Analysis on the Seasonality of European Tourism

First Author
Chengzhong Meng

Second Author
Kiarash Farivar

Third Author
Michael Cai

Fourth Author
Zhiqing Hu

Abstract

This document explores how seasons and features of a country impact tourism in European countries. Seasonality is the repetitive, annual changes which cause a shift in tourists' behaviour and imbalances in social resource allocation. This document focuses on exploration of tourism data provided by Eurostat[1].

1 Introduction

Europe is one of the most popular tourist destinations thanks to its rich cultures, diverse landscapes, and quality public infrastructures. Tourism, being one of the fastest growing industries, has contributed more than 782 billion Euros to GDP in 2018 for the EU economy. Thus, tourism forecasting is becoming an increasingly important activity in planning and managing the industry. Our goal is to investigate the nature of tourism and the patterns in the popularity of the countries. We will investigate the change in nights spent by people at hotels between seasons, predict future trends, and analyze reasons for shifting tourism. The insights obtained could be valuable for employees, government officials, and businesses involved in the tourism sector.

2 Data

We will primarily be using public dataset provided by the Eurostat[1], mainly focusing on the datasets under Database by themes>Industry, trade and services>Tourism (tour)>Monthly data on tourism industries. The data we used is split among many tables. Different ones were used on the tasks. Additional climate data came from the World Bank[2].

3 Research Questions

The three main questions we decided to answer/explore are:
1. What patterns are in the time series data for nights spent. And how can we use them to predict the future behaviour of tourists.
2. What factors contribute to monthly tourism in any given country.
3. Can various economic, geographic, and societal indicators be used to predict tourism popularity.

4 Seasonal Analysis

4.1 General seasonal patterns

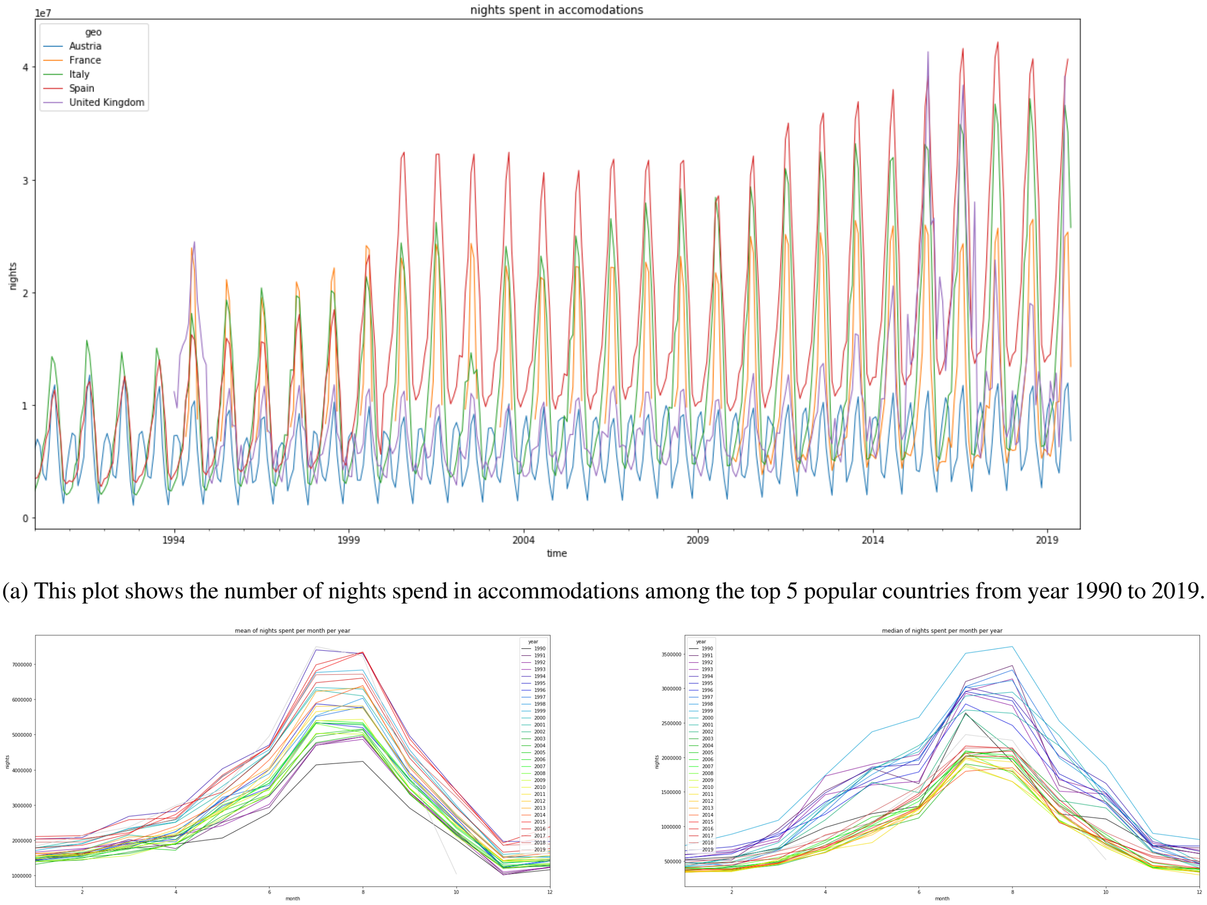
Figure 1(a) shows a nice dependence of nights spent on the season. There is a great seasonal pattern with same yearly frequency. The time series for the top 5 countries with the most nights spent is shown in Figure 1(a). From the whole time line, we can see a gradual increase of the nights of travelling in the top 5 popular countries from year 1990 till 2019. With Figure 1(a), it is hard to see the details among months.

