# Application Survey of Time Series Trend Prediction Models on Hotel Booking Dataset

**Sambit Mukherjee  Lingjia Shi**

**Abstract:**

Time series is the task of fitting a model to historical, time-stamped data to predict future values. This research explored the precision of diverse time-series models for Hotel Booking Trend Prediction in all the months in a year over a span of 12 years, i.e., 2010-2022, with the maximum count of booking confirmed as 622 and the minimum count confirmed as 102, as of December 2021. The results obtained showed that based on RMSE and MAPE metrics, Holts Winter Multiplicative Model and Seasonal Auto Regressive Integrated Moving Average Model obtained the best values for both sets, training, and test respectively. While in this project we focused on a trend prediction for hotel bookings, this method of survey can and are being applied to various other impactful and interesting fields of research including Demand forecasting.

## Introduction:

In the previous two decades, tourism has been cited as one of the most important economic sectors in the global market. Its position in global trade places is second to only oil, setting it apart from other sectors [1]. Making a sound decision today, especially in a situation of uncertainty, is essential in the tourism industry, like hotels. If it were feasible to predict changes in the number of bookings of a hotel by looking at current and previous tourism demands, promotion in the tourism industry would've been made simpler. It is necessary to keep in mind and improve upon the prior standard of performances, due to the complex and competitive environment in the tourism industry. The value of accuracy in predicting tourism demand cannot be denied; However, choosing a model that fits an issue is just as important as getting accurate results.

The method used to forecast tourism demand can be categorized into three themes,

namely, causal econometric models, time series models [2] and artificial intelligence-based methods [3]. Since datasets of hotel booking forecasting are highly seasonal, time series models are considered to be the current best options.

Time series methods need only one data series, and are based on the level of complexity of the model. Autoregressive moving average (ARIMA) models [4,5], and exponential smoothing (ES) models have gained popularity.

In lieu of choosing the best model and getting as good accuracy of trend prediction as possible out of a plethora of models, the dataset was experimented on with different categories of time series models. Researchers have proposed various tourism forecasting models, and each model, of course, has both advantages and drawbacks, and we have tried the models which are most in trend for predictive analysis.

In this project, eight models were applied to a hotel booking forecasting problem dataset, and the best algorithm was found, after comparing their performances, namely, Holts Winter's Multiplicative Forecast, followed by Seasonal Auto Regressive Integrated Moving Average Forecast with the lowest RMSE and MAPE score, two prominent metrics for time series evaluation.

**Related Works:**

Back in 1990, Andrew and others used Box-Jenkins and exponential smoothing, for forecasting hotel occupancy rates and both models show a high level of predictive accuracy [6]. Later on, Jae H.Kim and his team members evaluates the performances of prediction intervals generated from alternative time series models, including bias-corrected bootstrap for autoregressive model, seasonal ARIMA models and state space models for exponential smoothing. Among all of them, bias-corrected bootstrap performs best in general, providing tight intervals with accurate coverage rates, especially when the forecast horizon is long [7].

Kuan-Yu Chen combined linear and nonlinear statistical models to improve forecasting accuracy for Taiwanese outbound tourism [8]. One of the widely used neural network techniques for forecasting is backpropagation. Sometimes this approach does not provide a desirable accuracy of forecasting. Noersasongko, etc. predict the foreign tourist arrivals in three cities of Indonesia based

on combination of backpropagation and genetic algorithms. This approach can increase forecasting accuracy, and the back propagation algorithm's parameters are tuned to reduce prediction error [9].

Subedi, A proposed an approach of combining biquadratic polynomial function and autoregressive model on tourist arrival time series data with seasonal fluctuation. From the results, it considered to be an alternative approach for picking up a good model for such type of data [10].

**Data:**

This is a timeseries dataset which lists the number of bookings from the year 2010-2022. It consists of the month/year and the number of hotel bookings made during then. This dataset is downloaded from Kaggle. The values looks like it has been generated randomly by others, as there is no possible way there can only be 100-300 bookings in a month. We are working with a relatively small dataset. The reason for this is that we would partition the dataset into 200-300 samples anyway. Therefor we considered it best to, work with a smaller dataset, to remove the hassle. The data was preprocessed, imputing missing values, giving missing indexes to the columns, and finding seasonal trends and outliers in the dataset.

