Forecasting K-pop Popularity:

Application of Time Series Analysis

1. Introduction

For some fans the expansion of K-pop in the American market was a big step towards a broader audience and recognition. This analysis focuses on the popularity of the K-pop music genre as a whole rather than the comparison of popularity of groups. Though admittedly speaking that would have been an interesting topic to study as well. To see which groups will do well in the long run and which will not. However, as there are many groups to consider, with some not being in the business for long, we choose the music genre as our focal point. Why? As a fan, we would like to see how it would grow in the American market, which have been dominating the music scene for so long.

In order to model K-pop popularity, we pulled data from Google Trends. Google Trends is an analytical tool that allows users to generate the popularity of search terms over time. Data is available from 2004 to the present, and we chose to use data from January of 2010 to December of 2020. In this study, we included only the popularity in the United States. Also, we will not be taking into account any other variables as we could not justify how the variable can affect the data.

The objective of this study is to forecast K-pop popularity using univariate time series forecasting models in order to efficiently predict the trend popularity of the music genre.

2. Data and Description

The data was sourced from the Google Trends website. This data shows how the popularity of a term has changes over time in Google searches. We looked at the specific search term “K-pop” under the music genre. The data we generated from January 2010 to December 2020 was available at the monthly level, giving 132 observations at the time of writing. We filtered the data down to searches from only the United States. The trends are scored using relative index of 0-100, with 100 being the point at which the term being looked at peaked in popularity. A value of 50 is 50% as popular as the peak.

3. Model

Box and Jenkins introduced ARIMA models in 1970. This type of models encompasses three classes of models, the Autoregressive (AR), Moving Average (MA) and Autoregressive Moving Average (ARMA) models. The general shorthand notation for SARIMA model is ARIMA(p, d, q)(P, D, Q)$_s$ where p = order of non-seasonal AR term, q = order of the non-seasonal MA term, d = order of non-seasonal differencing, P = order of the seasonal AR term, Q = order of the seasonal MA term, D = order of seasonal differencing, and s = number of season per year for monthly data.

The equation of the ARIMA(p, d, q)(P, D, Q)$_s$ model is given as

$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{h=1}^{q} \theta_h W_{t-h} + \sum_{j=1}^{P} \Phi_j X_{t-js} + \sum_{l=1}^{Q} \Theta_l W_{t-ls} + W_t$

where c is an intercept term and depends if the series has a non zero mean or not; ϕ_i , θ_h, Φ_j, and Θ_l are the coefficients of the non-seasonal AR, non-seasonal MA, seasonal AR, and seasonal MA terms respectively.

SARIMA models depend on the pattern of the autocorrelation and partial autocorrelation functions and are based on stationary, model identification, estimation, and forecasting. If the original series is not stationary, non-stationarity is removed by differencing or using linear regression. The Augmented Dickey Fuller (ADF) test is used for checking stationary condition.

4. Forecasting Performance Measures

The models discussed above will be fit for the monthly K-pop data and then the best model will be selected. The overall performance of the models fitted will be measured using the following measures of forecasting performance:

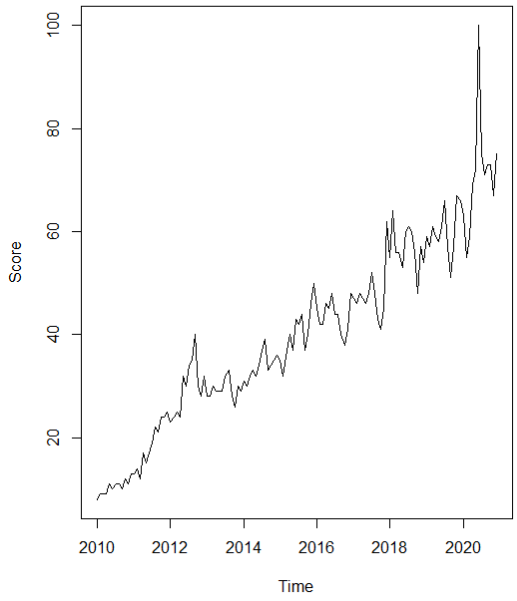
1. Mean Square Error (MSE)
2. Root Mean Square Error (RMSE)
3. Mean Absolute Error (MAE)
4. Mean Absolute Percentage Error (MAPE)
5. Akaike Information Criterion (AIC)
6. Bayesian Information Criterion (BIC)

The best model is the ones that has the lowest AIC and BIC and the lowest (minimum) values in the remaining four criteria. The four criteria involving errors will be used specifically to measure forecast accuracy.

