Bayesian Time Series Analysis

Gerardo R.

2023-11-05

Abstract. By using a smoothed Non-Dynamic Linear Model (DNLM) based on two components: an increasing linear trend and a yearly-seasoned period we have fitted a time-series model for the monthly Google searches performed for the term "cough" between the years 2004-2023. Our prediction shows that in 2025 the number of searches will likely surpass the current maximum, which happened in 2020.

Introduction and data description. The goal of this project is to predict how many hits will have the word "cough" in January 2025. Will it be higher than in January 2020? The data represents the number of Google searches registered between the years 2004-2023 of the term "cough". As we can see in the data the word "cough" has not done other thing that to be more and more searched while the time passed. This increase seems to have a pattern all years but in 2019-2023, which could (only could!) be related with the impact that covid19 had in the world's media.

Searches for "cough" in Google 2004–2023
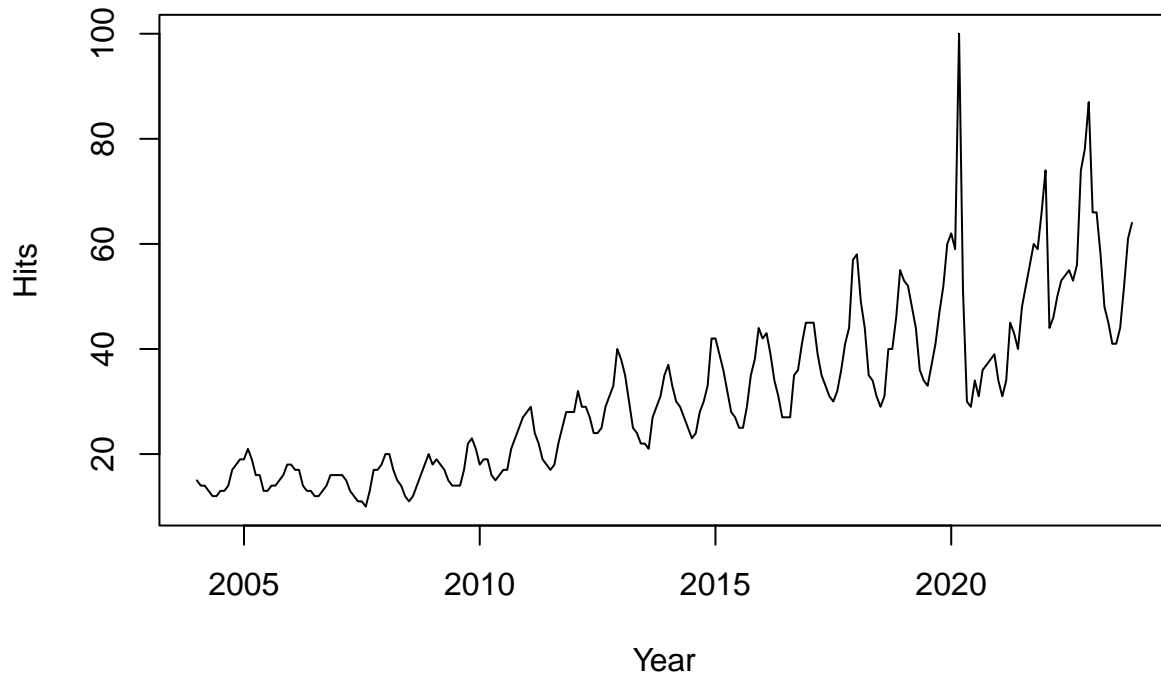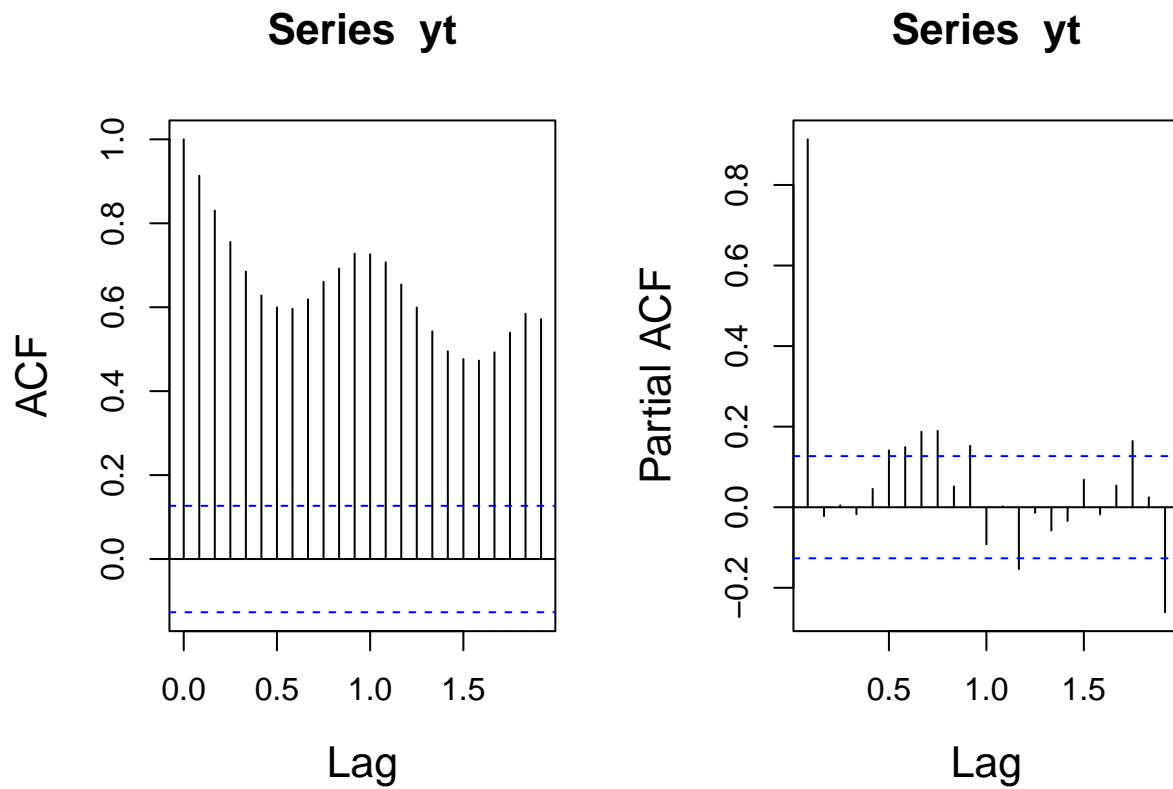
Data exploration. The ACF plot shows significant correlations between time series in all lags (so the model we are making will be with all time series). The plot also presents a decay, suggesting the presence of a **linear trend**. On the other hand, based on the decaying PACF plot we conclude there is **seasonality** in the data.