**Methods:**

The Models that we used for Time series forecasting are as follows:

1. Simple Forecasting Methods: As the name suggests, this is a simple method, but also can be very effective in many cases, just not this project, due to the availability of small data samples and outliers on top of that. Regardless, they were worth discussing, as a conclusion of importance of trend and seasonality prediction became apparent. Some Simple forecasting method that was used were:
    1.1 Naive Method: The naïve method of forecasting dictates that we use the previous period to forecast for the next period. The naïve method has its benefits and its limitations. The benefits are it is inexpensive to develop, store data, and operate. The limitations are that it does not consider any possible causal relationships that underly the forecasted variable. The naïve method is the

simplest form of business forecasting, but there are slightly more complex forecasting models that are almost as simple as naïve forecasting [11]. Some of them are listed below.

1.2 Simple Average Method: In Simple Average Method, just average the data by months, or quarters or years and then calculate the average for that period. Then find out what percent it is to the grand average.

Seasonal Index = (Monthly or Quarterly Average/ Grand Average of Months or the quarters) *100

1.3 Simple Moving Average Method: Simple moving average (SMA) or rolling average is the arithmetic mean of observations of the full data set and uses the arithmetic mean as the predictor of the future period. This method is used to smooth out short-term deviations of time series data and indicate long-term trends or cycles [12]. The equation of SMA is as follows:

$$F_t = MA_n = \sum_{i=1}^{n} \frac{Di}{n}$$

Where $F_t$ is the is the forecast for the time period t, is demand in period I, and n is the number of periods in the moving average.

2. ARIMA Method: ARIMA is a method for forecasting or predicting future outcomes based on a historical time series.

Some of the ARIMA Methods include:

2.1 Autoregressive method (AR Method): Autoregressive models operate under the premise that past values influence current values, which makes the statistical technique popular for analyzing nature, economics, and other processes that vary over time. Autoregressive process is one in which the current value is based on the immediately preceding value, while an AR process is one in which the current value is based on the previous two values. An AR process is used

for white noise and has no dependence between the terms. In addition to these variations, there are also many different ways to calculate the coefficients used in these calculations, such as the least squares method [13].

2.2 Seasonal Auto Regressive Integrated Moving Average Method: SARIMA is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality [14]. Therefore, this method takes the seasonal variations in time series trends in different months.

3. Exponential Smoothing Methods:

The most popular forecasting techniques are exponential smoothing techniques. The original work of Brown and Holt [15, 16] provided the foundation for the development of exponential smoothing forecasting methods in the 1950s. It gives the observed time series different weights. More importance is placed on recent observations than on older ones. One or more smoothing parameters are used to achieve the unequal weighting, and they control how much weight is assigned to each observation.

3.1 Simple exponential smoothing (SES), the most basic technique of this type, is appropriate for a series that moves randomly above and below a constant mean (stationary series). There is no trend or seasonal pattern. This model is given by the model equation $y(t) = \beta(t) + \varepsilon(t)$, where $\beta(t)$ takes a constant at the time t and may change slowly over the time; $\varepsilon(t)$ is a random variable and is used to describe the effect of stochastic fluctuation [18].

3.2 Holts method with trend (Double exponential smoothing model): This addresses both the level(l) and trend (b) component of the time series. Thus, two smoothing constants are used i.e. alpha for the level component and beta for the trend component. The equations are as follows:

$$l(t) = \alpha * y(t) + (1-\alpha)*(l(t-1)+b(t-1)) \text{ — — — — —} \textbf{ Level } l$$

$$b(t) = \beta * (l(t) - l(t-1)) + (1-\beta)* b(t-1) \text{ — — — —} \textbf{ Trend } b$$

$$y(t+1) = l(t) + b(t) \text{ — — — —} \textbf{ Forecast}$$

Models with low values of beta assume that the trend changes very slowly over time while models with larger value of beta assume that the trend is changing very rapidly [19].