5. Results and Discussion

i. Stationarity

Initially, we plotted our data to see if there are any trends and/or seasonality that is apparent. In the figure below, it can be seen that there is an increasing trend that is clear. Hence, this is not stationary.

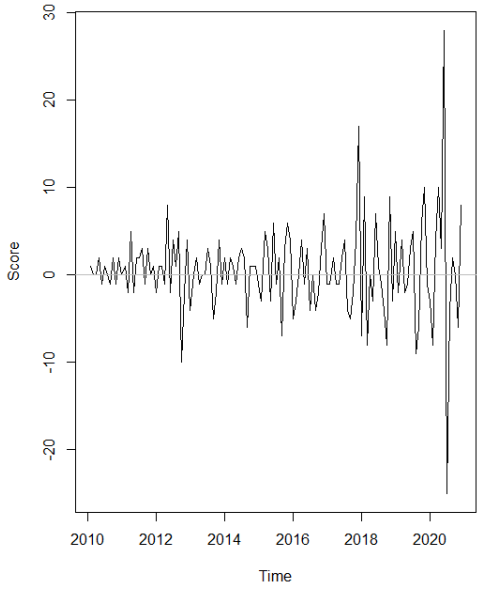


The plot also shows signs of seasonality. There are humps that are more apparent in the plot staring from 2017 and so on. The starting values of this humps also coincide with the beginning of the years; hence, seasonal models will be considered in the analysis as well.

Since the data is not stationary, we first have to make it as one. Hence, for this study, we used two kinds of de-trending: fitting a linear model and taking the first difference.

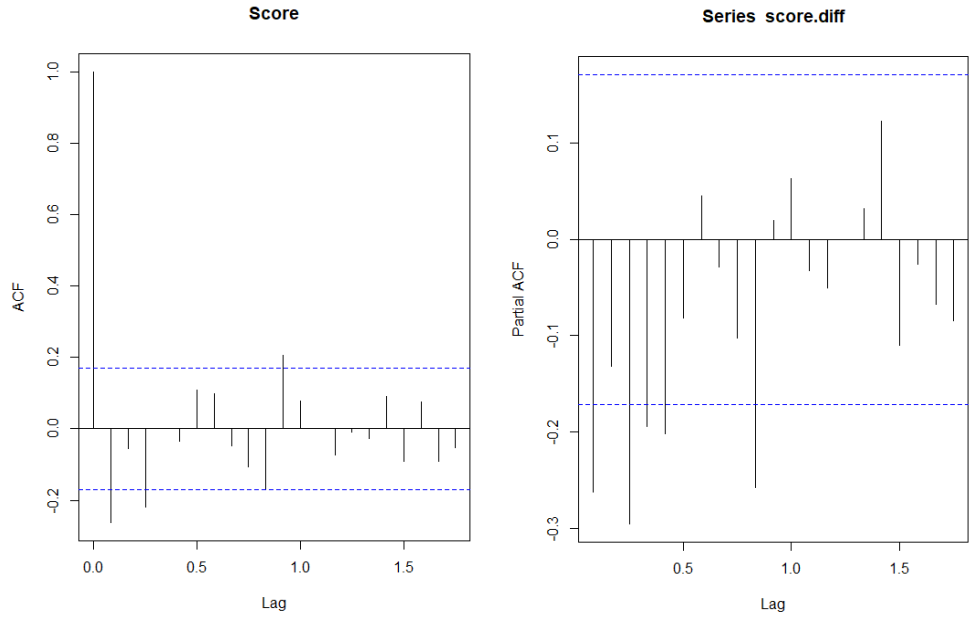
Taking the ADF test, we see that after removing a linear trend and an intercept that the data is still not stationary. The p-value is 0.1886 and, hence, we could not reject the null hypothesis that the data is not stationary. Since the ADF test does not infer that the data would be stationary after removing the linear trend, it is decided that the we would de-trend the data by taking the first difference.

