Time Series Class Final Exam - Winter 2022

Zach Horton

1. Soccer Search Trend Analysis

Below we show a plot of the monthly Google trends index for the term "soccer" ranging from Jan. 2004 to Mar. 2022.

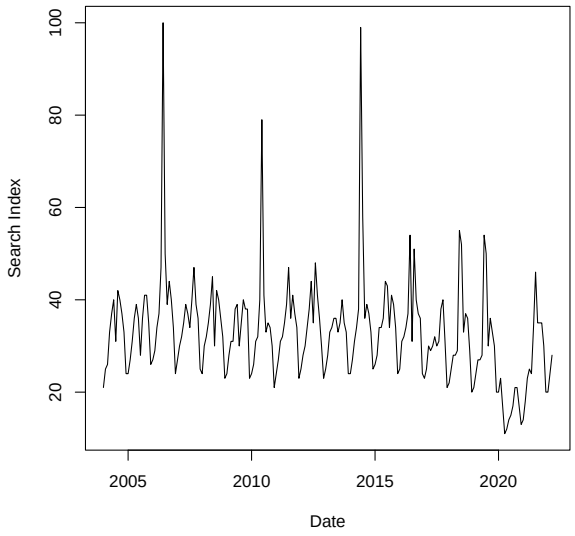

Figure 1: Monthly search index values for "soccer", comrpised of 219 observations from Jan. 2004 to Mar. 2022.

These data serve as a rough measure of global interest in soccer. Our goal in this section is to model and understand the patterns behind general soccer interest. Just from visual inspection, we see several patterns including an annual cycle with overall raised interest in the summer with peaks in June and August, and a 4-year cycle of major spikes in June around the time of the world cup. We also see a disruption to the otherwise consistent pattern in 2020 due to the pandemic.

1.1 DLM Construction

A natural choice to model the search index data is a dynamic linear model (DLM), considering it's time-based evolutionary nature. In this section we discuss possible DLM components and ultimately settle on a structure for analysis. Recall that a general normal DLM is given by the following form:

$$y_t = F_t'\boldsymbol{\theta}_t + \epsilon_t, \quad \epsilon_t \sim N(0, v_t)$$
$$\boldsymbol{\theta}_t = G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(\mathbf{0}, W_t)$$

which can be summarized by the quadruple $\{F_t, G_t, v_t, W_t\}$ and where the dimension p describes the length of $\boldsymbol{\theta}$. For our purposes, we only consider the set of regression time-series DLMs which have $G_t = G$ fixed and known and F_t known for all t. Additionally, we handle W_t through discount factors and we assume $v_t = v$ is fixed but unknown, all of which we discuss later.

In this framework, our task is to construct G and F_t such that the implied forecasts aligns with the data. We do this via the superposition principle, meaning we collect desirable canonical models components G_j^* and aggregate them into $G = \texttt{blockdiag}(G_1^*, G_2^*, \ldots)$.

1.1.1 Seasonal Component

As a preliminary study of seasonality, consider the data periodogram below:

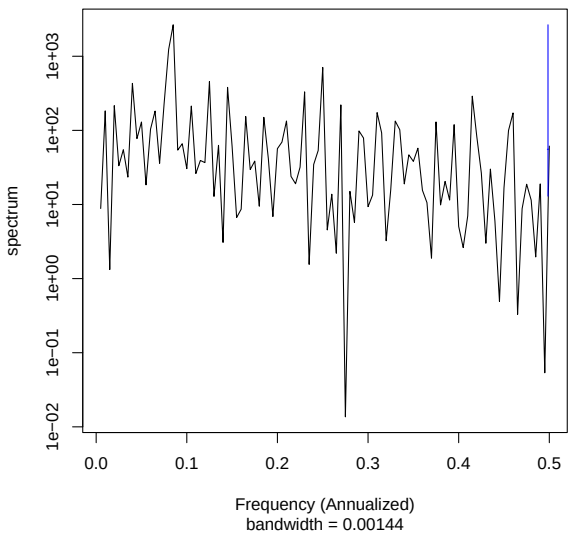

Figure 2: Periodogram of search index data.

Both the annual cycle and the 4-year cycle are easy to spot in the data. The periodogram shows

that the annual frequency is most pronounced, but doesn't clearly indicate other specific patterns. The below ACF plot may be helpful as well:

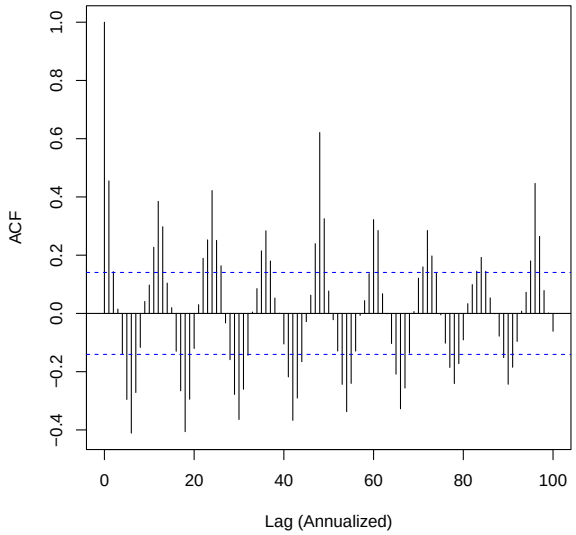

Figure 3: ACF of search index data.

This also shows a clear annual cycle. Interestingly, the 4-year autocorrelation is somewhat elevated above the pattern, but not in a cyclical way. For completeness, take a look at the PACF plot:

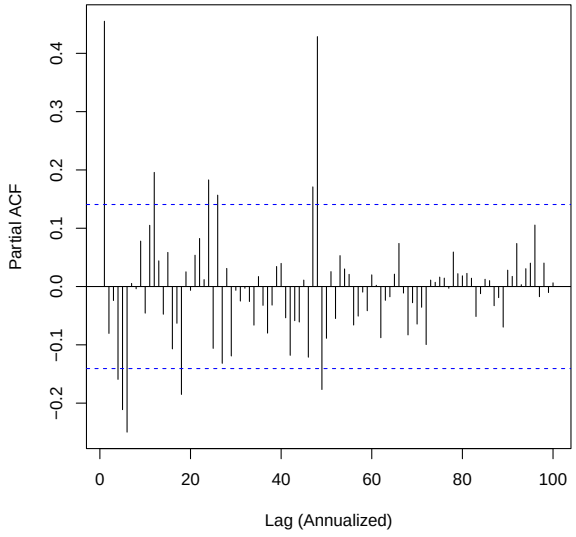

Figure 4: PACF of search index data.

Again we see an annual cycle at play, plus an anomaly at 4-years. This suggests that "cycle" may be the wrong description of the pattern. Indeed, it seems as though the annual cycle dominates, but every 4th year there is an outlier caused by the world cup.

Rather than the more standard 4-year seasonal effects model, which is likely too flexible, we model the annual cycle and the 4-year pattern separately. We treat the 4-year pattern as a regression, where the covariate is an indicator variable, but we discuss this later. We proceed here with some details on the annual seasonal model. Recall that the full Fourier representation of a 12-period seasonal effect is given by:

$$G^* = \begin{bmatrix} J_2(1,\omega) & & & & \\ & J_2(1,2\omega) & & & \\ & & \ddots & & \\ & & & J_2(1,5\omega) & \\ & & & & -1 \end{bmatrix}$$

where $J_2(1,\omega) = \begin{bmatrix} \cos(\omega) & \sin(\omega) \\ -\sin(\omega) & \cos(\omega) \end{bmatrix}$, where multiples of $\omega = \frac{2\pi}{12}$ span the Fourier frequencies, and where $F^* = (1, 0, \ldots, 1)'$. Also recall the Fourier representation property that each seasonal factor can be expressed as:

$$\psi_j = \sum_{r=1}^{6} \theta^*_{2r-1} \cos(rj\omega) + \theta^*_{2r} \sin(rj\omega)$$

where $\boldsymbol{\theta}^*$ is the state-vector for the Fourier form DLM $\{F^*, G^*, \cdot, \cdot\}$ and where $\theta_{12} = 0$. In a later section we consider removing insignificant harmonics, but we begin by considering all 6.