3.3 Triple Exponential Smoothing / Holt Winter's Method :

In this method, smoothing to seasonal component in addition to level and trend components were applied. The smoothing is applied across seasons. That is, the fourth component of one season is smoothed against the fourth component of the previous season, fourth component of two seasons ago and so on. Thus, there will be four equations — one each for level, trend, seasonality and the final equation with all the individual components.

The four equations of the multiplicative Holt Winter's Method is given as :

$$y_{t+h|t} = (\ell_t + hb_t)s_{t+h-m(k+1)}$$

$$\ell_t = \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1}+b_{t-1})$$

$$b_t = \beta*(\ell_t - \ell_{t-1}) + (1-\beta*)b_{t-1}$$

$$s_t = \gamma \frac{y_t}{(\ell_{t-1}+b_{t-1})} + (1-\gamma)s_{t-m}$$

Here s is the season length i.e. the number of data points after which a new season begins [19].

**Experiments and Results:**

1. Preprocessing:

   1.1 Imputation: To deal with the NULL values, we use Mean Imputation to fill the NULL values averaging from its previous values.

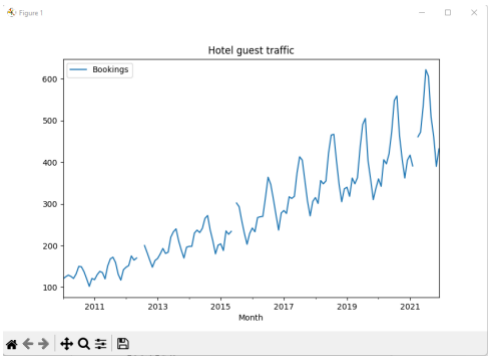


Figure 1:Missing Values

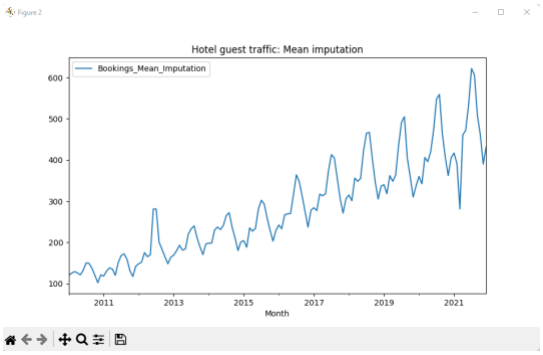


Figure 2: Imputed Valued through Mean Imputation

Comparing with figure 1, the gap has been closed in figure 2 after implementing Mean Imputation.

1.2 Seasonal Decomposition: To find a seasonal trend for hotel bookings in each respective years, we apply seasonal decomposition to our model. Seasonally adjusted value removes the seasonal effect from a value so that trends can be seen more clearly.

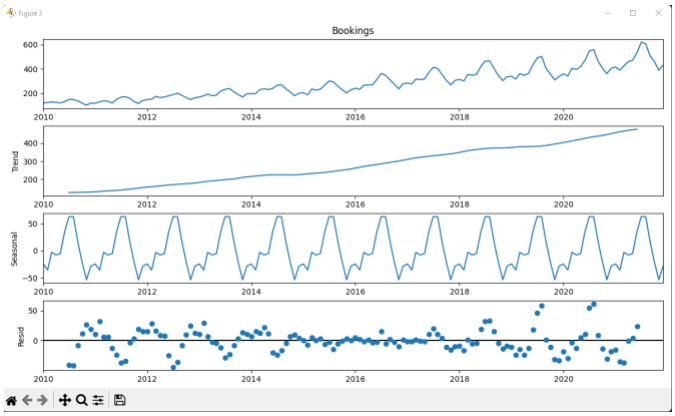


Figure 3: Finding the seasonal trend using additive seasonal decomposition

The plot shows the observed series, the smoothed trend line, the seasonal pattern, and the random part of the series

1.3 Outlier Detection: To detect points that are considered “abnormal,” or which don't fit a particular seasonal trend, we use outlier detection. Outlier detection estimators thus try to fit the regions where the training data is the most concentrated, ignoring the deviant observations.

