A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

FORECASTING AND ECONOMIC IMPLICATIONS OF TOURIST ARRIVALS TO SOUTH AFRICA

Miguel Gomes

Work project carried out under the supervision of:

Leid Zejnilovic

Lénia Mestrinho

20/12/2023

Table of Contents

1. Abstract

This research presents an analysis and forecast of tourist arrivals to South Africa and their economic implications. It explores the reciprocal role of tourism as a driver of economic development and an outcome of the country's progress.

Miguel's individual part focuses on building time series forecasting models (SARIMA and SARIMAX) to accurately predict monthly tourist arrivals to South Africa.

Hugo's part then moves to quantify the potential economic impacts of these predictions, analyzing them against current governmental practices.

Keywords (South Africa, Tourism Forecasting, Economic Analysis, SARIMA, SARIMAX)

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

2. Introduction

Tourism is not just a leisure activity; it is a vital economic catalyst. Globally, the sector shapes economies, influences social and cultural landscapes and plays a crucial role in environmental conservation and sustainability. The economic impact of tourism is extensive, stimulating job creation, driving infrastructure development, and fostering international trade. This multifaceted sector not only boosts economic growth but also cultivates cultural exchange and understanding, making it an essential area of study and development, especially in diverse and evolving regions like the African continent. Also, through responsible tourism practices, which minimize negative environmental and social impacts, promotes the preservation of natural habitats and cultural heritage.

This paper emerged from a collaboration with the WiTH Africa project, a partnership between Nova SBE's Westmont Institute of Tourism and Hospitality (WiTH) and Nova SBE's Data Science Knowledge Center. WiTH Africa aims to create tools and share knowledge on the tourism industry in Africa through data-driven approaches. As students of the Business Analytics Masters programme, we recognized that collaborating with WiTH Africa was an excellent opportunity to apply our academic learnings to a real-world context in a practical and insightful way.

The initial proposal for this work-project was to explore the evolution of tourism in Africa in the context of a post-pandemic world. However, to ensure a more thorough and detailed analysis, we narrowed our study to a specific country. South Africa was selected as the focus of this research, primarily due to the availability of structured and regularly updated tourism-related data. The tourism sector's significance in South Africa further influenced this decision. Since their first democratic elections in 1994, South Africa has made significant investments in

the tourism industry, recognizing it as an essential driver for economic growth and establishing itself as a tourism-dependent nation (Lundahl and Petersson 2009).

Given the importance of the sector to the country's development, this paper centers around forecasting international tourism demand in South Africa and evaluating the sector's economic implications, taking into account the evolving dynamics of the tourism industry. We examine different time series models to accurately forecast tourist arrivals to South Africa and incorporate exogenous factors that could improve predictive accuracy. We then assess the viability of using the forecasted tourist arrivals to predict related economic variables and examine how these can impact the country's economy. These forecasts can serve as valuable tools for the South African government and policymakers, possibly assisting in formulating effective tourism policies, efficient allocation of resources, and strategic infrastructure planning.

3. Literature Review

A literature review was conducted to further understand the extensive role of tourism as an economic catalyst. It delves into the multiple dimensions of how tourism impacts economies around the world and specifically in South Africa. It also examines the complexities of forecasting tourism demand and explores the stakeholders involved in this sector.

3.1. Tourism as an Economic Driver

Tourism is a significant driving force of many countries' economies worldwide, impacting them both directly and indirectly. According to the World Travel & Tourism Council (WTTC),

before the COVID-19 pandemic in 2019, travel and tourism accounted for 10.4% of the global Gross Domestic Product (GDP) at US$9.2 trillion.

As the leading economic benefits, tourism generates significant revenues through travelers' spending on accommodation, food, transportation, attractions, and shopping. Also, growing tourism has the capability of directly tackling unemployment due to its labor-intensive nature. This growth not only fosters job opportunities in the service sector, but also stimulates development in related fields, such as agriculture, manufacturing, and retail. Due to this multiplier effect, the United Nations World Tourism Organization (UNTWO) estimates that one job in the core tourism sector creates about one-and-a-half additional or indirect jobs in the tourism-related economy. Remarkably, in 2019, tourism collectively accounted for a sizable portion of employment, constituting approximately one in ten jobs worldwide (UNWTO 2021). Furthermore, the sector requires considerable investment in infrastructure. Airports, roads, hotels, resorts, and entertainment facilities are essential areas that witness substantial investments. Governments also directly benefit from tourism through various avenues, including taxes, entry fees, permits, and levies imposed on tourists and tourism-related businesses. The resulting revenues can be directed towards public services and investment, amplifying the sector's positive influence (Tisdell and Wilson 2013).

Beyond the direct impacts, tourism sets off other collateral effects, generating indirect economic contributions. For instance, tourism fosters economic diversity and entrepreneurship through demand for local crafts, souvenirs, and restaurants, promoting the growth of small businesses. Cultural preservation is also encouraged as way to attract visitors, indirectly contributing to the protection of a country's cultural identity and history. This multifaceted influence of tourism showcases its role as a powerful economic driver with the potential to shape diverse aspects of

a nation's economic landscape, making it a critical area for policy focus and strategic development (Lemma, 2014).

3.2. South Africa's Economy and the Significance of Tourism

The economic landscape of South Africa is often described as diverse and unevenly developed, where advanced industrial sectors contrast with underdeveloped rural areas. Baffi, Turok, and Marcuzzo (2018) note that these disparities are deeply rooted in historical, social, and political contexts that have shaped the nation's economic path. They argue that this diversity makes it challenging to categorize South Africa strictly as a developing or emerging economy.

The landmark of the first democratic elections in 1994 marked a significant shift in South Africa's development trajectory. The post-apartheid era represented not only a political transition but also a crucial phase of economic restructuring. The country embarked on a journey of economic growth characterized by efforts in poverty reduction, economic stabilization, and policy reforms in areas such as trade, investment, and labor markets. These initiatives aimed to create a more inclusive and prosperous economy, diversify industries, and integrate into the global market. This set the foundation for growth and development in various sectors, notably in tourism (Lundahl and Petersson 2009). Despite these advancements, in recent years, the momentum of this progress appears to have slowed. The World Bank (2023) points out that while South Africa has made considerable strides in specific sectors, namely the mining industry and automotive sector, the country continues to grapple with deep-rooted wealth inequality. These disparities are not merely a reflection of the current economic conditions but an indicator of systemic issues that perpetuate inequality across generations.