The final aspect our seasonal model is a discount factor. In the updating equations, we have $R_t = G_t C_{t-1} G_t' + W_t$ where W_t is the system variance matrix. A discount factor is employed to simplify $R_t = \frac{1}{\delta_S} G_t C_{t-1} G_t'$ where $\delta_S \in (0, 1]$ is the seasonal discount factor. These are a computationally tractable way of including uncertainty about the system variance, with values very close to 1 indicating little to no evolution over time.

1.1.2 World Cup Regression Component

As mentioned previously, modeling the world-cup outlier effect seen in the search index data

may be more naturally handled by a regression model. Specifically the component we consider is very simple, having $G^* = [1]$ and $F_t^* = x_t$ where $x_t = 1$ indicates the world cup is occurring that month, and otherwise $x_t = 0$. We also assign a separate discount factor δ_W to handle evolution of the world-cup effect.

One reason to prefer this structure over a seasonal one is because the upcoming world cup is not occuring in June, but rather November. This breaks seasonal models, but the indicator regression handles it with ease. The other reason has to do with small sample size. Only 4 scheduled world cups appear in the data. By separating its effect from the seasonal model, we can reduce noise coming from this somewhat "rare" event.

It is worth noting that this component brings our composite super position model out of the time-series DLM space because the implied vector F_t is not constant over time. However, it is the only component with time-varying vector F_t^*.

1.1.3 Trend Component

Looking at the data plot reveals additional variability beyond annual seasonality and the outlier effect. Indeed, there are periods where it appears the whole process is gradually rising or falling, not to mention the epidemic effect. By including a trend component we not only capture those movement, we also enable assessment of the deseasonalized patterns of general soccer interest.

The component we add takes the form $F^* = (1, 0)'$ and

$$G^* = J_2(\lambda = 1) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

We chose $\lambda = 1$ for two reasons. First, the ACF plot doesn't show any obvious exponential decay, so it didn't seem that $|\lambda| < 1$ would be appropriate. Second and perhaps more importantly, it corresponds to the linear trend or second-order polynomial model which is a powerful framework. It suggests that forecasts are linear, which seems to be reasonable given the data shape.

As with the other components, we incorporate a discount factor δ_T, which will allow us to gauge how much evolution variance appears in the process.

1.2 Model Selection

In this section we discuss selecting harmonics from the seasonal component and overall discount factor selection. Our approach to use AIC to simultaneously optimize all elements.

To summarize the previous section, our initial model is built using the super position principle. The vector F_t is given by:

$$\begin{aligned} F_t &= (F_T, F_W, F_S)' \\ &= (1, 0, x_t, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1)' \end{aligned}$$

and the matrix G is given by:

$$\begin{aligned} G =& \texttt{blockdiag}(G_T, G_W, G_S) \\ G_T =& \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \\ G_W =& [1] \\ G_S =& \begin{bmatrix} J_2(1,\omega) & & & & \\ & J_2(1,2\omega) & & & \\ & & \ddots & & \\ & & & J_2(1,5\omega) & \\ & & & & -1 \end{bmatrix} \end{aligned}$$

for a total parameter count of $p = 14$. In a similar fashion we handle the system covariance W_t through discount factors. Let $P_t = GC_{t-1}G'$ and let $P_t[j:k]$ denote the $(j-k)\times(j-k)$ sub block of P_t with only rows and columns $j, \ldots, k$. Then we set the time t system covariance by:

$$\begin{aligned} W_t =& \texttt{blockdiag}(W_{t,T}, W_{t,W}, W_{t,S}) \\ W_{t,T} =& (\frac{1}{\delta_T} - 1)P_t[1:2] \\ W_{t,W} =& (\frac{1}{\delta_W} - 1)P_t[3:3] \\ W_{t,S} =& (\frac{1}{\delta_S} - 1)P_t[4:14] \end{aligned}$$

where $\delta_T, \delta_W, \delta_S \in (0, 1]$ are to each be optimized over a grids of points $(0.9, 0.901, \ldots, 0.999, 1)$. Finally, recall that we consider fixed but unknown observation variance $v_t = v$. Thus our

DLM can be expressed by $\{F_t, G, v, W_t\}$ plus the reference prior[1].

In this framework, the goal of model selection is to choose the discount values and decide whether any of the blocks of G_S should be dropped. To proceed with model selection, we provide the likelihood:

$$L(\delta_T, \delta_W, \delta_S, \boldsymbol{I}|\boldsymbol{y}) = \prod_{t=15}^{n} p(y_t|D_{t-1})$$

where $\boldsymbol{I} = (I_1, \ldots, I_6)$ is the configuration vector indicating which seasonal harmonics remain and where $p(y_t|D_{t-1})$ is the usual Student-t density arising from the update equations, with mean $f_t = F_t' G \boldsymbol{m}_t$ and variance $Q_t = F_t'(P_t + W_t)F_t + S_{t-1}$, with degrees of freedom $n_t = n_{t-1}+1$ and data standard deviation $S_t = S_{t-1} + \frac{S_{t-1}}{n_t}\left(\frac{(y_t - f_t)^2}{Q_t} - 1\right)$.

The reference prior dictates that the distribution for $t \leq p$. We use the corresponding reference updating equations until observation $p+1$, where we set $n_{p+1} = 1$ and begin using the usual update equations. Of course, removing harmonics will reduce p, but we do not adjust the likelihood in order to keep the amount of data constant between comparisons.

To perform optimization we maximize the likelihood but introduce a penalty to account for differences in model complexity. The expression to be optimized amounts to a shifted AIC:

$$\text{AIC} = \log(L(\delta_T, \delta_W, \delta_S, \boldsymbol{I}|\boldsymbol{y})) - 2\sum_{j=1}^{6} I_i$$

1.2.1 Optimization Results

Given the unexpected deviations caused by the pandemic in March 2020, we perform the optimization under two scenarios. The first is under "normal" conditions where we only consider data up through Feb 2020. The second is the full dataset including the pandemic dip. The results are given below, noting that $I_i = 1$ indicates the ith harmonic is included:

[1]Due to sparsity of world cup events, we add 1 to the 3rd diagonal element of K_t for $t = 1$ which prevents singular update matrices. Consequently, the world cup effect is unreliable until the first actual event.

Table 1: Optimized Model Configurations

Parameter	δ_T	δ_W	δ_S	I_1	I_2	I_3	I_4	I_5	I_6
Subset	0.948	0.759*	1.000	1	1	1	1	1	0
Full Data	0.938	0.756*	1.000	1	1	1	1	1	0

The optimal configuration in both cases removes the 6th harmonics corresponding to the 2-month cycle, leaving the rest. Both settings also have no discount for the seasonal components, suggesting no seasonal evolution over time. The biggest difference is in the trend discount δ_T, with the full data fit permitting more system variability. Although the optimal world-cup discounts δ_W were around 0.75, we fix it to 0.9. This is because of the small data issue with only 4 world cup events in the data. Overfit is already an issue for this, but we curb it slightly by restricting the discount factor.

1.3 Filtering-Based Inference

Given these configurations we fit the models again, this time decreasing the model dimension from 14 to 13 to account for the removed harmonics. This is easily accomplished simply be deleting their contributions to the blocks of G and the components of F_t. In this section we present various posterior results and contrast the effect that COVID had on the model fit.

1.3.1 Observational Variance

The updating equations we used involve Student-t distributions for the observation level, which is equivalent to marginalizing out the unknown variance v under normally distributed observations and an inverse-gamma prior. Thus, given the update quantities for the final time point $t = T$, the posterior for v is of the form:

$$v \sim IG\left(\frac{n_T}{2}, \frac{n_T}{2} S_T\right)$$

This provides means for inference. Results are displayed in the table below.

These results are somewhat what we would expect. The effect of the pandemic was unanticipated, adding unexplained variance to the sys-

Table 2: Observation Variance Posterior

Quantity	n_T	Mean	95% Low	95% High
Subset	180	17.54	14.27	21.56
Full Data	208	24.80	20.42	30.10

tem. In the next section we examine the fits in more detail.