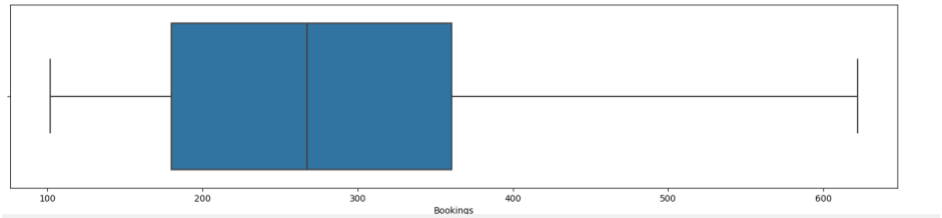


Figure 4: Using Box plot to detect outliers

In this figure, the outlier values can be detected at data points of the quartiles, Q1 and Q3, whose values are 180 and 350, respectively.

2. Building the Models:

2.1 Simple Forecasting Methods

2.1.1 Naïve Model (NM):

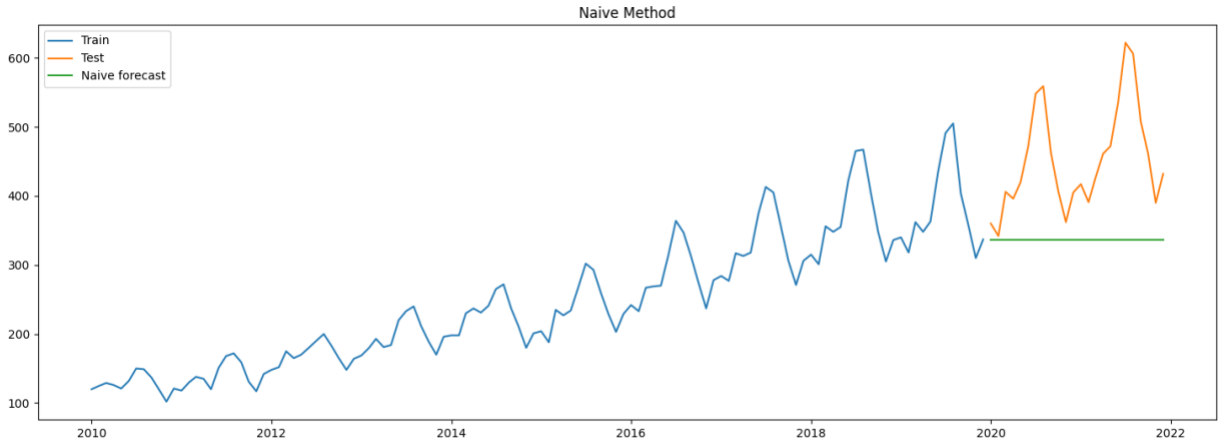


Figure 5: Naive Forecast for Train and Test Set

This model does not consider any possible causal relationships that underly the forecasted variables, and therefore is not a good model for this dataset, as represented by a horizontal line which does not go through any correct data points and hence does not correctly predict the trend.

### 2.1.2 Simple Average Method (SAM):

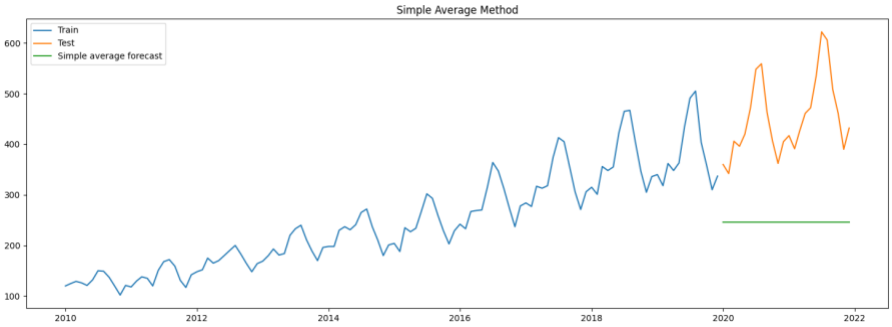


Figure 6: Simple Average Forecast for Train and Test Set

This model shows clearly, what effect outliers can have on prediction of a trend, only using a simple forecasting method like SAM. As the model averages out the values within a range, outliers which is within the range of averages, but are very apart from the common averaged data points, result in an off-putting forecast as seen in the figure above, where the forecast line is way apart from the actual trend.