Taking the first difference of the data and plotting it, we can see from the graph below that it looks approximately stationary, at least the mean and the variance are more-or-less constant. Hence, de-trending by taking the first difference is good enough for this data.



ii. Arima Models

Taking the ACF and PACF functions of this differenced score, the autocorrelations lying outside of the noise line are of lags 1, 3, and 11 for the ACF and 1, 3, 4, 5, and 10 for the PACF. Lag 3 seems to be the most significant looking at the PACF graph.



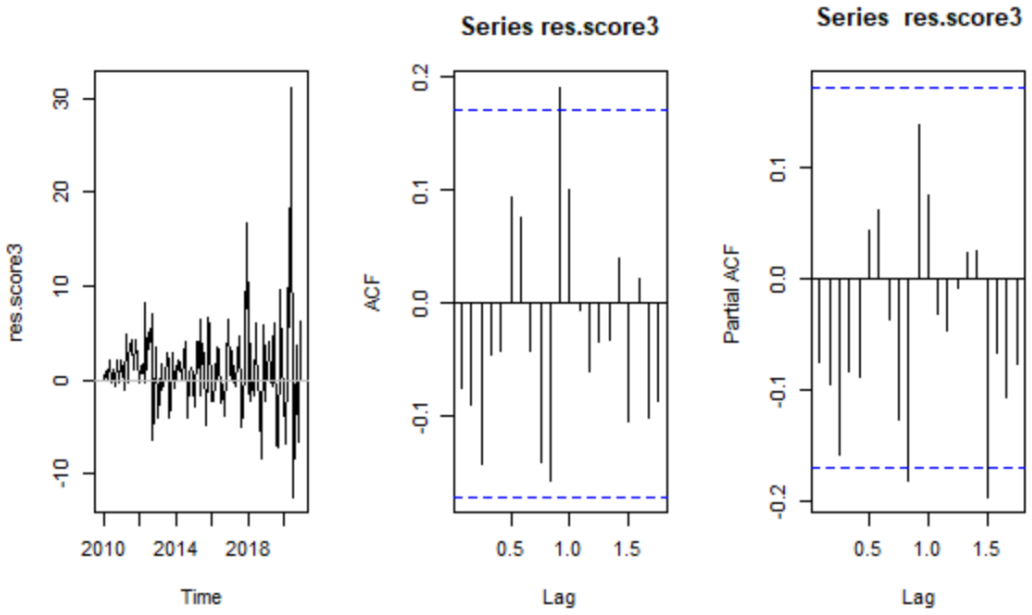
Looking at both graphs above, we can infer that since the ACF graph seems to go to zero immediately and that the PACF graph goes to zero after the third lag, this could be a moving average process. To check this, we tried some Arima models for the

differenced scores where p and q are given the order 3. The following table shows the results:

Model	AIC	BIC	ME	RMSE	MAE	MAPE
Arima(0,1,0)	810.89	813.76	0.508	5.284	3.477	8.425
Arima(3,1,0)	796.62	808.12	0.871	4.884	3.214	7.934
Arima(0,1,3)	**795.68**	**807.18**	**1.165**	**4.867**	**3.212**	**7.975**
Arima(3,1,3)	787.91	808.04	0.994	4.517	3.012	7.528

From this, we can see that there are two ARIMA models that we can use. There is the Arima(0,1,3) as it has the minimum values for both AIC and BIC; and the Arima(3,1,3) model as it has the minimum values for the error. Though if we were to pick one model, we would choose Arima(0,1,3) as it has the lower number of parameters.