The service sector in South Africa has also experienced significant growth, with tourism emerging as a pivotal component. Far from being just another economic segment, tourism is increasingly recognized as a fundamental pillar in the country's development. Renowned for its wildlife, stunning landscapes, and cultural diversity, South Africa has established itself as a globally recognized travel destination. The country caters to a wide range of tourist interests and preferences with its varied offerings, including mountains, deserts, safari options, and beaches (Britz and Venter 2016).

The sector's economic impact is substantial. According to a report by the WTTC, before the COVID-19 pandemic, the direct contribution of travel and tourism to South Africa's GDP was estimated at 2.9% in 2019, equating to about R139 billion. The influence of tourism extends beyond GDP contributions. It is a significant source of employment, supporting both direct and indirect jobs across various sectors. Additionally, the influence of tourism expands to improve related industries, including hospitality, transportation, retail and cultural enterprises (Balkaran and Maharaj 2014). This sector's impact on socio-economic development, especially in rural areas, is noteworthy, providing opportunities for small and medium enterprises and contributing to infrastructure improvements (WTTC, 2020).

3.3. Impact of COVID-19 on Tourism

The COVID-19 pandemic has drastically affected the global tourism industry, imposing travel restrictions that stopped traditional patterns of international travel. South Africa was no exception to these challenges. Statistics South Africa (2023) highlighted a steep decline in visitor numbers, from 14.8 million foreign arrivals in 2019 to 2.7 million in 2021. This had significant economic repercussions, notably reflected in a 330% decline in inbound tourism expenditure in 2021.

Despite these challenges, South Africa's tourism sector is showing a promising recovery trajectory. Statistics South Africa's 2023 annual report on tourism indicates that 2022 foreign arrivals have marked a notable increase. However, they still lag 44.3% behind the pre-pandemic levels of 2019. The first quarter of 2023 continues this positive trend, with significant improvements in tourist arrivals and spending. While the sector is on the path to full recovery, projections suggest it may take until 2025-2026 to return to pre-pandemic levels (S&P Global 2021).

The South African government has taken proactive steps in response to the pandemic's impact on tourism. The Tourism Sector Recovery Plan (2021), aligned with the South African Economic Reconstruction and Recovery Plan (2020) presented measures to stimulate economic growth and job creation, along with specific initiatives to revive the tourism industry. These strategies included marketing efforts, infrastructure development, and targeted support for small and medium enterprises within the sector.

3.4. Tourism Demand Forecasting

This chapter explores diverse methods and models used for predicting tourism demand worldwide. The precision of these forecasts plays an important role in the industry's strategic planning and policy preparation. Accurate forecasts are helpful for a broad range of stakeholders, including governmental bodies, businesses within the sector and investors (Vanhove 2010). The evolution of forecasting methods in tourism reflects the industry's dynamic nature and the need for analytical tools to anticipate future trends.

In 2019, Song, Qiu, and Park published a comprehensive review of 211 studies, spanning from 1968 to 2018, to examine the evolution of forecasting methods in tourism. This study highlights

the diversification and improvement in accuracy of forecasting methods, while adapting to the needs of tourism analysis. Notably, there has been a shift from non-causal time series models towards incorporating exogenous variables and multivariate dimensions, reflecting the need to account for structural changes and improve explanatory power.

This trend is evident in the frequent use of models like the Autoregressive Integrated Moving Average (ARIMA) and its variations, such as ARIMA with exogenous variables (ARIMA-X), Seasonal ARIMA (SARIMA), and Multivariate ARIMA (MARIMA).

Song, Wong, and Chon (2003) used the SARIMA and MARIMA models to forecast tourism in Hong Kong, incorporating significant external factors like visa issuance policies, financial crises, and epidemics. Similarly, a study on forecasting Zimbabwe's tourism arrivals employed the SARIMA model while emphasizing the lack of scientific methodologies in African nations for forecasting tourism demand (Makoni and Chikobvu 2018).

Moreover, the integration of AI-based methods, particularly those leveraging big data analytics, represents a significant evolution in forecasting techniques. However, this development brings challenges in creating new modelling techniques and understanding consumer behaviour in the context of big data (Song and Liu 2017). More recently, the COVID-19 pandemic has intensified these forecasting challenges due to its unpredictable nature. To account for the unpredictability of this period a study by Bespalova (2022) integrated both Google search data and crucial economic variables into an ARIMA-X model, aiming to enhance the accuracy of short-term forecasts in Aruba, a highly tourism-dependent economy. Similarly, a study that employed a Random Forest to forecast tourism inbound in China proved to improve accuracy by incorporating web search traffic data (Feng et al. 2019).

3.5. Economic Assessment of Tourism in South Africa

Tourism in South Africa is not just a vital economic activity but also a significant contributor to national development. Accurately assessing its economic impact is crucial for informed policy-making and strategic planning.

Central to South Africa's approach in quantifying tourism's economic impact is the Tourism Satellite Account (TSA). The TSA is an internationally recognized framework that offers detailed data on the contribution of tourism to key economic indicators like Gross Domestic Product (GDP) and employment. It provides a structured approach to measure tourism's direct and indirect economic effects, from spending patterns to job creation.

The latest TSA report covers the results for 2019 and provides provisional figures for 2020 and 2021, capturing the ongoing impact of the COVID-19 pandemic on the tourism sector. According to the report, the Tourism Direct Gross Value Added (TDGVA) experienced a significant increase, rising from R108,757 million in 2020 to R128,746 million in 2021, marking an 18% increase. Similarly, the Tourism Direct Gross Domestic Product (TDGDP) showed substantial growth, escalating from R117,946 million in 2020 to R140,095 million in 2021, a 19% increase. In terms of employment, the tourism sector directly employed 492,561 persons in 2021, with the tourism share of total employment rising from 30% in 2020 to 34% in 2021. These figures show the resilience and significant role of the tourism sector in the South African economy, even in the face of global challenges as the pandemic.