1.3.2 Trend Terms

Under our DLM model, $\theta_{1,t}$ corresponds to the deseasonalized process level, $\theta_{2,t}$ is the process slope, and $\theta_{3,t}$ is the world-cup effect. In this section we look at the filtering distributions for each. Below are the $\theta_{1,t}$ estimates:

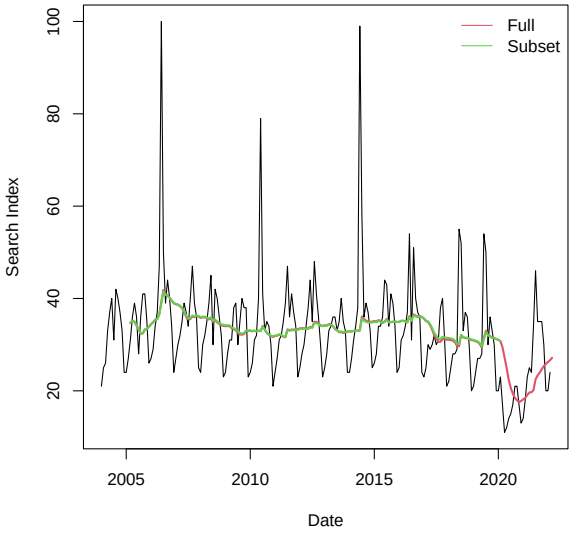

Figure 5: Estimated Trend Levels

The filtered trend estimates are very similar in the two cases, with just a little added variability in the full-data case.

In Figure 6 we plot corresponding 95% interval bands. The intervals show slightly less variance overall in the subset model. Overall, these suggest that general interest in soccer is somewhat constant, but took a big hit during the pandemic.

Next we look at the slope term and associated intervals in Figure 7. Both data sets tell the same story here: before the pandemic, general soccer interest was not growing. In fact, the only points with significant trend slope are the pandemic and

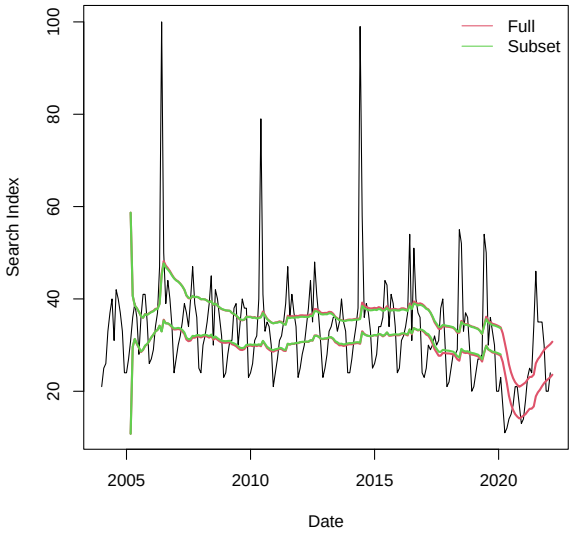

Figure 6: Interval Band for Trend Levels

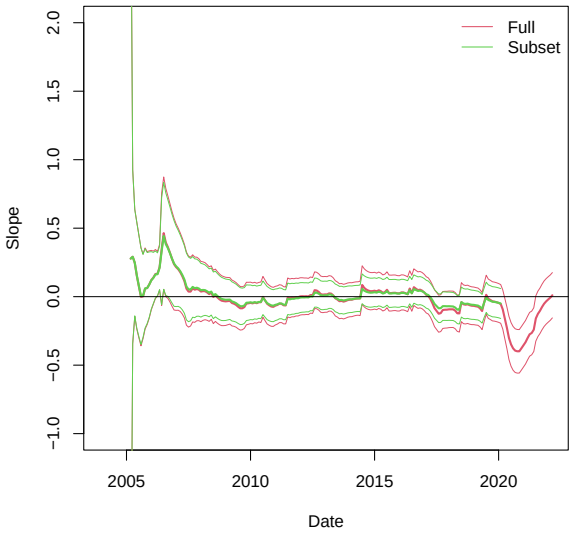

Figure 7: Trend Slope Estimates and Intervals

maybe brief periods around world cups. And as before, the pre-pandemic model estimate has less variability. The need for a dynamic estimate is obvious once the pandemic arrived, so this result makes sense.

The final trend element to consider is the world-cup effect. Unlike the other components, this functioned like a regression parameter. Figure 8 below shows it's estimate and 95% bands for the two data scenarios.

It is important to note that this term caused

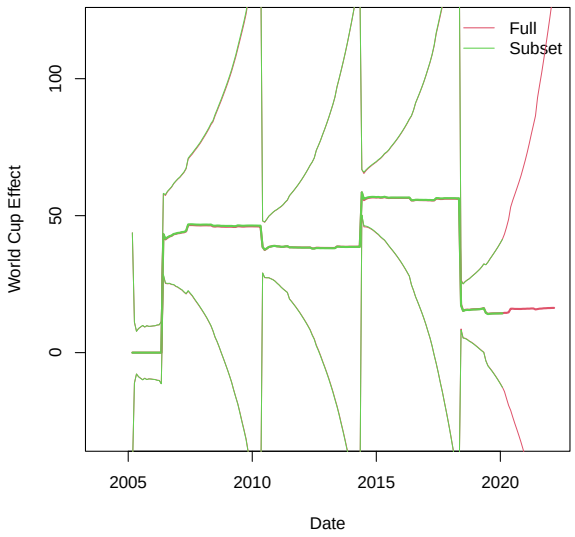

Figure 8: World Cup Effect Estimates and Intervals

computational issues since it isn't observable until the first world cup observation. It is also obvious that the small number of world cup data points leads to overfitting. However, its inclusion leads to more stable estimates for the trend level and slope, which is fitting.

Overall, both data scenarios produced very similar trend estimates and uncertainty bands, the most notable difference being in the slope term.

1.3.3 Seasonal Terms

Now we turn attention to the seasonal terms. Recall that we started with a 12-month seasonal factors model, written in Fourier harmonic DLM representation with 6 total harmonics, and that through model selection we removed the 6th harmonic. The terms $\theta_{4,t}, \theta_{6,t}, \theta_{8,t}, \theta_{10,t}, \theta_{12,t}$ correspond to the effects of the 1st-5th harmonics. The estimated aggregate seasonal effect with corresponding 95% band is given in Figure 9. The estimated seasonal patterns are very similar, both in effect and uncertainty.

We note that although the discount $\delta_S = 1$, evolution is still expected for the filtering distributions. In the smoothing inference section we see the static seasonal components.

Decomposing the seasonal effect into it's harmonic components reveals a similar story. Figure 10 shows all 5 harmonics displayed together, and they are extremely similar between both fits. From an amplitude standpoint, the annual pattern is most prominent. We will examine these more closely to determine significance.

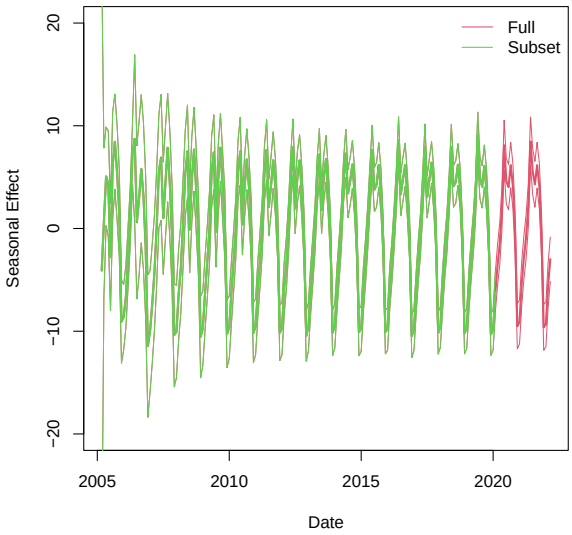

Figure 9: Aggregate Seasonal Effect with 95% Band

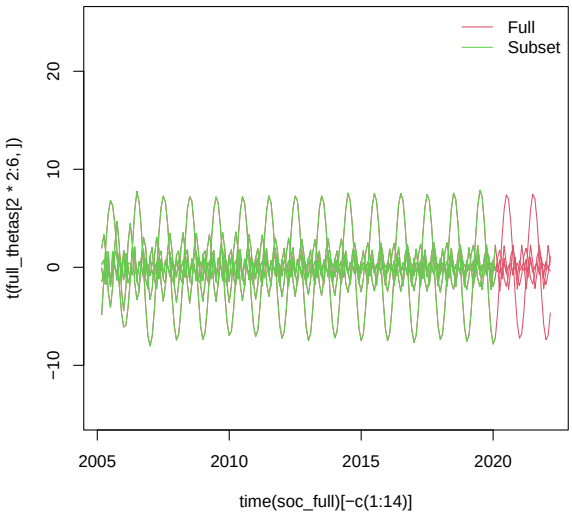

Figure 10: Estimated Harmonic Effects

Specifically, we will look at the effect significance of the harmonics at each observed time. This is done by examining the estimates at each time t of $\theta_{4,t}, \theta_{6,t}, \theta_{8,t}, \theta_{10,t}, \theta_{12,t}$ and their corresponding uncertainty. Figure 11 below shows the plot of the 1st harmonic ($\theta_{4,t}$) and correspond-