Figure 1(b-left) displays the average nights spent per month for each of the year. Summer is always the popular season for travelling as shown in the Figure 1(b). A dramatic peak is seen for July and August followed by a dramatic decrease. We could also see a small increase on December for all the years. This slight increase could by explained by the Christmas holidays, a popular travelling period for western countries.

By comparing Figure (b-left) and Figure (b-right), we also found that the median of recent

[1] https://ec.europa.eu/eurostat/data/database

[2] https://datahelpdesk.worldbank.org/knowledgebase/articles/902061-climate-data-api



(a) This plot shows the number of nights spend in accommodations among the top 5 popular countries from year 1990 to 2019.

(b) These two plots shows the mean VS median number of nights spent per month from different year to compare from the year domain. Different colors refer to different year and more recent years have more redish colors.

Figure 1: seasonal data

years are consistently lower than many past years. However, the mean of recent years are higher than the past years. This suggests that, in recent years, a few selected countries attracts a considerably more visitors; while many other countries loose their popularity.

A coloured map is included in the appendix to display each country's seasonality [7]. Seasonality is measured by the standard deviations of nights spent among the months.

4.2 Time series analysis

To predict the future number of nights spent by the tourists, we have analyzed the time series of each of the top 3 countries with the most nights spent by foreign tourists: Italy, Spain and France.

We will focus on Italy since the analysis for all 3 are similar. Figure 2 shows a general increasing trend of both tourism overall and the seasonal fluctuations. A clear seasonal pattern can be seen, and confirmed by analyzing the autocorrelation.

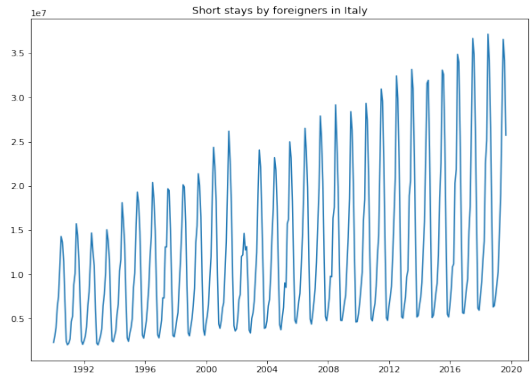


Figure 2: Seasonal trends in Italy

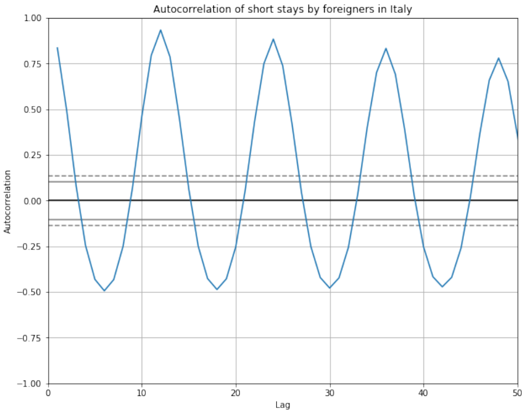


Figure 3: Autocorrelation plot for Italy. The time lag between the peaks is 12 months

4.2.1 Prediction Task (SARIMA)

As shown in Figure 4, the data can be decomposed into 3 distinct components. Under this assumption, Seasonal ARIMA (SARIMA) models works well for predicting future nights spent by tourists. The ARIMA model tries to make the data stationary so that it is possible to perform linear regression on the data. It achieves this by using the moving average and the autoregression terms in the model. We perform cross validation with a $66\% - 34\%$ split; and use grid search to identify the optimal hyper-parameters for the model under RMSE. The full data is then used to train final model, shown in Figure 5. The final RMSE is 1.2×10^6 which is around 10% of the order of number of nights spent.