However, assessing the economic impact of tourism in South Africa also presents a few challenges. One of the main difficulties lies in accurately capturing the full spectrum of tourism's indirect and induced impacts. While the direct effects, such as spending in hotels or restaurants, are more straightforward to measure, the indirect benefits, like supply chain

expenditures and the induced impacts resulting from the spending of those employed in tourism are more challenging to quantify. This complexity underscores the need for sophisticated statistical models and a collaborative approach among government departments and stakeholders.

4. Preliminary Data Analysis

In this section, we delve into a preliminary analysis of the data which forms the foundation of our work-project. The decision to focus solely on international tourism data, rather than domestic tourism, was primarily influenced by the data's completeness. While domestic tourism volume is higher, international tourists' spending in South Africa is more substantial (Louw 2011). This approach allows for a more comprehensive analysis of the economic implications of tourism in South Africa and facilitates the development of targeted strategies to attract high-value tourists.

4.1. Data Collection

The data collection process began by sourcing the most detailed data on tourist arrival to South Africa, aiming to acquire as much information on arrivals as possible. The Statistics South Africa website provided annual tourism reports dating back to the year 2000, which included monthly figures for total arrivals and departures in the country. The data on arrivals is categorized into South African residents and foreign travelers, where the latter is further subdivided into visitors and non-visitors. We collected the monthly numbers of foreign arrivals classified as visitors for the available time frame: from January 2000 to September 2023. For the period from 1995 to 2000, only the total annual foreign arrivals were available.

In addition to the primary focus of our study, we aimed to explore other tourism related data. The United Nations World Tourism Organization (UNWTO) database, spanning from 1995 to 2021, included a wide range of annual statistics relevant to our study. Key among these were the total expenditure by foreign tourists, employment statistics within the tourism sector, occupancy rates of hotels and length of stay.

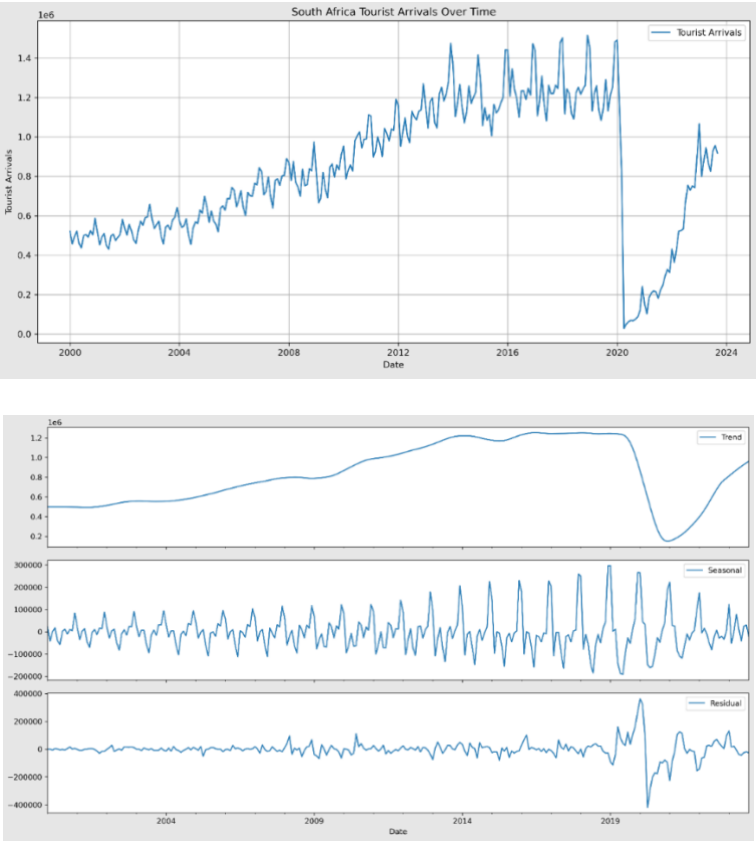
Furthermore, to gain insights intro the economic impact of tourism in South Africa a few macroeconomic indicators were collected. Real Gross Domestic Product (GDP) in domestic currency (Rand, ZAR) was obtained from the Federal Reserve Bank of St. Louis database. The exchange rate of Rand per US dollar was collected from the South African Reserve Bank. International Tourism Receipts, Inflation Rate and Foreign Direct Investment were derived from the World Bank database.

The real GDP data, available only in quarterly format, was converted to monthly frequency. This transformation was an essential step in aligning the data with the periodicity of the tourist arrivals in order to later be integrated in our forecasting model. To do so, Seasonal-Trend decomposition was employed using Loess (STL) followed by linear interpolation, as outlined by Theodosiou (2011). By first decomposing the data, we ensure that each aspect of the time series – trend, seasonality, and residual – is appropriately accounted for and interpolated, leading to a more nuanced and accurate monthly dataset. This approach preserved the integrity of the original quarterly values while smoothing the transitions in the estimation of monthly values for real GDP.

4.2. Data Exploration

The initial step in analyzing the monthly tourist arrivals to South Africa involved employing the Seasonal and Trend decomposition using Loess (STL), decomposing the observed time series into its constituent components: seasonal, trend, and residuals.

Figure 1: Monthly Tourist Arrivals to South Africa and STL Decomposition



The provided graphs in Figure 1 effectively capture the major trends and disruptions within the South African tourism sector. The observed data from the beginning of the century until the onset of the COVID-19 pandemic in March 2020 shows a growth trajectory with a period of relative stability before the pandemic. The substantial drop in arrivals due to the pandemic's travel restrictions is visibly marked in the trend and residual components of the time series decomposition. The seasonal component clearly reveals a seasonal pattern that not only persists

over time but also exhibits an increasing amplitude, suggesting potential changes in seasonality intensity. This pattern aligns with expectations for the tourism industry, which is known to be sensitive to seasonal fluctuations. Factors such as climate conditions, public holidays, and school break periods significantly shape tourist behaviors throughout the year (Hylleberg et al. 1990).

5. Miguel Gomes: Predictive Analysis of Tourist Arrivals to South Africa

5.1. Introduction

This individual part chapter focuses on the methodology and analytical approach employed in conducting a predictive analysis of tourist arrivals to South Africa. The primary objective is to develop a model that can accurately forecast the number of monthly foreign tourist arrivals to South Africa. Opting to forecast on a monthly basis provides a larger dataset to train our models and allows for a more granular view of tourism trends, which is crucial for a country where tourism is a vital economic contributor.