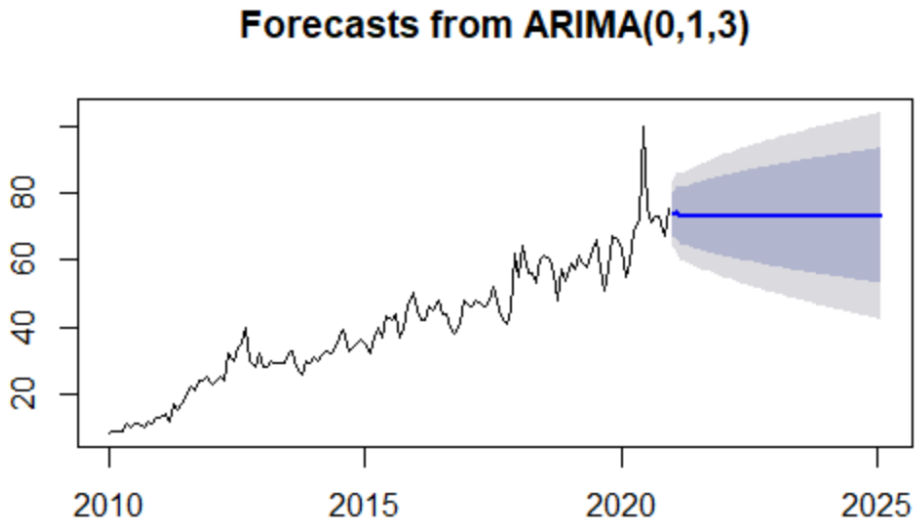
Taking the residual of the Arima(0,1,3) model and plotting it, we can see from the figure below that the data while may not look like a white noise, it kind of looks like it. Some of it lies near zero, though most are above zero. However, the ACF and PACF graphs has some lags that are significant, specifically at 10. Suggesting that the residuals may be correlated.



To analyze the independence of residuals, we used the Box-Pierce and Ljung-Box tests. Using the Box-Pierce Test to test the correlations among the residuals of the differenced scores, we get a small p-value of 0.0292. Hence, we could reject the null hypothesis that the residuals of the differenced scores are independent of each other.

We also get the same conclusion using the Ljung-Box method with a p-value of 0.02652. Hence, Arima(0,1,3) is not a good fit for our data.

Since we already know that Arima(0,1,3) is not a good fit and looking at the forecast plot below shows why. It should be going up, yet it went flat. The forecast tells us nothing about the possible scores of the data. For this reason, a seasonal component might be necessary.



iii. SARIMA Models

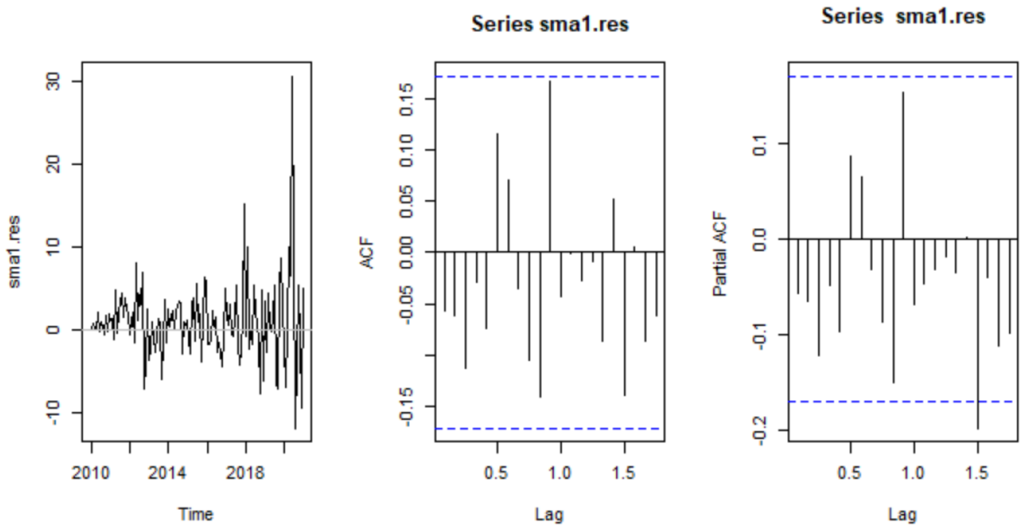
Now, at the beginning of this study, we said that there might be a seasonality in our times series. Hence, we also tested some seasonal models for our differenced scores data. Using the same order for the non-seasonal part of the SARIMA model, we tested some seasonal models when the period is 10 and 12. We chose 10 as looking at the graphs of the ACF and PACF above, it is always above the noise line; and 12 because we are looking at monthly data. The results of our seasonal models are below:

Model	AIC	BIC	ME	RMSE	MAE	MAPE
$Arima(0,1,3)(1,0,0)_{10}$	795.34	809.72	1.208	4.819	3.164	7.869
$Arima(0,1,3)(0,0,1)_{10}$	795.03	809.4	1.231	4.812	3.175	7.903
$Arima(0,1,3)(1,0,1)_{10}$	797.02	814.27	1.229	4.182	3.173	7.899
$Arima(0,1,3)(1,0,0)_{12}$	**792.58**	**806.95**	**0.979**	**4.757**	**3.162**	**7.863**
$Arima(0,1,3)(0,0,1)_{12}$	794.09	808.47	1.053	4.793	3.182	7.901
$Arima(0,1,3)(1,0,1)_{12}$	789.43	806.68	0.799	4.623	3.047	7.671

The best model for this is $Arima(0,1,3)(1,0,0)_{12}$. It has the minimum values for AIC and BIC and $Arima(0,1,3)(1,0,1)_{12}$ has the minimum values for the errors. Nonetheless, we chose this as it has the lower number of parameters.

The graph of the residuals of this model shows that it looks like white noise. More importantly though, the ACF and PACF graphs lies most under the noise line. 10 is no longer a significant lag in this model. To make sure that we can use this model, we analyzed the residuals. Using the Box-Pierce method, we got a p-value of 0.06532. Hence, we could not reject the null hypothesis. Hence, we could not assume that the residuals are dependent. We also get the same conclusion using the Ljung-Box method. Therefore, this is the first model that could possibly be a fit for our time series.

This model is: $X_t = 0.2617X_{t-12} - 0.4189W_{t-1} - 0.0554W_{t-2} - 0.1352W_{t-3}$

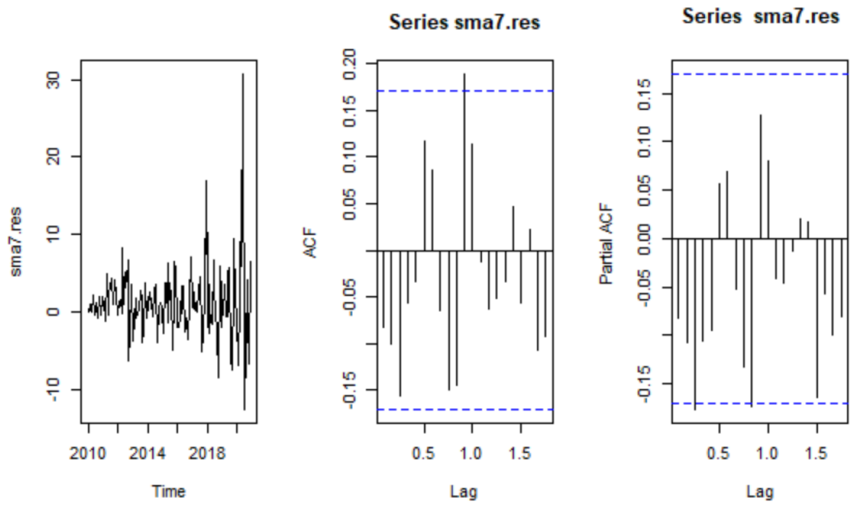


Interestingly, if we look at the graph of the PACF of this model, we can see that there is a significant lag at order 18. Even though we tried several Box-Pierce and Ljung-Box method to test the independence of the residuals, we still get the same conclusion that they are uncorrelated. This is even after testing with varying increasing values of lag.

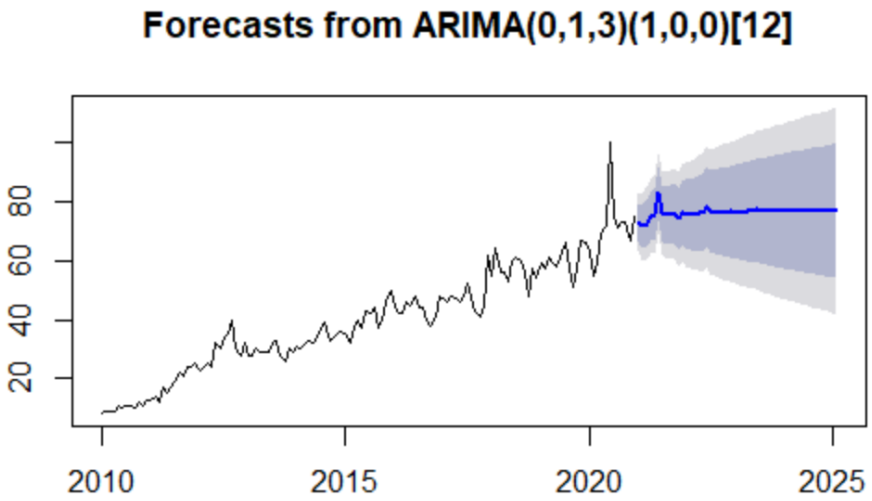
Hence, to see if maybe there is a seasonality at every 18^{th} period, we also tried a few seasonal models for it:

Model	AIC	BIC	ME	RMSE	MAE	MAPE
$Arima(0,1,3)(1,0,0)_{18}$	**797.02**	**811.4**	**1.218**	**4.852**	**3.214**	**7.949**
$Arima(0,1,3)(0,0,1)_{18}$	797.8	811.56	1.207	4.857	3.216	7.959
$Arima(0,1,3)(1,0,1)_{18}$	798.67	815.92	1.216	4.843	3.203	7.921

Comparing these results to the one above, we can conclude that seasonal models of period 12 is still better. In fact, if we get the ACF and PACF graphs for the residuals, below, of the best model for the 18^{th} period, $Arima(0,1,3)(1,0,0)_{18}$, we get that there are correlated residuals. This is backed up by the Box-Pierce and Ljung -Box methods as both have small p-values.



Let's go back to our analysis of the seasonal model of period 12, Arima(0,1,3)(1,0,0)$_{12}$. Looking at the forecast plot 50 months ahead of this model, it can be seen that there is a fluctuation that we expected from the seasonal component, but it also decays off immediately. Going farther, the variance spreads out and the forecasts stays constant.



iv. Models with Drift

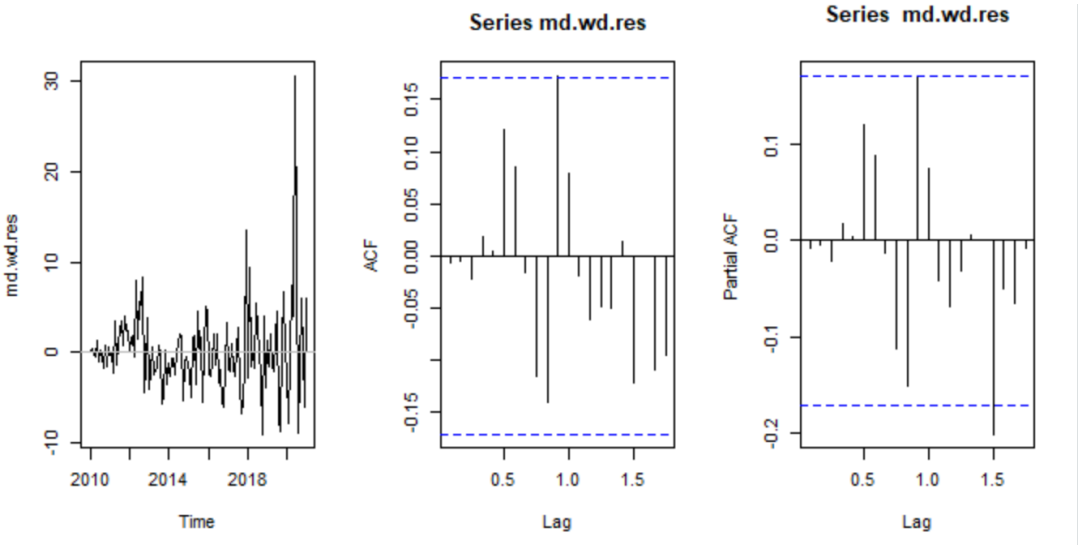
Trying more possible models for our time series data, we decided to see what happens if we added a drift to the models that we have tried so far. The following table shows the results:

Model	AIC	BIC	ME	RMSE	MAE	MAPE
Arima(0,1,3) with drift	**783.12**	**797.49**	**0.047**	**4.588**	**3.034**	**7.631**