ing uncertainty intervals. This pattern appears significant, as would be expected given our data structure. Overall the effects are similar between the full-data fit and the subset data fit.

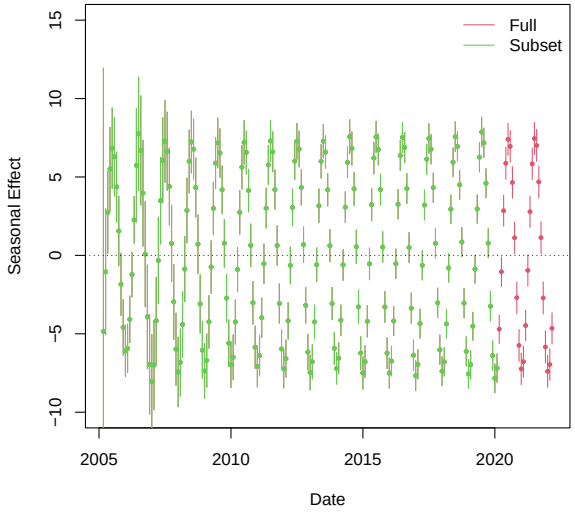

Figure 11: Estimated 1st Harmonic (annual cycle) Effects with 95% Intervals

Now we look at the 2nd harmonic ($\theta_{6,t}$), which produces a 6-month cycle and has second highest amplitude. The plot below shows the corresponding estimates.

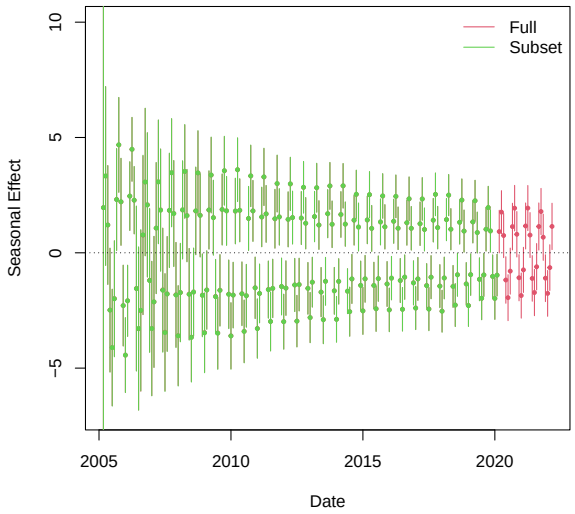

Figure 12: Estimated 2nd Harmonic (6-month cycle) Effects with 95% Intervals

This harmonic also appears significant. The two data scenarios are extremely similar. Next we look at the 3rd harmonic ($\theta_{8,t}$).

Figure 13: Estimated 3rd Harmonic (4-month cycle) Effects with 95% Intervals

This harmonic is overall significant, but it appears that only every-other month effect is significant. As before, the seasonal effect looks similar in the two scenarios. Next, in the plot below we look at the 4th harmonic ($\theta_{10,t}$) which defines a 3-month cycle.

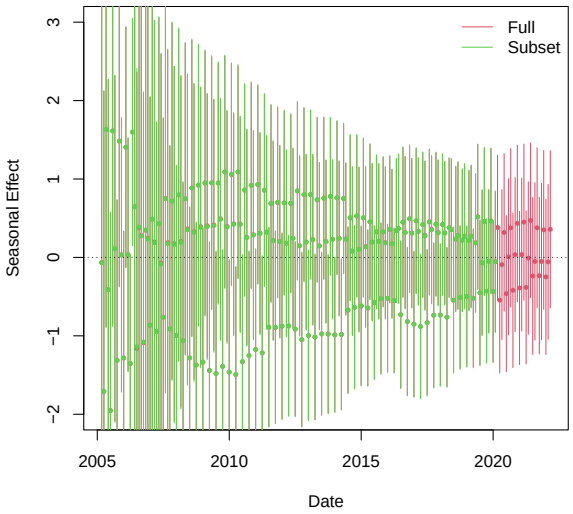

Figure 14: Estimated 4th Harmonic (3-month cycle) Effects with 95% Intervals

It may not be the case that the 3-month cy-

cle component is significant. Only during a few periods of time is one of the three month effect different from zero. Further, the post-pandemic time definitely does not treat this as significant. Lastly, we look at the 5th harmonic in $\theta_{12,t}$ which defines a 2.4-month cycle.

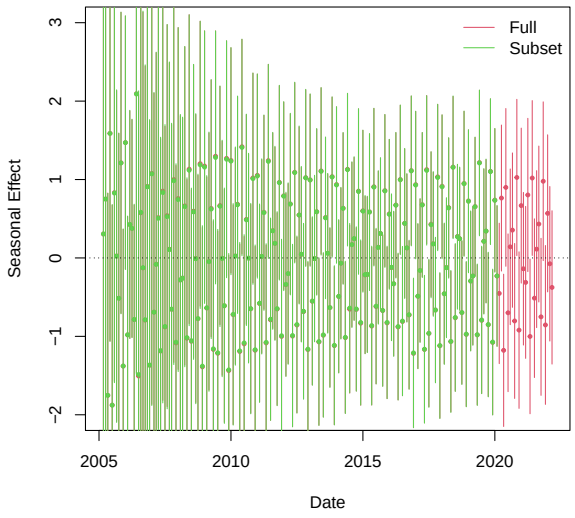

Figure 15: Estimated 5th Harmonic (2.4-month cycle) Effects with 95% Intervals

It is difficult to assess this plot given that the 5th harmonic divides the year into non-integer month segments. However, other than some early uncertainty, this harmonic appears significant since many intervals stay above/below zero over time.

Overall, the model appears to be adequately capturing data features. And the model fits are very similar between the two data cases, which is expected given the similar discount factors and forward propagating nature of the filtering distributions.

1.4 Smoothing-Based Inference

Results in the previous section based on the filtering/update equations are very insightful. However, the smoothing equations can be used to borrow information from later time-points and remove more noise from model parameter estimates. We use the standard smoothing equations from the latest time point $T = 219$ all the way back to time $p + 1$ when filtering distributions first become proper. Due to the redundancy of these results with respect to the overall analysis goal, discussion will mostly be limited to comparison with the matching filter-based plots.

1.4.1 Trend Terms

Figure 16 shows the smooth trend levels and corresponding uncertainty intervals. As expected, overall the pattern is much smoother, but we start to see how the smoothing through the pandemic impacts earlier trend level estimates.

