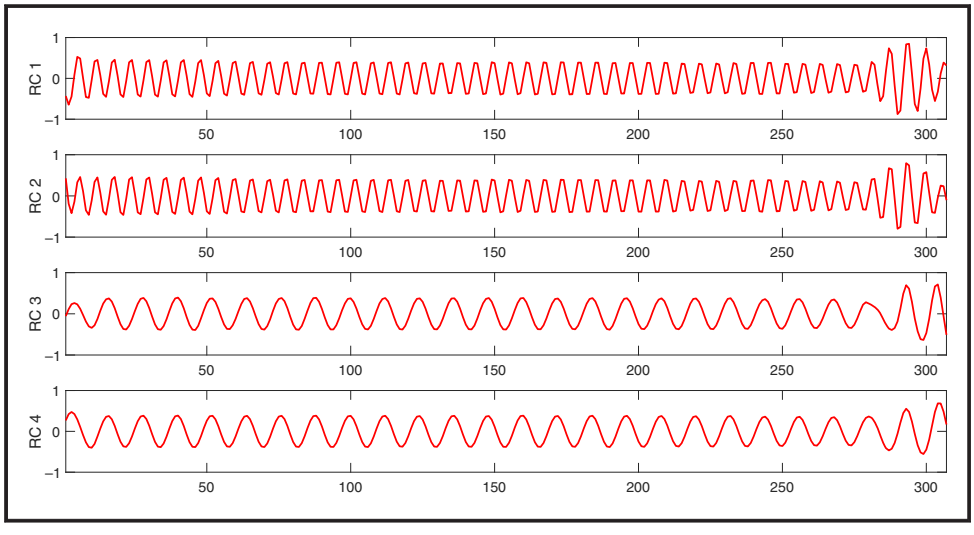
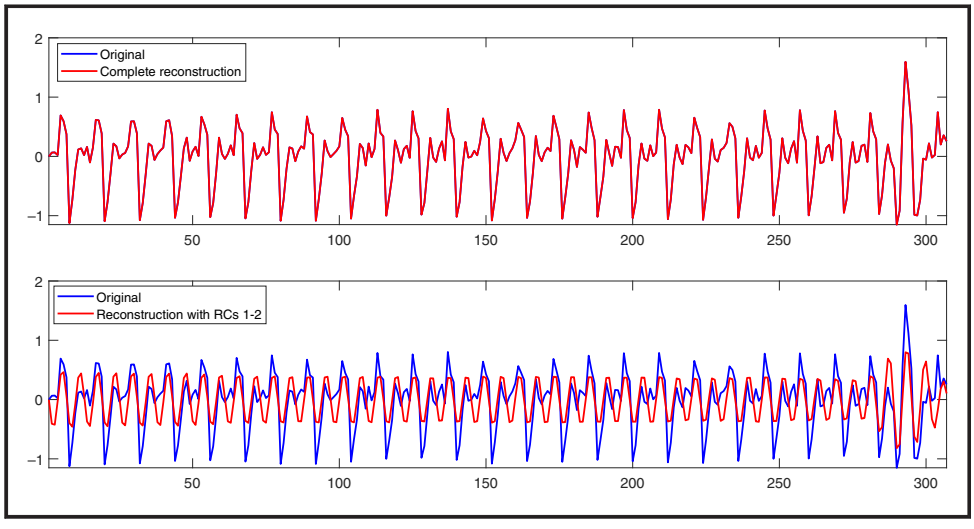


Figure A5 Upper graph: original vs reconstructed signal; and lower graph: original vs denoised signal



Corresponding author

Giuseppe Orlando can be contacted at: giuseppe.orlando@uniba.it