Python was the chosen programming language to build our models due to its data analysis and statistical modeling capabilities. Also, the advanced AI system ChatGPT was used for coding assistance and model optimization.

5.2. Forecasting Models

In this section, we discuss the specific models considered in our time series forecasting analysis. Our analysis revolves around two models: Seasonal Autoregressive Integrated Moving Average (SARIMA) and Seasonal Autoregressive Integrated Moving Average with exogenous factors

(SARIMAX). These models are chosen for their ability to handle time series data with seasonality and potential influence from external variables. Also, these models have achieved accurate results in several similar studies, as mentioned in our literature review.

5.2.1. SARIMA

At the core of the SARIMA framework is the Autoregressive Integrated Moving Average (ARIMA) model. First introduced by Box and Jenkins (1976), ARIMA has been a fundamental tool in time series analysis, valued for its effectiveness in leveraging delays and shifts in historical data to discover underlying trends and predict future patterns. The model is composed by three components:

- Autoregressive (AR): This part of the model takes past values in the series to predict future observations. The term *p* represents the number of past observations, also known as lags, used in this prediction process.

- Integrated (I): Involves achieving stationarity in the time series, which refers to the statistical properties of the series, such as the mean and variance, being constant over time. We achieve stability through a process called differencing, where the current value is subtracted from the previous one. It is represented by *d*, which corresponds to the order of differencing required to make the series stationary.

- Moving Average (MA): This component is designed to account for random fluctuations in the series, often referred as white noise. It incorporates the errors from previous forecasts into the current prediction, basically learning from past deviations. The value *q* indicates the number of lagged forecast errors included in the model.

In our analysis, we used the SARIMA model, an extension of the ARIMA model, which is particularly appropriate for data with seasonal and cyclical patterns. This is achieved by adding a seasonal component to the standard ARIMA structure:

- Seasonal Component (P, Q, D, s): This part of the model captures the patterns that occur over specific intervals. In our case of monthly data with annual seasonality, s is set to 12 to account for the yearly cycle. The parameters P, Q and D are the seasonal equivalents of the ARIMA model's components, specifically targeting the seasonal variations observed in the data.

In summary, the SARIMA model, denoted as SARIMA(p, d, q)(P, D, Q, s), effectively combines both non-seasonal and seasonal components. This combination allows for a comprehensive analysis, capturing both the cyclical patterns and the overall trends, of the monthly tourist arrivals to South Africa.

5.2.2. SARIMAX

The SARIMAX model extends the capabilities of SARIMA by incorporating exogenous factors into the model. This addition allows for the inclusion of variables that goes beyond the internal dynamics typically shown in time series data, providing a more comprehensive understanding of the factors that drive changes in the data. This extra layer of complexity has the capability of offering a more power forecasting tool, especially valuable in exploring the complex dynamics of foreign tourist that arrive to South Africa.

5.3. Model Assessment Metrics

To evaluate and compare the accuracy of our forecasting models for monthly tourism arrivals in South Africa, we used two primary metrics: Mean Absolute Error and Root Mean Squared Error.

5.3.1. Mean Absolute Error (MAE)

MAE measures the absolute average difference between the forecasted and the actual values of the model, providing a straightforward measure of prediction accuracy. The lower the MAE, the closer the model's forecasts are to actual outcomes. The MAE is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where y_i represents the actual values, $\hat{y}_i$ the forecasted values, and n the number of observations.

5.3.2. Root Mean Squared Error (RMSE)

On the other hand, RMSE is employed for its sensitivity to larger errors in forecasts. This metric squares the forecasting errors before averaging them, thus giving more weight to larger deviations. This characteristic of RMSE is particularly significant in the context of tourism forecasting, where incorrect predictions of large changes in arrivals can have considerable implications. Minor errors, while still important, might be less impactful in this context. The RMSE is calculated using the formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

5.3.3. Additional metrics

Alongside MAE and RMSE, we used Mean Absolute Percentage Error (MAPE) and Root Mean Square Percentage Error (RMSPE), which calculate the same measure but in percentage terms. Normalizing the metrics allows for an intuitive interpretation and direct comparison across different models.

The Akaike Information Criterion (AIC) measure was also considered during the model selection process. It is measure that primarily serves as a model fit indicator that penalizes excessive complexity. However, given that our primary goal is forecasting accuracy, particularly the model's ability to predict unseen data, RMSE and MAE were selected as the main metrics for model assessment.

5.4. Model Employment

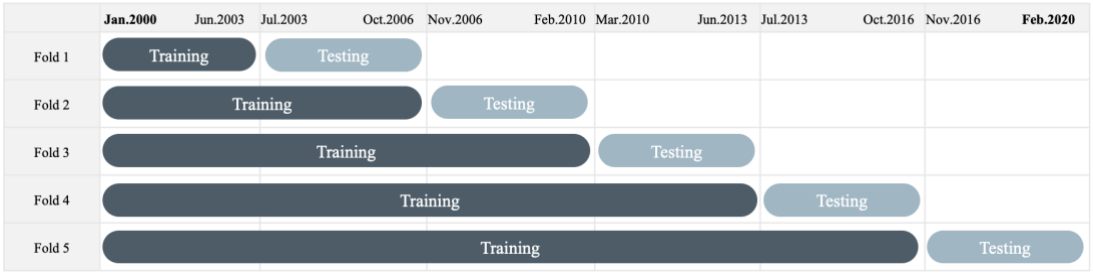
In developing our time series forecasting models, an important assumption was made regarding the impact of the COVID-19 pandemic on tourism data. An attempt was made to incorporate this unprecedented shock in our model. However, this integration severely affected the model's forecasting accuracy, particularly when predicting the future trend of the post-pandemic recovery period which is still underway. Given the ongoing recovery phase and the unprecedented nature of this global event, there is still uncertainty regarding the trajectory and duration of the recovery process. Therefore, to assure stability and enhance predictive accuracy, we focused our models exclusively on pre-pandemic data.