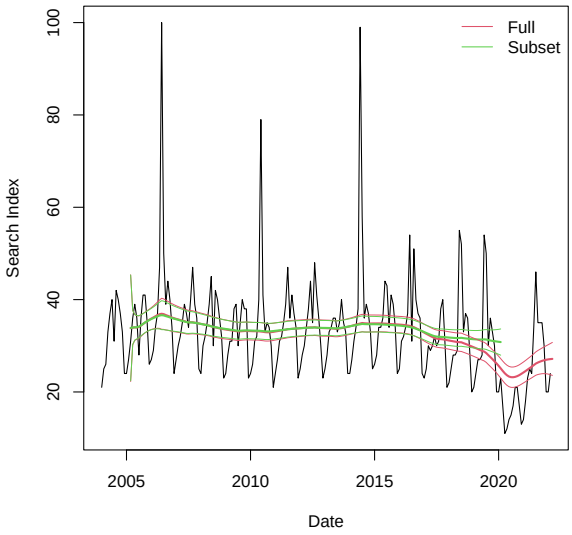

Figure 16: Smoothed Trend Levels

Figure 17 below shows the smoothed estimates and intervals of the trend slope term. Notice that the smoothing has caused the pandemic decline to begin much earlier. A feature that is missing are bumps due to the world cup (except potentially for the first one).

Finally, the smoothed world cup estimates and intervals are below in Figure 18. Here we see that the two models produce slightly different estimates, but this is unsurprising given the computational struggles and small-data overfitting.

Overall, the smoothed parameter distributions tell similar stories as the filtering distributions, just with less variability and uncertainty.

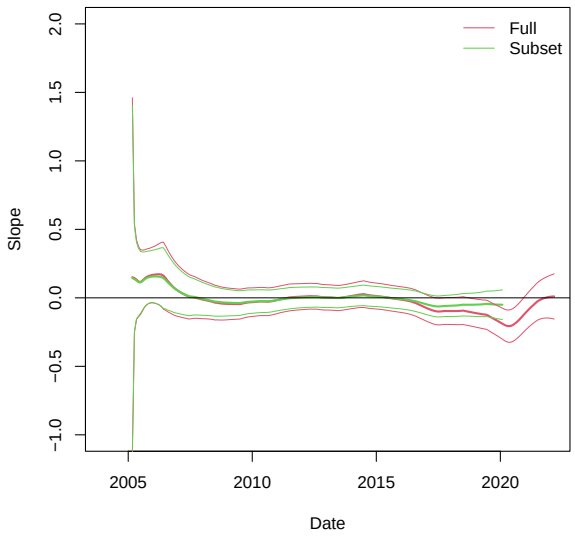

Figure 17: Smoothed Trend Slope Estimates

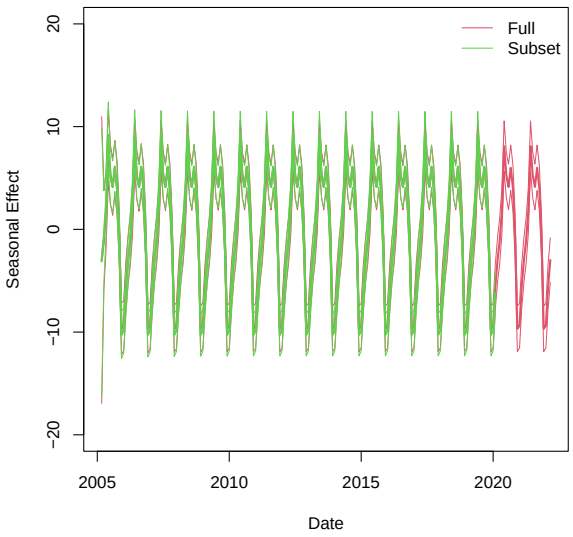

Figure 19: Smoothed Aggregate Seasonal Effect

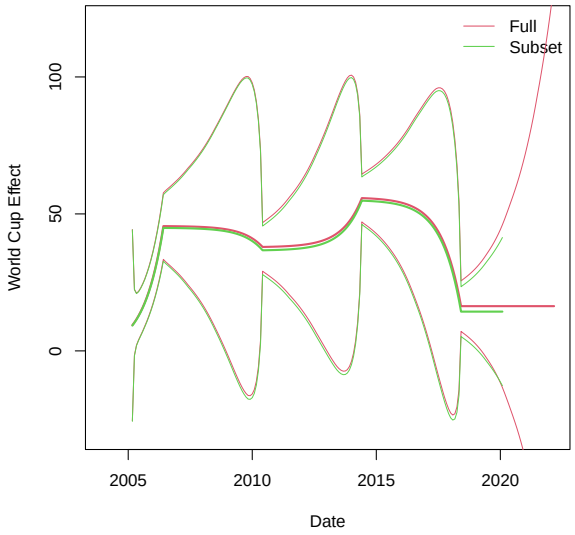

Figure 18: Smoothed World Cup Effect Estimates

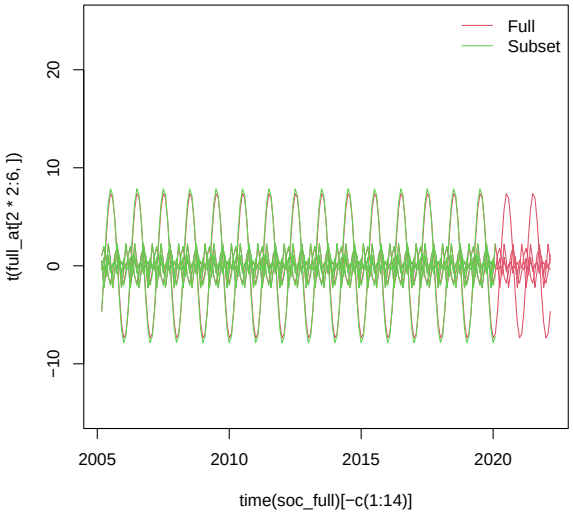

Figure 20: Smoothed Harmonic Effects Estimates

1.4.2 Seasonal Terms

We proceed similarly looking at smoothed seasonal estimates. Figure 19 shows the smoothed aggregate seasonal pattern, which shows the overall similarity between the data scenarios. Also, Figure 20 shows the estimated harmonic decomposition of this pattern. Note that there is no evolution in the pattern due to the static model of $\delta_S = 1$.

Compared to the filtered estimates, some minor discrepancies appear between the data cases, which makes sense considering smoothing for the full-data case takes into account the pandemic. As before, we take a look at each harmonic. We display all 5 in figures 21-25, then highlight notable features.