### 2.1.3 Simple Moving Average Method (SMAM):

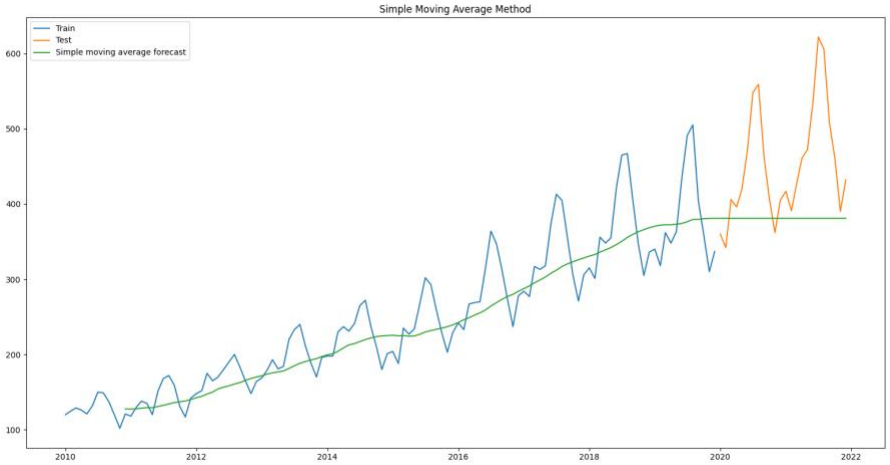


Figure 7: Simple Moving Average Method forecast over a window period of 12 years

For Simple Moving Average method, we added a window period of 12 years and fit the model through the test and training sets. SMAM's method forecasts the results more accurately than both our previous models, as indicated by the forecast line. However, the accuracy of forecast on training set is much higher than the test set. Hence, we concluded there has to be ways different from simple forecasting methods which predicts the increasing trend of our dataset, and in the best case scenario, also takes into account, the changing seasonality while forecasting.

## 2.2 ARIMA Models

### 2.2.1 Autoregressive method (AR):

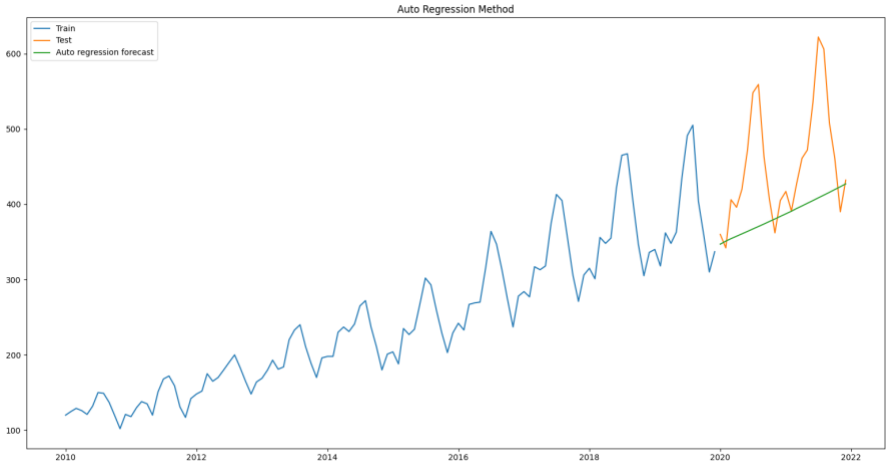


Figure 8: Auto regressive model for a period of 24 months

The AR method forecasts a good trend for our dataset over a period of the last 24 months, i.e., the test sets period. However, since it's a strictly increasing regression model, it does not consider, the shifting trends and only predicts a linearly increasing trend over time.