5.4.1. Training and testing

Our target variable ranges from January 2000 to February 2020, providing a total of 242 observations. We started by partitioning the series into a training set, 80% of the data, spanning

from January 2000 to January 2016 and a testing set (20%) from February 2016 to February 2020. This method of data splitting is recommended to avoid unwanted overfitting as it allows for a comprehensive evaluation of the model's performance on unseen data (Feng et al. 2019). Also, cross-validation, as recommended by Bergmeir and Benítez (2012), was performed in all models under consideration. This technique trains and test the data across multiple folds and stores the accuracy metrics for each fold. However, unlike traditional cross-validation methods where data is randomly split into training and test sets, time series cross-validation maintains the chronological order of observations. This approach involves sequentially expanding the training dataset fold by fold, testing each time on a subsequent period. Figure 3 offers a visual guide to this process, which depicts the cross-validation training and testing periods employed across the dataset.

Figure 3: Cross-validation Training and Testing periods



5.4.2. Parameter Selection

Selecting the optimal parameters for our SARIMA model is a crucial step in ensuring accurate and reliable forecasts. Accoding to Hyndman and Athanasopoulos (2021) in their book "Forecasting: Principles and Practice", the order of differencing, both non-seasonal (*d*) and seasonal (*D*), is first established to ensure stationarity of the time series. To do this an Augmented Dickey-Fuller (ADF) test was conducted on the monthly arrivals. This test evaluates the presence of a unit root, indicative of a time series where the fluctuations are not

constant and exhibit random walk behavior, thus affecting the predictability of the series. The test is represented by the following equation:

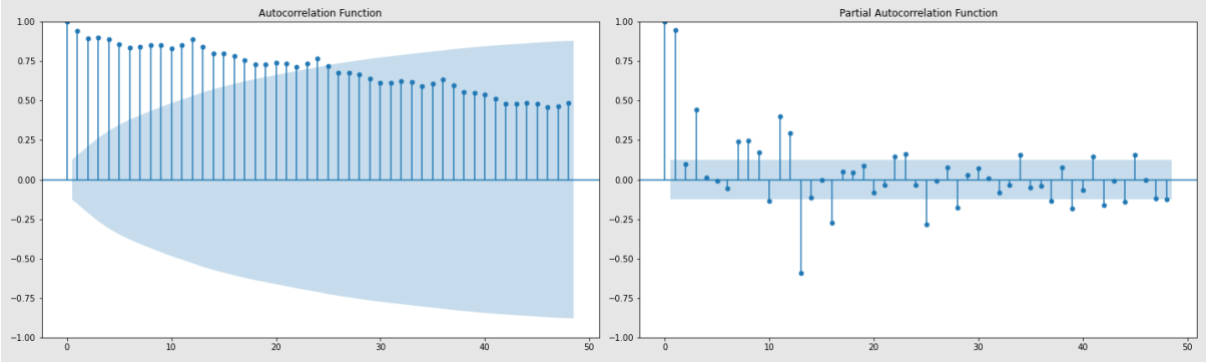
$$X_t = \beta X_{t-1} + \ + \alpha_1 \Delta X_{t-1} + \alpha_2 \Delta X_{t-2} + \ldots \ + \alpha_p \Delta X_{t-p} + \mu$$

The null hypothesis $H_0: \beta = 1$ suggests a unit root is present, and the alternative hypothesis $H_1): \beta < 1$ suggests stationarity.

The ADF test results yielded a t-statistic of -1.063775 and a p-value of 0.7293. Given the high p-value, we fail to reject the null hypothesis, indicating that the series is non-stationary. To address this, seasonal differencing was applied, a method that removes the repetitive seasonal patterns by subtracting the observation from the same period in the previous cycle. For our monthly data, this meant employing 12th order differencing to account for the annual cycle. The ADF test was reconducted and the results were significant, with a t-statistic of -3.215481 and a p-value of 0.019106. This allowed for the rejection of the null hypothesis at the 5% level, suggesting that the series became stationary after seasonal differencing. These findings imply that the parameter for seasonal differencing, *D*, should be set to 1, confirming the necessity of seasonal differencing to achieve stationarity in the series.

The next step in the parameter selection involved identifying the appropriate Autoregressive (AR) and Moving Average (MA) parameters for the SARIMA model, requiring an analysis of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) (Figure 4). The ACF measures the correlation between observations at different lags, providing insights into the general pattern of correlation over time. On the other hand, PACF identifies the correlation between residuals which remain after removing the effects already explained by earlier lags.

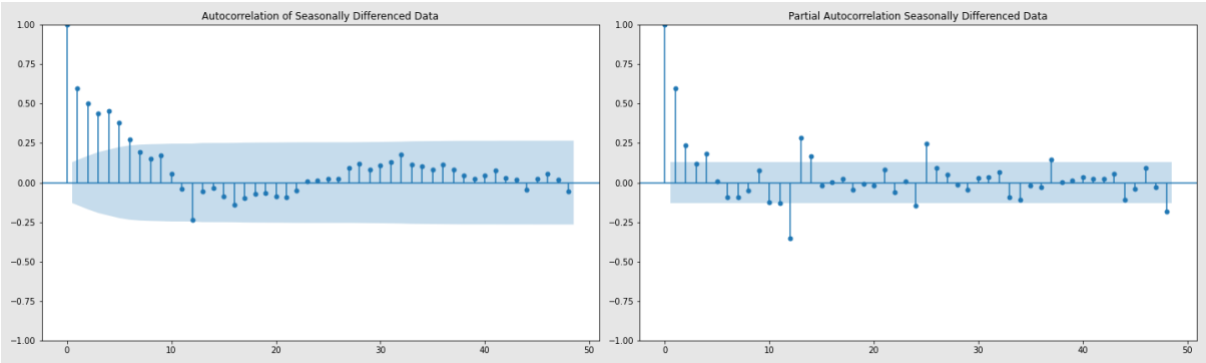
Figure 4: Autocorrelation and Partial Autocorrelation Functions



In Figure 4, the ACF plot exhibits pronounced autocorrelations at the seasonal lags of 12, 24, and 36, persisting above the significance threshold given by the blue shadow. This observation implies the presence of a seasonal pattern, necessitating the inclusion of seasonal Moving Average components (*Q*) to address these cyclical fluctuations. Regarding the PACF plot, it also shows significant spikes at the same seasonal intervals, underscoring the potential requirement for seasonal Autoregressive components (*P*).