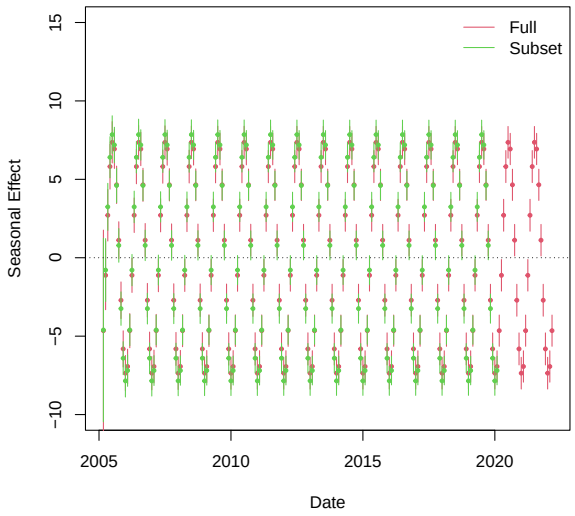

Figure 21: Smoothed 1st Harmonic (annual cycle) Effects with 95% Intervals

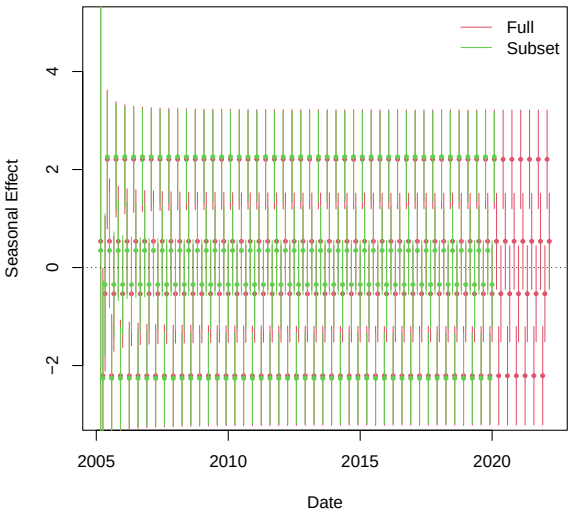

Figure 23: Smoothed 3rd Harmonic (4-month cycle) Effects with 95% Intervals

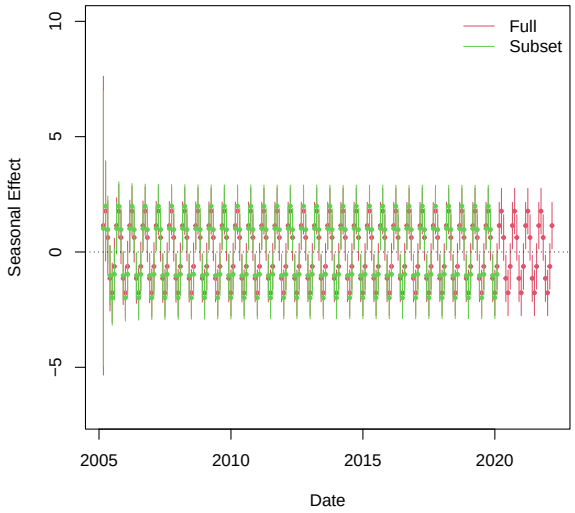

Figure 22: Smoothed 2nd Harmonic (6-month cycle) Effects with 95% Intervals

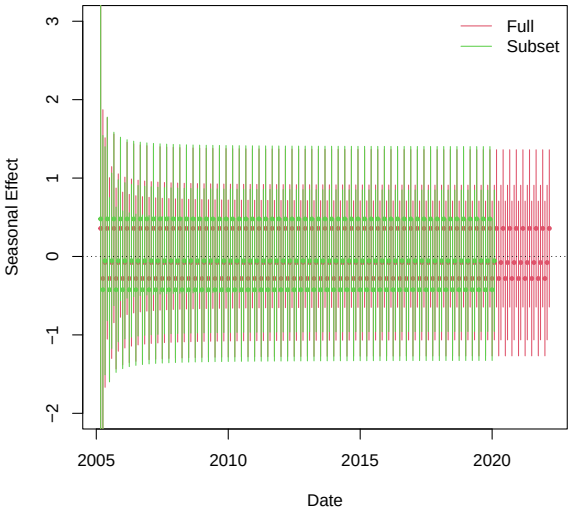

Figure 24: Smoothed 4th Harmonic (3-month cycle) Effects with 95% Intervals

The differences between filtered and smoothed are much more apparent for the seasonal effects, and potentially that is due to the way the pandemic messed with the seasonal components. We also see how the static element in the model leads to constant seasonal effects. Significance is also impacted, namely in 4th harmonic. Overall, it is interesting to see how back-propagating the pandemic data into otherwise very similar models leads to noticeable differences in smoothing.

1.5 Forecasting-Based Inference

It is insightful to consider forecasting as an additional model use-case. We first use the subset data to forecast March 2020 to Feb 2021, and compare/validate against the observed data. Then, we consider forecasting 12 months out from April 2022 to March 2023 using the full data model.

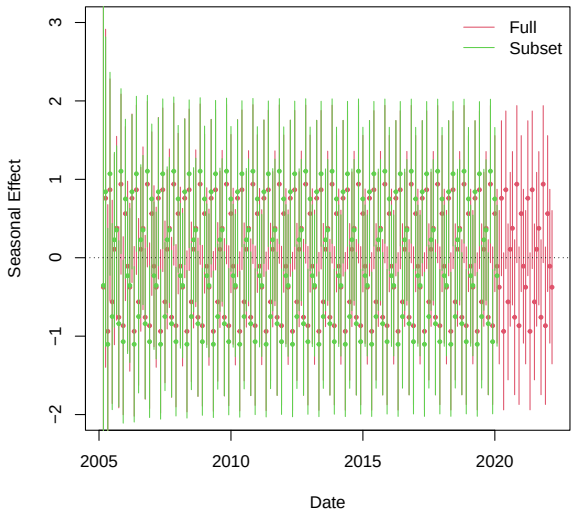

Figure 25: Smoothed 5th Harmonic (2.4-month cycle) Effects with 95% Intervals

1.5.1 Pre-Pandemic Forecast

Forecasting is accomplished by the usual set of equations. In particular, because no world cup occurs within March 2020 to Feb 2021, the vector F_t is constant, and we repeat the discounted variance W_{T+1} for all forecast periods. The plot below shows the predicted and observed patterns with uncertainty bounds.

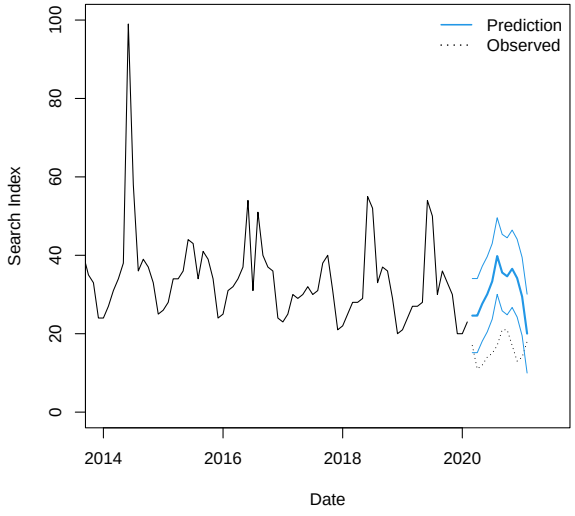

Figure 26: 12-Month Forecast of Subset Data

Clearly, the model did not anticipate the pandemic drop. However it is interesting that the seasonal pattern still seemed to hold somewhat, which lines up with the optimized discount factor $\delta_S = 1$.

As an additional exercise in forecasting, we consider looking 12 months beyond the full data. This is interesting because the 2022 world cup is being held in November, which is a departure from the normal schedule and is included in the forecast period. Our forecast of April 2022 to March 2023 is shown in the plot below, and reflects this event.

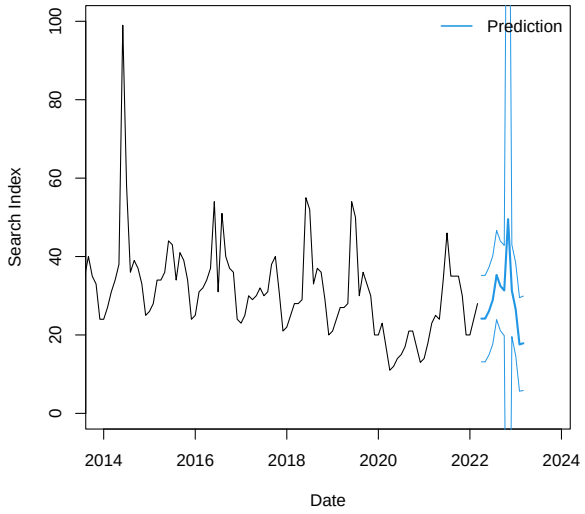

Figure 27: 12-Month Forecast of Full Data

This forecast result includes a world cup effect comparable to the most recent 2018 effect, but the uncertainty around it is huge. That comes partly from the low amount of data defining that effect. But overall it exhibits behavior fitting to the underlying process.

This forecasting problem raises a more general question: when an unprecedented event disrupts an otherwise consistent pattern, how do you adjust the model? There are potentially several available avenues, but the most sensible to me is an "intervention"-based approach. This essentially modifies the filtering equations at a certain point in time, mainly by heavily inflating the variance. This allows a practitioner to manually

intervene when an unexpected deviation occurs. Indeed, this approach seems more natural than building a generalized model which accounts for many types of unpredictable black-swan events. Due to time constraints, we don't implement this here, but we mention it as a possible extension.

1.6 Residual Analysis

Up to this point we have seen a lot of useful inference on model parameters and forecasts. Before proceeding with any sort of conclusion or decision, we do a brief residual analysis to verify our results are grounded in reasonable assumptions. The standardized residuals come from the one-step forecasts in the update equations.