### 2.2.2 Seasonal Auto Regressive Integrated Moving Average (SAIMA):

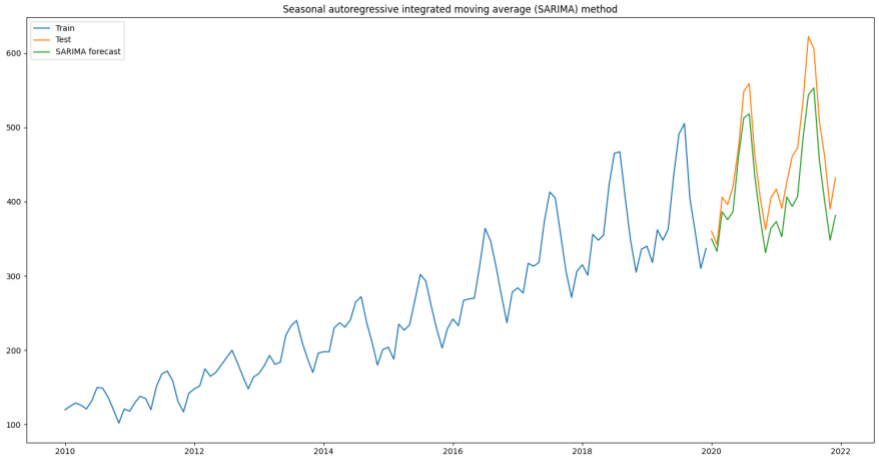


Figure 9: Auto regressive model for a period of 24 months

To fix the problem with the previous ARR forecast, SAIMA model was applied to the dataset. This model considers, the positive and negative shift in trend and seasonality over time and gives a far more accurate forecast of the changing trend in the dataset. This the best model applied to the dataset to this point of the experiment.

## 2.3 Exponential Smoothing Methods

### 2.3.1 Simple Exponential Smoothing Method (SESM):

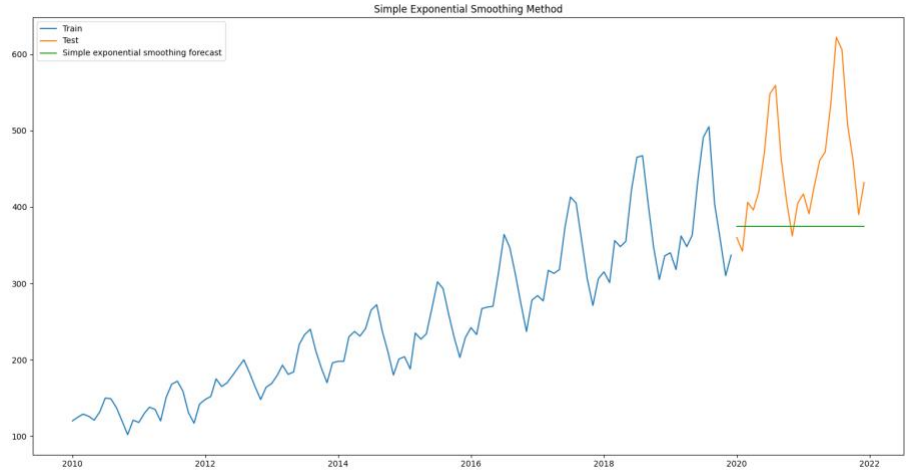


Figure 10: Simple Exponential Smoothing Forecast

Having forecasted the values on the training and test set for SMAM, a window size of 24 months was set, to predict the trend for data. The conclusion was that, since the dataset is relatively small, the model does not have enough information to predict an accurate enough trend and smoothen out over time. However, applying this model on a larger dataset is likely to have a better result.