To identify the non-seasonal elements within the series, we took a second layer of analysis, involving the ACF and PACF plots of the data seasonally differenced at order 12, as exhibited in Figure 5.

Figure 5: Seasonally Differenced Autocorrelation and Partial Autocorrelation Functions



The seasonal differentiated ACF plot reveals substantial autocorrelation at the initial lags, suggesting the need for a non-seasonal Moving Average component. This is corroborated by the PACF plot, where a significant peak at the first lag implies the inclusion of a non-seasonal Autoregressive component of at least order one. The diminishing partial autocorrelations beyond the first lag do not provide sufficient evidence for additional non-seasonal AR terms.

To refine the parameter selection for our SARIMA model, a grid search was employed on different combinations of order parameters in order to consider potentially more effective configurations:

- Non-seasonal orders (p, d, q): Ranging from (1, 0, 1) to (2, 1, 2).
- Seasonal orders (P, D, Q, s): Seasonal AR order P and MA order Q will be explored from 0 to 2, in line with the observed seasonal lags, while the seasonal differencing order D is established at 1. Parameter s is set to 12 given the monthly data.

Upon performing the grid search, the best performing model was the following SARIMA(2,1,1)(0,1,1)_12, with an average cross-validation RMSE and MAE of 78,863.7 and 65,352.4, respectively.

5.5. Application of Exogenous Variables in the SARIMAX model

This section explores the integration of relevant external factors into our forecasting model to potentially improve accuracy beyond the SARIMA model. Our goal is to include variables that directly impact tourism demand in South Africa. While selecting these variables, we were cautioned to maintain the model's integrity and prevent overfitting. Overfitting occurs when a model becomes excessively complex, tailored specifically to the training set rather than capturing underlying data patterns. This can lead to poor predictive performance on unseen

data. To mitigate this risk, we decided to select only two exogenous variables that could potentially impact tourism demand.

Our approach draws on the findings of Odhiambo and Nyasha (2020), who highlight the role of economic growth in South Africa not just as an outcome of tourism development but also as a determinant of it. Their study reveals a short-run bidirectional causality between tourism and economic growth in South Africa, with tourist arrivals in the country as a proxy for tourism development. This insight underscores the importance of considering an economic variable in our forecasting model, to accurately capture the dynamics of tourism demand.

In selecting an appropriate economic variable to incorporate into our model, we decided on using Real Gross Domestic Product (GDP) as a proxy for the nation's economic development. This decision was grounded in the recognition of real GDP's substantial correlation with tourist arrivals, coupled with its availability in a consistent and monthly format. However, it's important to address the issue of correlation versus causation in this context. While real GDP is highly correlated with tourist arrivals, this does not necessarily imply a direct cause-and-effect relationship. Although the bidirectional relationship observed in the Odhiambo and Nyasha study provides a basis for including real GDP in our model, we primarily regard the variable as a predictive factor, rather than a causal determinant.

Alongside real GDP, we considered the possibility of an additional measure that could capture the varying dynamics of changes in South Africa's tourism demand as a destination country. Google Trends provides insights into the popularity of specific keywords in Google web searches worldwide. This tool analyses the search frequency for selected terms, normalizing the data on a scale from 0 to 100. This scale represents the search interest relative to the highest point on the chart for a given region and time, allowing for an effective comparison of relative

popularity over time. As demonstrated by Feng et al. (2019), incorporating Google Trends data can significantly enhance the accuracy of tourism demand forecasting models. This is attributed to the ability of Google Trends to capture real-time changes in public interest and intent, offering a valuable digital perspective that complements traditional economic indicators. The data extracted is monthly, providing a consistent temporal alignment with our target variable, and dates back to 2004. Also, it has a feature to categorize the keyword popularity within a specific context, such as "travel" or "air travel". The selected keywords and their categories are detailed in Table 1.

Table 1: Google Trends keywords searches

Category	Search keywords
Travel	*South Africa*
	South Africa Beach
	Kruger National Park
Air Travel	*Cape Town*

The integration of exogenous variables into our model can potentially influence the underlying patterns and dynamics within the monthly tourist arrivals data. Therefore, it became necessary to reassess the model's optimal parameters. To address this, the seasonal and non-seasonal parameters were defined with the use of the *auto.arima* function, from *pmdarima* python package. This tool automates the process of model selection by iteratively exploring various combinations of parameters and selecting the optimal model based on the AIC measure.

We decided to train and test the SARIMAX model with each variable individually, as well as in combination with others. Specifically, we examined the influence of real GDP alone and together with each Google Trends keyword. The results of these models, including their non-seasonal and seasonal optimal parameters, cross-validated accuracy metrics and the coefficients

of each exogenous variable, are presented in Figure 6, in the appendix. We should note that not all exogenous variables achieved statistical significance in the model, indicated by their p-values.

Among the various configurations tested, the SARIMAX(2,0,1)(0,1,1)_12 model, incorporating real GDP and the keyword *South* Africa from Google Trends, presented the best results with the lowest RMSE and MAE values, 63 681.37 and 49 267.7, respectively. The corresponding RMSPE is 6.29% and MAPE 4.89%. The real GDP, measured in millions of Rands, had a coefficient of 0.4526, while the "South Africa" keyword had a coefficient of 2634.0, both coefficients with p-values lower than 5%.

This demonstrates the synergetic effect of coupling a traditional economic measure with digital search traffic data, a combination that has proven to enhance the accuracy of forecasting tourist arrivals to South Africa. By recognizing the importance of these variables in the predictive analysis of tourism, stakeholders are now equipped with insights that enables the formulation of strategies which are not only grounded in the current economic landscape but also aligned to the evolving digital interests of tourists.

5.6. Model Performance Benchmarking

Benchmarking is an essential step in assessing the efficacy of our forecasting model. It involves a comparison of our model's performance against established standards and other predictive models within the same field. Defined by Lewis (1982), forecast accuracy is categorized as follows: a Mean Absolute Percentage Error (MAPE) of less than 10% is considered highly accurate, 11–20% is good, 21–50% is reasonable, and over 51% is deemed inaccurate.

According to this definition our model with a MAPE of 4.89% can be considered highly accurate.