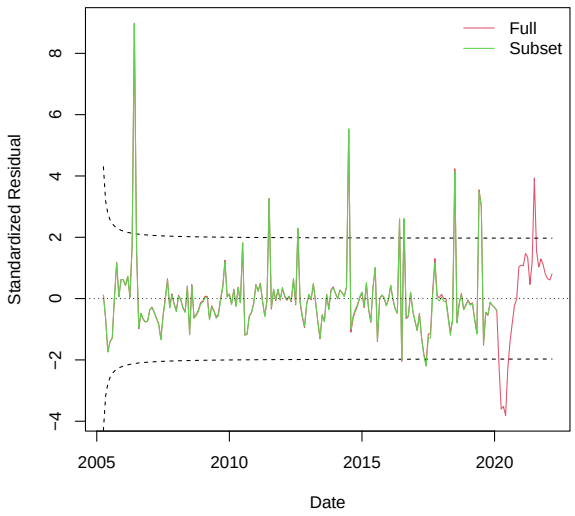

Figure 28: Standardized Residual Plot

The dashed lines correspond to 95% intervals using the Student-t distribution. Given the amount of data, we expect to see about 11 data points exceed those thresholds and we actually observe 13. We don't see residual drift, changing volatility, or other alarming patterns.

Overall this is suggests the model assumptions are reasonable. Notice that the majority of extreme residuals are either world-cup events or pandemic-related. This makes sense given that these are one-step ahead forecasts, not smoothed distributions.

1.7 Conclusions

In this problem we have analyzed soccer search trend data using a DLM. The model contained a second order trend component, a regression-based world cup effect, and a 12 period seasonal effects model. Through optimization we picked discount factors and which harmonics to keep from the seasonal component. We did the model fit under the full data as well as a subset occurring before the pandemic. We found that the model fit well under both data scenarios according to residual analysis. Also, the filtering fits were very similar between data cases, but the smoothing estimates were not.

Our results surrounding the application are interesting. We revealed that national soccer interest, as measured by search index, is relatively predictable within a strong annual pattern with little actual growth. When the pandemic hit, interest fell, but has risen back to normal levels.

We also found that the world-cup has an effect, although this was very difficult to estimate due to the small number of events in the data. The model overfit this effect, however it did remove the event's influence from the other trend parameters which was the goal. Some alternatives may have been better to consider such as providing a stronger prior, forcing a static discount factor, including it as part of an intervention regiment, or even just ignoring it and letting the trend terms pick up the added variability.

2. Conditionally Gaussian DLM Sampling

Consider the following model:

$$
\begin{aligned}
y_t &= x_t + \epsilon_t, \quad \epsilon_t \sim N(0, 9^{\gamma_t}) \\
x_t &= \phi_1 x_{t-1} + \phi_2 x_{t-2} + \nu_t, \quad \nu_t \sim N(0, \omega) \\
\omega &\sim IG(a, b) \\
\boldsymbol{\phi} &= (\phi_1, \phi_2) \sim N(\mathbf{0}, 0.25\boldsymbol{I}_2) \\
p(\gamma_t = 1) &= 0.2 \\
(x_0, x_{-1}) &\sim N(\mathbf{0}, \boldsymbol{I}_2)
\end{aligned}
$$

Notice that x_t follows a standard AR(2) process and that y_t introduces mixture Gaussian noise. Our goal in this problem is to develop a sampling scheme to fit this model to the data displayed below:

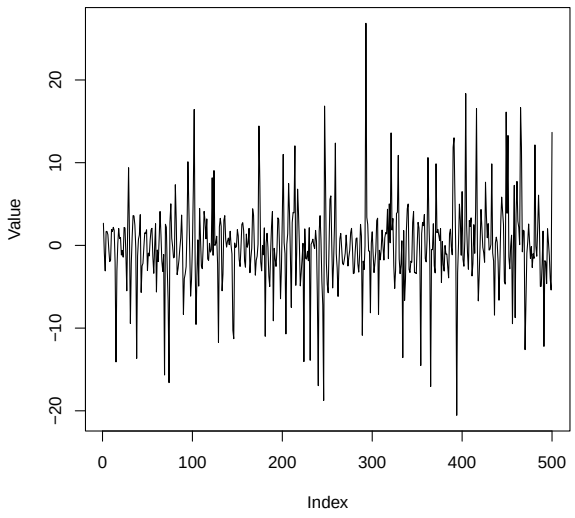

Figure 29: Simulated Data

Overall, our strategy is to appeal to Gibbs sampling. More specifically, our goal is to simulate from the joint posterior

$$p(\omega, \boldsymbol{\phi}, \gamma_{1:T}, x_{-1:T} | y_{1:T})$$

where $T = 500$ is the number of available data points. We do so by iteratively sampling from the full conditional distributions.

2.1 Full Conditional Distributions

In this section we derive each of the complete conditional distributions and give a corresponding sampling method. Sampling from each iteratively forms the basis for Gibbs sampling from the full joint posterior. For brevity, let $[\theta|\mu, -]$ arbitrarily denote the full conditional of some parameter θ, with emphasis placed on the μ conditioning.

2.1.1 ω Full Conditional

The AR(2) level variance ω has conjugate form:

$$
\begin{aligned}
[\omega|-] &\propto p(\omega) \prod_{t=1}^{T} p(x_t|\omega, x_{t-1}, x_{t-2}, \boldsymbol{\phi}) \\
&\propto IG(\omega|a, b) \prod_{t=1}^{T} N(x_t|\boldsymbol{x}'_{t-1}\boldsymbol{\phi}, \omega) \\
&\propto \omega^{-a-1} e^{b/\omega} \omega^{-T/2} \exp\left(\frac{1}{2\omega} \sum_{t=1}^{T} (x_t - \boldsymbol{x}'_{t-1}\boldsymbol{\phi})^2 \right) \\
&\propto \omega^{\frac{-T}{2}-a-1} \exp\left(\frac{1}{\omega} \left(b + \frac{1}{2} \sum_{t=1}^{T} (x_t - \boldsymbol{x}'_{t-1}\boldsymbol{\phi})^2 \right) \right) \\
&\propto \omega^{-a^*-1} e^{b^*/\omega} \\
&= IG(\omega|a^*, b^*)
\end{aligned}
$$

where $\boldsymbol{x}_{t-1} = (x_{t-1}, x_{t-2})'$ is notation used hereafter. To sample from this full conditional, simply generate an inverse gamma value with parameters a^* and b^* as above. For our data analysis, we use $a = b = 1$.

2.1.2 γ_t Full Conditional

The latent variable $\gamma_t \in \{0, 1\}$ determines which variance a particular y_t is generated with. The full conditional is given in terms of discrete probabilities:

$$
\begin{aligned}
p(\gamma_t = 0|-) &\propto p(\gamma_t = 0) p(y_t|\gamma_t = 0, x_t) \\
&\propto 0.8 N(y_t|x_t, 1) \\
p(\gamma_t = 1|-) &\propto p(\gamma_t = 1) p(y_t|\gamma_t = 1, x_t) \\
&\propto 0.2 N(y_t|x_t, 9)
\end{aligned}
$$

To sample from the complete conditional, simply generate from $\{0, 1\}$ with weights given above. If probabilities are desired instead of unnormalized weights, just divide each value by their sum.

2.1.3 $\boldsymbol{\phi}$ Full Conditional

The AR(2) level parameters determine some of the autocorrelation structure within the data, and they also have a conjugate form:

$$
\begin{aligned}
[\boldsymbol{\phi}|-] &\propto p(\boldsymbol{\phi}) \prod_{t=1}^{T} p(x_t|\omega, x_{t-1}, x_{t-2}, \boldsymbol{\phi}) \\
&\propto N(\boldsymbol{\phi}|\mathbf{0}, 0.25\boldsymbol{I}_2) \prod_{t=1}^{T} N(x_t|\boldsymbol{x}'_{t-1}\boldsymbol{\phi}, \omega) \\
&\propto \exp\left(\frac{1}{0.25}\boldsymbol{\phi}'\boldsymbol{\phi} + \frac{1}{\omega}\sum_{t=1}^{T}(x_t - \boldsymbol{x}'_{t-1}\boldsymbol{\phi})^2\right) \\
&\propto \exp\left(4\boldsymbol{\phi}'\boldsymbol{\phi} + \frac{1}{\omega}(x_{1:T} - X\boldsymbol{\phi})'(x_{1:T} - X\boldsymbol{\phi})\right) \\
&\propto \exp\left(\boldsymbol{\phi}'(4\boldsymbol{I}_2 +' X'X)\boldsymbol{\phi} - 2x'_{1:T}X\boldsymbol{\phi}\right) \\
&= N(VX'x_{1:T}, V)
\end{aligned}
$$

where $X = (x_{0:t-1}, x_{-1:t-2})$ is $T \times 2$ dimensional and $V = (4\boldsymbol{I}_2 + X'X)^{-1}$ is the covariance matrix. Thus the full conditional for $\boldsymbol{\phi}$ can be sampled from the above bivariate normal distribution.