### 2.3.2 Holts Method with Trend (HMWT):

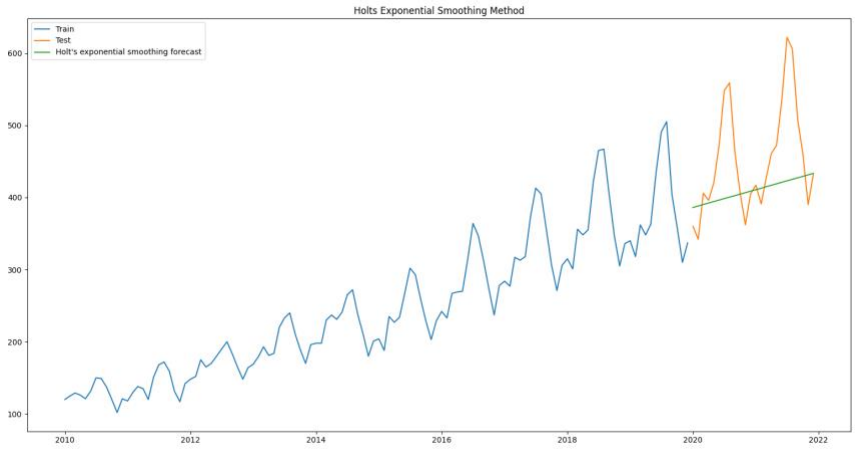


Figure 11: Holts linear trend forecast

As seen in the figure above, the result is positively affirming. Even on a smaller dataset like this, a trend begins to emerge. Using a seasonal period of 12 months, with an additive trend, over a total period of 24 months, it can be seen that Holts method forecasts the result quite well. However, this is still a positively regressive model and does not forecast the steep increase and decrease in seasonality, even though it gives a better result than the previous models.

### 2.3.3 Holts Winter's Multiplicative method with trend and seasonality (HWMM):

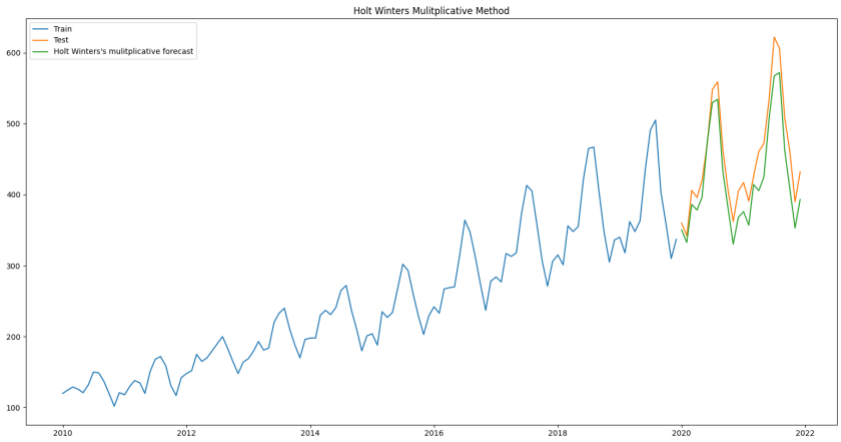


Figure 12: Holts Winters Multiplicative Forecast with Seasonality and Trend

This is the best result from the conducted experiments. As seen, this method forecasts the trend with a very close approximation of the shifting seasonality, but overall increasing trend in the dataset in a promising way. This is because this method is best for the datasets, where the trend and seasonality are increasing over time. Thus, a curved forecast reproduces the seasonal changes as seen in the figure above.

The metrics used in the dataset to measure the accuracy of the models, were Root Mean Square Error (RMSE) and Mean Absolute Percentage errors, the two prominent accuracy metrics for computing the accuracy of time series predictions. The results are shown in Table 1.

Table 1: RMSE and MAPE of the models

| Method | RMSE | MAPE |
|---|---|---|
| NM | 137.51 | 23.63 |
| SAM | 219.40 | 44.23 |
| SMAM | 103.33 | 15.54 |
| AR | 95.71 | 14.18 |
| SARIMA | 42.86 | 8.39 |
| SESM | 107.65 | 16.49 |
| HESM | 82.31 | 11.57 |
| HWMM | 33.72 | 6.71 |

**Conclusion:**

To summarize, the models with the highest RMSE and MAPE scores, are the Simple Forecasting Models. This is because, these models do not take the causal factors into account and are heavily affected by seasonal outliers. The models which can forecasts an increasing trend, but does not consider, the steep changes in seasonality, give better results than the Simple Forecasting Models but the accuracy is moderate. The RMSE and MAPE scores are the lowest for the models that seem to predict an increasing trend and considers, the changes in seasonality, and therefore as seen from the results, it is apparent that the best model is the Holts Winter Multiplicative Model with Seasonal Trends, followed by Seasonal Auto Regressive Integrated Moving Average Model. There are more advanced models, that are suitable for a bigger data set which includes models like LSTM, N-BEATS and Temporal Fusion Transformer but may not give the best output for limited dataset like ours.

Code Availability:
https://github.com/Siryouka/CSCE822_Final_Project.git

**Declaration of Contribution:**

The contribution of each team member:
Sambit Mukherjee: 50%
Lingjia Shi: 50%

We had weekly team meetings and discussed the various models to be implemented, preprocessing strategies, and methods to calculate accuracy and coded it together after bouncing off ideas.
We also wrote the report and created the PowerPoint together.