The plots for the walk forward evaluation are in the appendix. [6]

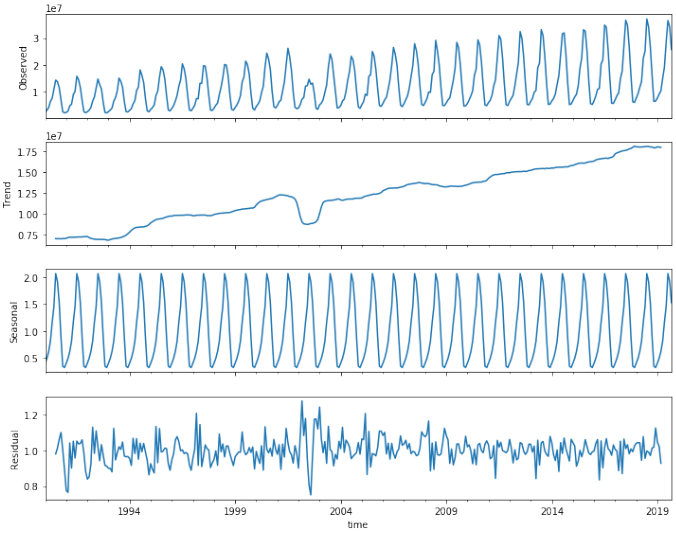


Figure 4: The time series can be decomposed into 3 parts under a multiplicative model: a main trend (moving average), a seasonal trend, and noise. Multiplying the 3 signals results in the original time series

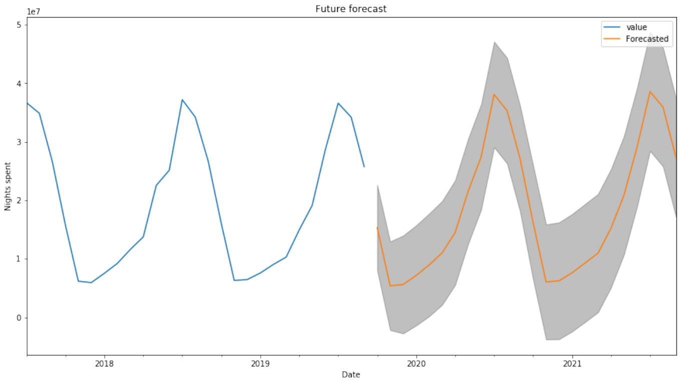


Figure 5: Prediction of nights spent by tourists, going 2 years into the future. The 95% confidence interval is shown in gray.

5 Analysis of Factors Contributing to Tourism

5.1 Concept

During training, supervised machine learning models are designed to discover hidden relationships between the features and the predicted variable. By analyzing trained models, we attempted to gain a better understanding of what factors contribute to a given country's tourism from month to month. We used regression models, including various societal and weather metrics as features, to predict the number of overnight stays by foreigners in a given region. We then conducted feature selection, and analyzed the relationships that the model had discovered.

5.2 Methodology

The predicted variable was the number of overnight stays. The input features were various metrics about the country during that month. In other words, there was 1 data point for each country and month. The features represented many aspects of the country and its citizens, including climate, land usage, distribution of people, wealth and income, education, crime, price of products and services, and purchasing power.

Multi-layer perceptions (with ReLU activation, and MSE loss) were used to capture the complexity of the relationships. To evaluate the importance of the features, we used two methods. One was forward recursive feature selection, where we trained a new model for every unused feature in each iteration. Given that thousands metrics made up the features, this became extremely computationally expensive. A faster method was a variant of permutation importance, where we scrambled a feature to see what impact it had on a trained model's performance. If randomizing any given feature did not affect the model's predictions, we could then assume the feature to be unimportant. On the other hand, if manipulating the feature causes a massive drop in performance, then we can assume that feature to be important to the model's predictions. Performance was mostly measured using Mean Squared Error and R Squared.

For the most popular countries, the predictions came close to the real values. The data provided by large countries tend to come in greater quality and quantity. Their massive size and population also made them less vulnerable to noise and temporary events, making their data more consistent. The prediction results for some of these countries are in the appendix.

5.3 Results of Feature Analysis

Forward recursive feature selection and permutation importance provided pathways to gain some insights into what factors contribute to the monthly number of overnight stays in a country. While general trends emerged, the nondeterministic nature of neural networks, multicollinearity, and the lack and quality of data made analysis difficult, described in greater detail in the notebook.

The first feature selected by the forward recursive method was always the amount of land in the country occupied by buildings with 3 or more floors (followed by temperature). It is a good predictor of annual tourism, and suggests that tourists to European countries tend to stay within cities and urban areas.