In an extensive study analyzing and forecasting the tourism demand to South Africa, a SARIMA model was used with monthly tourist arrivals as the dependent variable. The forecasts, made from specific different global regions to South Africa, exhibited results ranging from a RMSPE of 3.31% for Europe to 9.11% for Asia (Louw 2011). Equivalent studies in other countries have also shown effectiveness with similar models. Prilistya, Permanasari, and Fauziati (2021) applied a SARIMAX model for Indonesia, incorporating Google Trends as an exogenous variable, and achieved a MAPE of 5.46%. Also in Indonesia, Pratiwi et al. (2021) reached a MAPE of 3.6% in forecasting tourist arrivals in Indonesia using a SARIMA model.

The comparative analysis with other studies highlights the robustness of our SARIMAX model in forecasting tourist arrivals. With a MAPE of 4.89% and a RMSPE of 6.29% our model's performance is relevant, particularly in the context of its long forecasting horizon ranging from January 2004 to February 2020.

5.7. Limitations

It is crucial to acknowledge the limitations of our model. The primary constraint lies in its exclusive reliance on pre-pandemic data, up until February 2020, to train and test the model. The COVID-19 pandemic was a global shock in international travel, and its long-term effects are still unknown. This unprecedented event can potentially alter travel preferences and behaviors permanently. Our model sidesteps the pandemic's immediate impact to maintain stability and predictability. Therefore, it may not fully encapsulate the post-pandemic recovery and future tourism trends.

Additionally, our model intentionally avoids incorporating a broader set of exogenous variables, like political stability, environmental changes, and global economic conditions. This decision was made to prevent overfitting and to maintain the model's simplicity, ensuring reliable forecasts in the context of the unpredictable future of tourism. However, this simplification may make our model overlook essential tendencies that significantly impact tourist arrivals. This trade-off between model complexity and forecast reliability is a key consideration and underlines an area of future improvement in our predictive analysis.

5.8. Conclusion

Our study's use of SARIMA and SARIMAX models has proven to be highly effective in forecasting monthly tourist arrivals to South Africa. The inclusion of Real GDP and Google Trends data as exogenous factors increased the model's predictive accuracy, underlining the importance of integrating both economic indicators and web-search traffic metrics.

The positive coefficients of Real GDP and Google Trends data indicate a strong association with tourist arrivals. Real GDP, a broad metric indicative of the country's overall health, proves to have a significant predictive power on tourist arrivals. We can infer that this creates a reciprocal effect where tourism further boosts the economic vitality. Likewise, the influence of Google Trends data, specifically the web-search keyword *South Africa*, enhances the model's understanding of tourist interests and behavior. This highlights the importance of digital marketing strategies in promoting tourism destinations.

While aware of its limitations, the model stands as a strong tool, ready to be applied and further refined in the dynamic landscape of global tourism. The next step involves employing the model to project the future of foreign tourist arrivals to South Africa. It will not only test the model's

predictive capabilities in a real-world setting but also provide valuable and detailed insights for stakeholders in the tourism industry. Governmental bodies, tourism operators, and businesses can leverage these forecasts for strategic planning, resource allocation, and tailoring marketing initiatives. Particularly, the importance of establishing a strong online presence and targeted digital marketing efforts seems to be crucial for attracting foreign tourists to South Africa.

6. References

Africa, Statistics South. 2023. “SA Tourism Shows Slight Recovery after COVID-19 Pandemic. | Statistics South Africa.” May 22, 2023. https://www.statssa.gov.za/?p=16327.

———. n.d. “Publication | Statistics South Africa.” Accessed December 20, 2023. https://www.statssa.gov.za/?page_id=1854.

Baffi, Solène, Ivan Turok, and Celine Vacchiani-Marcuzzo. 2018. *The South African Urban System*.

Balkaran, R., and Shamina Maharaj. 2014. “A Comparative Analysis of the South African and Global Tourism Competitiveness Models with the Aim of Enhancing a Sustainable Model for South Africa.” *Journal of Economic and Behavioural Studies* 6 (April): 273–78. https://doi.org/10.22610/jebs.v6i4.490.

Bergmeir, Christoph, and José M. Benítez. 2012. “On the Use of Cross-Validation for Time Series Predictor Evaluation.” *Information Sciences*, Data Mining for Software Trustworthiness, 191 (May): 192–213. https://doi.org/10.1016/j.ins.2011.12.028.

Bespalova, Olga. 2022. “Modeling and Forecasting Monthly Tourism Arrivals to Aruba Since COVID-19 Pandemic.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.5089/9798400225871.001.

Box, George E. P., and Gwilym M. Jenkins. 1976. *Time Series Analysis: Forecasting and Control*. Holden-Day.

Britz, Peter, and Samantha Venter. 2016. “Aquaculture Review: South Africa.” *World Aquaculture*, December, 20–28.

Feng, Yuyao, Guowen Li, Xiaolei Sun, and Jianping Li. 2019. “Forecasting the Number of Inbound Tourists with Google Trends.” *Procedia Computer Science*, 7th International Conference on Information Technology and Quantitative Management (ITQM 2019): Information technology and quantitative management based on Artificial Intelligence, 162 (January): 628–33. https://doi.org/10.1016/j.procs.2019.12.032.

Garidzirai, Rufaro, and Michael Pasara. 2021. “AN ANALYSIS OF THE CONTRIBUTION OF TOURISM ON ECONOMIC GROWTH IN SOUTH AFRICAN PROVINCES: A PANEL ANALYSIS.” *Geojournal of Tourism and Geosites* 29 (June): 554–64.

Hylleberg, S., R. F. Engle, C. W. J. Granger, and B. S. Yoo. 1990. “Seasonal Integration and Cointegration.” *Journal of Econometrics* 44 (1): 215–38. https://doi.org/10.1016/0304-4076(90)90080-D.

Hyndman, Rob, and G. Athanasopoulos. 2021. “Forecasting: Principles and Practice.” https://research.monash.edu/en/publications/forecasting-principles-and-practice-3.

Lemma, Alberto F. n.d. “Evidence of Impacts on Employment, Gender, Income.”

Lewis, C. D. (Colin David). 1982. *Industrial and Business Forecasting Methods : A Practical Guide to Exponential Smoothing and Curve Fitting*. London ; Boston : Butterworth Scientific. http://archive.org/details/industrialbusine0000lewi.