## References


1. Samuel B. A Glance at Tourism Economics over the last decade. Mondes du Tourisme [Online], 18, 2020, URL: http:// journals.openedition.org/tourisme/3366; doi: https://doi.org/10.4000/tourisme.3366
2. Coshall, J. T., & Charlesworth, R. A management orientated approach to combination forecasting of tourism demand. Tourism Management, 2010, 32(4), 759–769.
3. Yao, Y., Cao, Y., Ding, X., Zhai, J., Liu, J., Luo, Y., & Zou, K. A paired neural network model for tourist arrival forecasting. Experts Systems with Applications, 2018, 114, 588–614. doi: 10.1016/ j.eswa. 2018.08.025
4. Claveria, O., & Torra, S. Forecasting tourism demand to Catalonia: Neural networks vs. time series models. Economic Modelling, 2014, 36, 220–228. doi:10.1016/j.econmod.2013.09.024
5. Gounopoulos, D., Petmezas, D., & Santamaria, D. Forecasting tourist arrivals in Greece and the impact of macroeconomic shocks from the countries of tourists' origin. Annals of Tourism Research, 2012, 39(2), 641–666. doi:10.1016/j.annals.2011.09.001
6. Andrew, W.P., Cranage, D.A., Lee, C.K. Forecasting hotel occupancy rates with time series models: an empirical analysis. Hosp. Res. J. 1990, 14 (2), 173–182.
7. Kim, J.H., Wong, K., Athanasopoulos, G., Liu, S. Beyond point forecasting: evaluation of alternative prediction intervals for tourist arrivals. Int. J. Forecast. 2011, 27, 887–901.
8. Chen, K.-Y. Combining linear and nonlinear model in forecasting tourism demand. Expert Syst. Appl. 2011, 38, 10368–10376.
9. Noersasongko, E.; Julfia, F.T.; Syukur, A.; Purwanto; Pramunendar, R.A.; Supriyanto, C. A Tourism Arrival Forecasting using Genetic Algorithm based Neural Network. Indian J. Sci. Technol. 2016, 9.
10. Subedi, A. Time Series Modeling on Monthly Data of Tourist Arrivals in Nepal: An Alternative Approach. Nepal. J. Stat. 2017, 1, 41–54.
11. Dhakal, Chuda. A Naïve Approach for Comparing a Forecast Model. 2018.
12. Hansun, Seng. A Novel Research of New Moving Average Method in Time Series Analysis. International Journal of New Media Technology (IJNMT). 2014, 1. 22.

13. Zhihua Wang, Yongbo Zhang, Huimin Fu, "Autoregressive Prediction with Rolling Mechanism for Time Series Forecasting with Small Sample Size", Mathematical Problems in Engineering, 2014, Article ID 572173, 9 pages, 2014.
14. Brownlee, Jason. “A Gentle Introduction to Sarima for Time Series Forecasting in Python.” MachineLearningMastery.com, 21 Aug. 2019, https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/.
15. BROWN, R. G.: Statistical Forecasting for Inventory Control, McGraw-Hill: New York, 1959.
16. HOLT, C. C.: Forecasting trends and seasonals by exponentially weighted averages, O.N.R. Memorandum 52/1957, Carnegie Institute of Technology, 1957
17. LI, Z. P. – YU, H. – LIU, Y. C. – LIU, F. Q.: An Improved Adaptive Exponential Smoothing Model for Short Term Travel Time Forecasting of Urban Arterial Street, Acta automatica sinica, 2008, 34, 11,1404– 1409
18. Ostertagová, & Ostertag, O. Forecasting using simple exponential smoothing method. Acta Electrotechnica et Informatica. 2012, 12(3), 62–66. https://doi.org/10.2478/v10198-012-0034-2
19. Khairina, Daniel, Y., & Widagdo, P. P. Comparison of double exponential smoothing and triple exponential smoothing methods in predicting income of local water company. Journal of Physics. Conference Series. 2021, 1943(1), 12102–. https://doi.org/10.1088/1742-6596/1943/1/012102