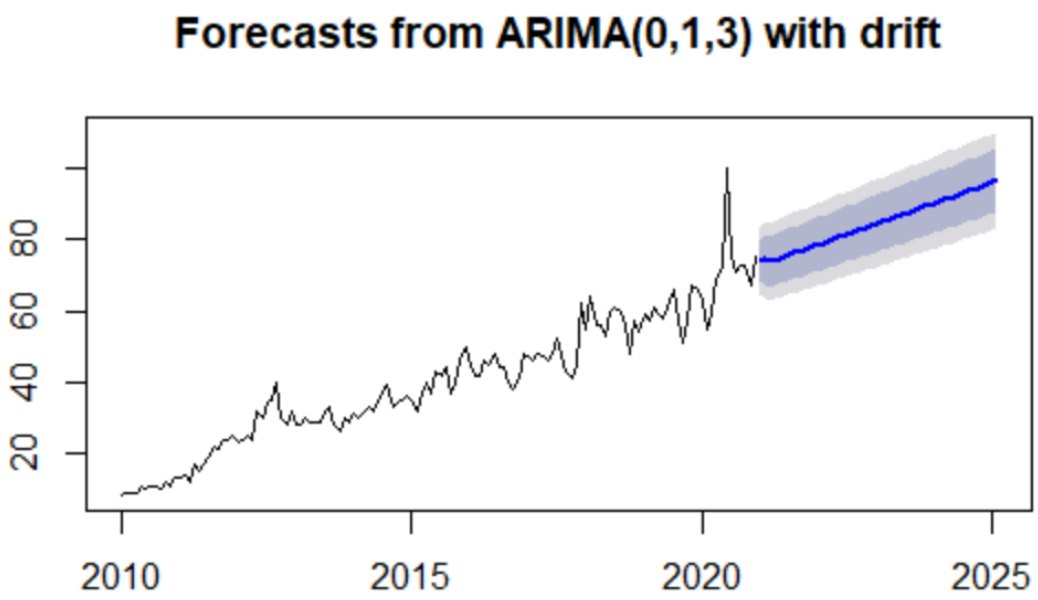
Arima(0,1,3)$(1,0,0)_{12}$ with drift	783.68	800.93	0.019	4.562	3.016	7.583

From the table above, Arima(0,1,3) with drift is the best model. As it has the minimum values for AIC and BIC. Though the other model has better forecast accuracy, we chose this model as it has lower number of parameters. Interestingly, the residuals of this model are independent compared to when there is no drift. Performing both the Box-Pierce and Ljung-Box methods, we got p-values 0.7294 and 0.7241 respectively. We performed these methods by fitting five lags as this model has four parameters. Increasing the values of the lag gives the same conclusion. Also, the ACF and PACF graphs shows that most of the lags are below the noise line, though the lag 11 seems unambiguous. Nonetheless, performing the box test function for both models with drift, we found that both have residuals that are independent. Therefore, our time series data now has a second possible model: Arima(0,1,3) with drift.

This model is: $X_t = 0.04835 - 0.4640W_{t-1} - 0.1763W_{t-2} - 0.2318W_{t-3}$



Since we have to choose one model to fit our time series data, we have to compare which model, Arima(0,1,3)$(1,0,0)_{12}$ or Arima(0,1,3) with drift, is best. As both models have the same number of parameters we have to compare them it terms of AIC, BIC, and forecast accuracy. In terms of AIC and BIC, Arima(0,1,3) with drift has the minimum values. It also has minimum values for the errors. Hence, for this time series data, we should fit it with a Arima(0,1,3) with drift model.



The forecast plot of this model 50 months into the future shows an increase in scores that we expected. Though there are no fluctuations, but that makes sense as we did not fit a seasonal component. If we were to forecast 500 months into the future (plot on the right), we will see that the variance does not blow up.

v. Auto Arima

Performing the Auto Arima function to our time series data, we got the same conclusion as above: Arima(0,1,3) with drift.

6. Conclusion

The main focus of this paper was to use time series data from Google Trends to predict the future popularity for the search term "K-pop". The models we used were Arima(0,1,3), Arima(0,1,3)$(1,0,0)_{12}$, and Arima(0,1,3) with drift. The Arima(0,1,3) with drift model provided a more better fit compared to Arima(0,1,3) and the Arima(0,1,3)$(1,0,0)_{12}$. Seeing the outcome from the forecast plots showed us a lot of interesting points about K-pop and its growing popularity in the American audience. Though in a sense, we should question the outcome generated from this study as we did not take into account any other variables. In a sense, as a fan we should take the results of our forecast plots with a grain of salt.