Data preprocessing. Based on our data exploration we decided to detrend (diff) and deseason (smoothering) the data. Based on

Model proposal

- The data shows non-stationarity behavior (no quasy-periodicity), thus we would prefer to use the Normal Dynamic Linear Model (NDLM) over the autorregression (AR).

- Based on ACF and PACF plots we need to create a linear component and a season component to model our data. We will use the polynomial of order two to capture this double behaviour.
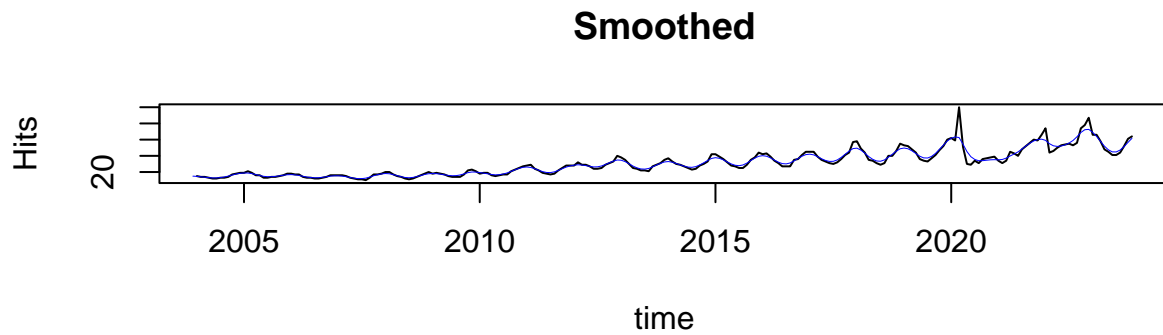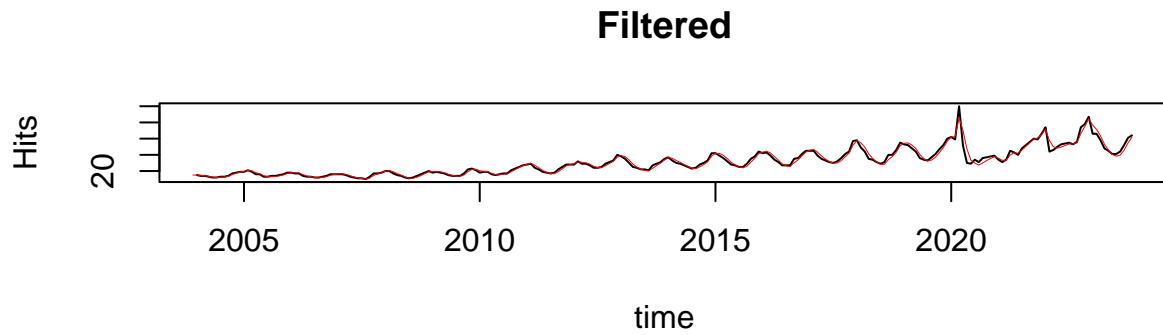
We are assuming a NDLM model where:

$y_t = F_t \phi_t + v_t$ being $v_t \ N(0, v_t)$

$\phi_t = G_t \phi_{t-1} + w_t$ being $w_t \ N(0, W_t)$

Additionally here are some details related with model values:

- For model trend component: dV = 10 and dW = 1
- For model season component: dV = 10, dW = 1, m0 = 15, C0 = 1

Filtered

Smoothed

Goal completion: Forecasting for 2025 (and conclusion)

In order to answer my original question I will try to forecast the year 2025 (that is, 12 points/months in advance). The predicted section within a 95% Confidence interval is depicted in blue:

Forecasting: Searching of "cough"
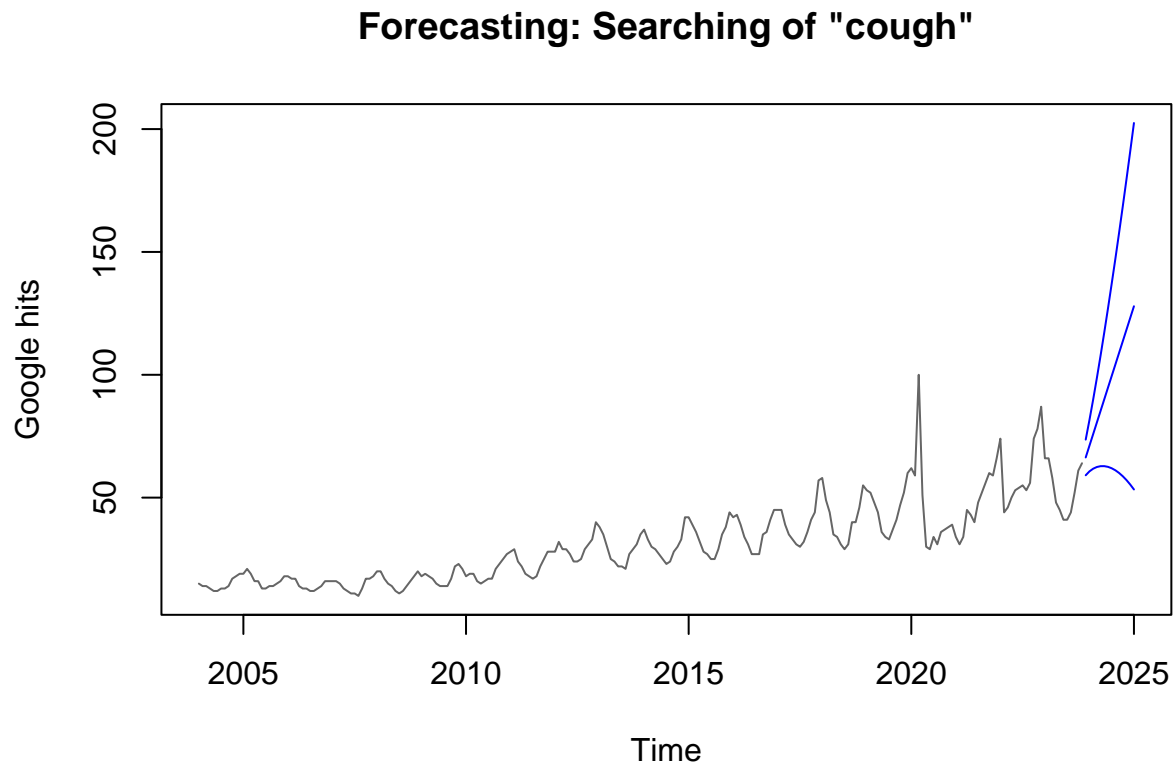
Conclusion. Based on this graph considering a 95% CI that the more likely situation based on our prediction is that in January 2025 the current maximum number of hits for the tearm "cough" will be surpassed. However, based on a 95% CI this does not seem to be certain.

Some extra notes for replication

Software details. The whole project was run using the R package: dlm. For replication purposes, you can just introduce the parameters specified in the model section. The data obtaiend was downloaded through the R package: gtrends. All code written was done using R programming.

Thank you for taking your time to review!

I tried to be brief and explicit. Hope it was easy to follow! :)