Given all the available features, the model prioritizes precipitation levels and temperature. Scrambling these features causes the most damage to the model's predictions, and indicates the high impact seasonality has on tourism trends. In general, temperature has a correlation of 0.88 with tourist numbers, while precipitation has -0.85. In other words, high temperatures and low precipitation are primary factors in the attractiveness of a country to tourists in Europe.

The next most important features are the price of tourism-related products, including cultural services, passenger transport via sea or waterway, accommodation, and holiday packages. However, the correlation of these features with the number of tourists is positive - in other words, the more expensive tourist activities are, the more tourists there are. This positive correlation is strongest with accommodation and holiday packages, with a value of 0.66. This may be because higher quality services are more expensive, or because high demand increased the price. Whatever the reason is though, it seems high prices are not able to deter tourists from a location.

A high proportion of prisoners plays a small role in determining tourism numbers. A high number of people in bad health within cities also seems to be a minor deterrent to tourism activity, maybe because it projects a negative image of the country.

Interesting enough, many factors that contribute to the well-being of a country's citizens do not affect tourism numbers much. The price of everyday goods and groceries, the education levels of the citizens, the median and mean incomes, and purchasing power seem to play little to no role. It appears that tourists are not particularly concerned about the well-being and wealth of the locals when choosing tourism destinations.

6 Conclusion

Tourism as measured by nights spent in the country is a highly seasonal time series. Looking at the total number of nights spent, there is a general increase. The seasonal patterns also enables us to predict the data in the future with good accuracy. But one limitation of this method is that it can't predict changes in the main trend of our data. For that, features relating to the countries should be used.

The development of urban areas, greater attention to tourism-related services and products, reducing crime seems to be viable paths of encouraging more tourism within a country. However, as temperature and precipitation are major factors, some countries may find overcoming their natural disadvantage difficult.

7 Appendix

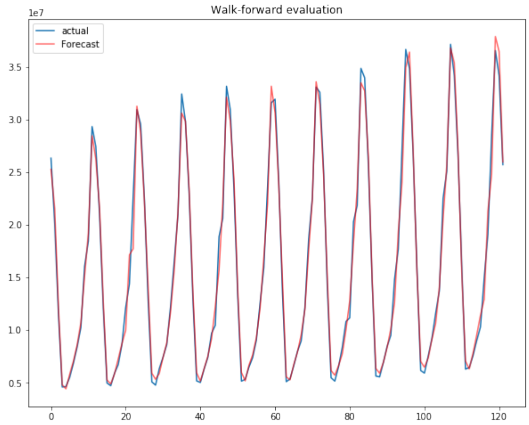


Figure 6: SARIMA predictions on Italy's overnight stays. SARIMA is able to capture the patterns in the data well, due to clear seasonal trends

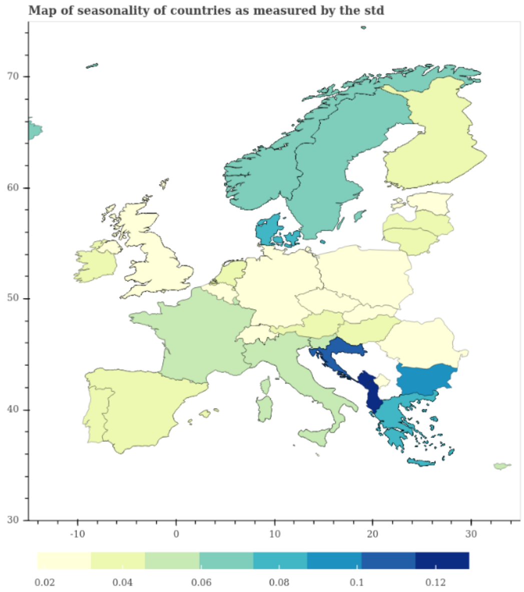


Figure 7: Impact of seasonality of each country, with darker countries being more affected. Impact is represented by seasonal fluctuations, and is measured by standard deviation of monthly data.

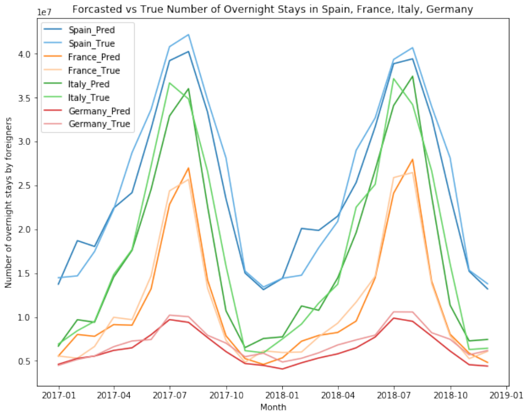


Figure 8: Regression prediction results for major countries. Using societal and climate data, the amount of tourism in these countries seems to be successfully captured