Louw, Riëtte. 2011a. “Forecasting Tourism Demand for South Africa.” Thesis, North-West University. https://repository.nwu.ac.za/handle/10394/7607.

———. 2011b. “Forecasting Tourism Demand for South Africa.” Thesis, North-West University. https://repository.nwu.ac.za/handle/10394/7607.

Lundahl, Mats, and Lennart Petersson. 2009. “Post-Apartheid South Africa: An Economic Success Story?” In *Achieving Development Success*, edited by Augustin K. Fosu, 232–64. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199671557.003.0012.

Makoni, Tendai, and Delson Chikobvu. 2018. “Modelling and Forecasting Zimbabweâ€™s Tourist Arrivals Using Time Series Method: A Case Study of Victoria Falls Rainforest.” *Southern African Business Review* 22 (November): 22 pages-22 pages. https://doi.org/10.25159/1998-8125/3791.

“News Article | World Travel & Tourism Council (WTTC).” n.d. Accessed December 20, 2023. https://wttc.org/news-article/south-africas-travel-and-tourisms-growth-to-outpace-the-national-economy-for-the-next-10-years.

Odhiambo, Nicholas M., and Sheilla Nyasha. 2020. “Is Tourism a Spur to Economic Growth in South Africa? An Empirical Investigation.” *Development Studies Research* 7 (1): 167–77. https://doi.org/10.1080/21665095.2020.1833741.

“Overview.” n.d. Text/HTML. World Bank. Accessed December 20, 2023. https://www.worldbank.org/en/country/southafrica/overview.

Prilistya, Suci, Adhistya Permanasari, and Silmi Fauziati. 2021. “The Effect of The COVID-19 Pandemic and Google Trends on the Forecasting of International Tourist Arrivals in Indonesia.” In , 1–8. https://doi.org/10.1109/TENSYMP52854.2021.9550838.

"Sharing Insights Elevates Their Impact." 2021. S&P Global. June 24, 2021. https://www.spglobal.com/marketintelligence/en/mi/research-analysis/subsaharan-african-tourism-industry-unlikely-recover.html.

Song, Haiyan, and Han Liu. 2017. "Predicting Tourist Demand Using Big Data." In *Analytics in Smart Tourism Design: Concepts and Methods*, edited by Zheng Xiang and Daniel R. Fesenmaier, 13–29. Tourism on the Verge. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-44263-1_2.

Song, Haiyan, Richard T. R. Qiu, and Jinah Park. 2019. "A Review of Research on Tourism Demand Forecasting: Launching the Annals of Tourism Research Curated Collection on Tourism Demand Forecasting." *Annals of Tourism Research* 75 (March): 338–62. https://doi.org/10.1016/j.annals.2018.12.001.

Song, Haiyan, Kevin K.F. Wong, and Kaye K.S. Chon. 2003. "Modelling and Forecasting the Demand for Hong Kong Tourism." *International Journal of Hospitality Management* 22 (4): 435–51. https://doi.org/10.1016/S0278-4319(03)00047-1.

"South African Economic Reconstruction and Recovery Plan | South African Government." n.d. Accessed December 20, 2023. https://www.gov.za/documents/other/south-african-economic-reconstruction-and-recovery-plan-15-oct-2020.

Theodosiou, Marina. 2011. "Forecasting Monthly and Quarterly Time Series Using STL Decomposition." 2011. https://www.researchgate.net/publication/227417758_Forecasting_monthly_and_quarterly_time_series_using_STL_decomposition.

Tisdell, Clement, and Clevo Wilson. 2013. "Public Economics and the Assessment of Tourism Developments and Policies." In , 417–41. https://doi.org/10.1142/9789814327084_0019.

"Tourism Sector Recovery Plan." n.d. Accessed December 20, 2023. https://www.tourism.gov.za/CurrentProjects/Pages/Tourism_Sector_Recovery_Plan.aspx.

Vanhove, Norbert. 2010. "Forecasting Tourism Demand." In *The Economics of Tourism Destinations*, 2nd ed. Routledge.

World Tourism Organization (UNWTO), ed. 2021. *International Tourism Highlights, 2020 Edition*. World Tourism Organization (UNWTO). https://doi.org/10.18111/9789284422456.

7. Appendix

Figure 6: SARIMA & SARIMAX model's results

Exogenous variables	Optimal model	Cross-validated error measures				AIC	Coefficients	
		RMSE	RMSPE	MAE	MAPE		Var1	Var2
-	SARIMA(2,1,1)(0,1,1)_12	78863.70	9.93%	65352.4	8.29%	4409.5	-	-
Var1: Real GDP	SARIMAX(1,0,1)(0,1,1)_12	70158.03	8.21%	55709.07	6.66%	4402.63	0.8108***	-
Var1: *South Africa*	SARIMAX(3,1,1)(1,1,1)_12	74862.65	7.77%	60494.15	6.17%	3516.29	1229.35	-
Var1: *South Africa Beach*	SARIMAX(3,1,1)(1,1,1)_12	82218.95	8.82%	67061.12	7.10%	3523.5	-79.11	-
Var1: *Kruger National Park*	SARIMAX(1,1,1)(0,1,1)_12	79423.67	8.47%	64950.96	6.85%	3517.88	393.07	-
Var1: *Cape Town*	SARIMAX(1,1,1)(0,1,1)_12	83301.14	8.97%	69647.93	7.43%	3518.93	437.51	-
Var1: Real GDP Var2: *South Africa*	**SARIMAX(2,0,1)(0,1,1)_12**	**63,681.37**	**6.29%**	**49,267.7**	**4.89%**	**3519.2**	**0.4526****	**2634.0****
Var1: Real GDP Var2: *South Africa Beach*	SARIMAX(3,1,0)(0,1,1)_12	98396.71	10.90 %	85704.68	9.45%	3493.35	0.7668***	392.62
Var1: Real GDP Var2: *Kruger National Park*	SARIMAX(3,1,0)(0,1,1)_12	70326.10	7.16%	55066.87	5.57%	3494.69	0.8061***	-83.02
Var1: Real GDP Var2: *Cape Town*	SARIMAX(1,0,1)(0,1,1)_12	65602.51	6.66%	51912.03	5.33%	3518.2	0.6344***	1508.7*

Significance levels: *** 1%, ** 5%, * 10%