2.1.4 $x_{-1:T}$ Full Conditional

Conditionally sampling the AR(2) level observations $x_{-1:T}$ is not as straight forward as the other parameters. We rely on the Forward-Filtering-Backward-Sampling algorithm. To begin, we rewrite the model in DLM form:

$$
\begin{aligned}
y_t &= (1, 0)\boldsymbol{x}_t + \epsilon_t \\
\boldsymbol{x}_t &= \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} \boldsymbol{x}_{t-1} + \boldsymbol{\nu}_t
\end{aligned}
$$

where it is obvious that $F' = (1, 0)$ and $G = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix}$ with state vector $\theta_t = \boldsymbol{x}_t = (x_t, x_{t-1})'$ and noise vector $\boldsymbol{\nu}_t = (\nu_t, 0)'$. With this established, the FFBS algorithm proceeds as follows:

1. Use DLM filtering equations to obtain the standard results for $t = 1 : T$.
2. Sample θ_T from $N(\boldsymbol{m}_T, C_T)$.
3. For $t = (T-1) : -1$, sample θ_t from $N(\boldsymbol{m}^*_t, C^*_t)$ where $\boldsymbol{m}^*_t = \boldsymbol{m}_t + B_t(\theta_{t+1} - a_{t+1})$ and $C^*_t = C_t - B_t R_{t+1} B'_t$.

Following this procedure will produce a single sample from the full conditional of interest. The Gibbs sampling element comes into play by changing the DLM parameters G, ω, and $v_t = 9^{\gamma_t}$.

Note that there is some redundancy in $\theta_t = \boldsymbol{x}_t$. However, structurally the second term in $\boldsymbol{x}_t = (x_t, x_{t-1})$ is unchanged during the update. Thus we only keep the first element x_t, waiting to sample x_{t-1} until the next recursion.

2.2 Results

Here we summarize the results from running the proposed algorithm on the provided data. We initialize the chain at reasonable starting values and run the Gibbs algorithm for 3000 iterations plus an additional 300 as burnin.

Before exploring numerical or inferential results, we briefly take a look at convergence. The plot below shows $\boldsymbol{\phi}$ and ω plotted together, shifted for visibility. Other parameters are omitted for space, but are similarly convergent.

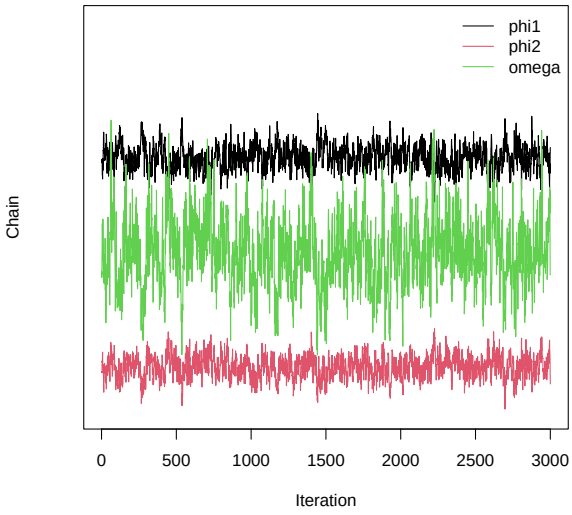

Figure 30: Traceplots Showing Convergence of Parameters

The traces appear stable. Some thinning or additional samples could be helpful, but these will be sufficient for our purposes.

The posterior distributions of the AR(2) parameters ω and $\boldsymbol{\phi}$ are summarized in the table and kernel density plots below.

Table 3: Posterior Summary of AR(2) Parameters

Param.	Est.	95% Low	95%High
ω	9.65	7.81	13.44
ϕ_1	0.44	0.34	0.54
ϕ_2	-0.41	-0.50	-0.32

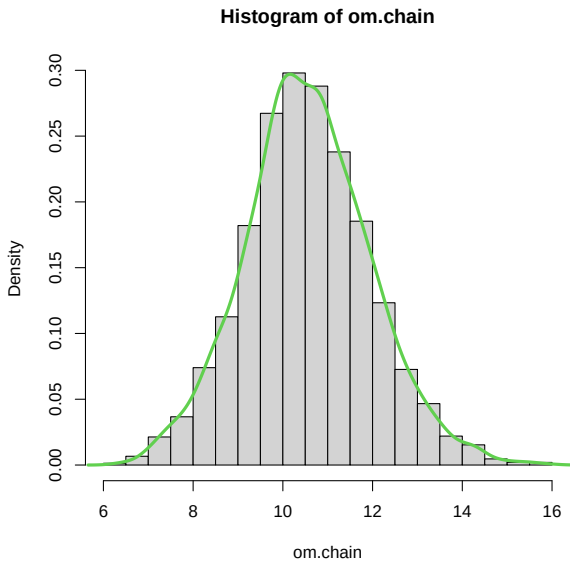

Figure 31: ω Posterior Samples

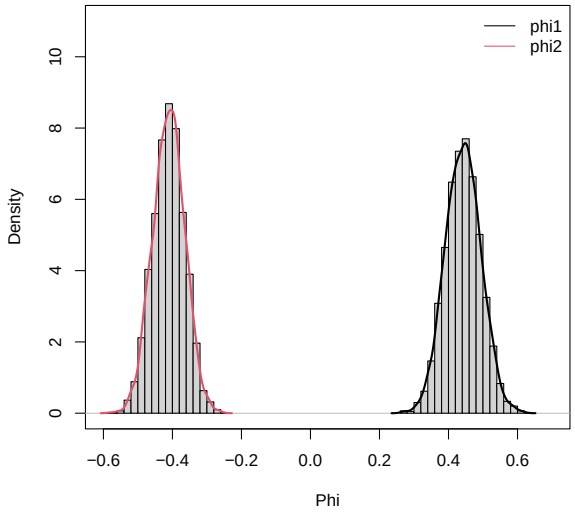

Figure 32: $\boldsymbol{\phi}$ Posterior Samples

Recall that several properties of an AR(p) model, including stationarity and quasi-periodicity, depend on the characteristic polynomial roots. Figure 33 below shows the posterior distribution of the reciprocal root modulus (recall they come in conjugate pairs).

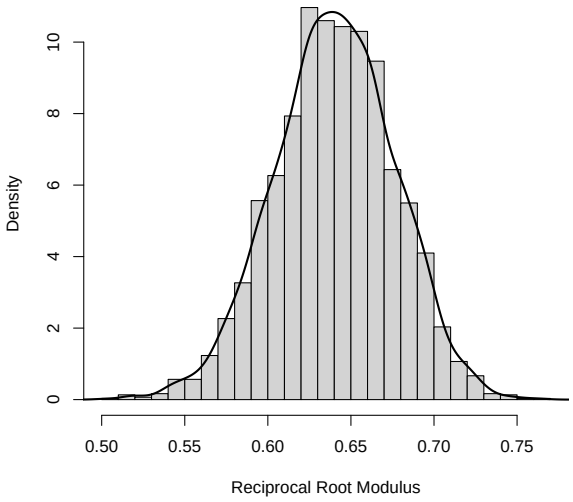

Figure 33: Posterior Reciprocal Root Modulus

The main finding here is that the reciprocal root modulus is well below 1, indicating stability. Furthermore, the roots were complex in 100% of posterior samples, suggesting quasi-periodicity in the underlying AR(2) process, which we return to later.

Now we turn attention to the latent indicator variables $\gamma_{1:T}$. The histogram in Figure 34 shows the distribution of posterior probabilities that $\gamma_t = 1$. The shape suggests the data are somewhat reasonably separated. In particular, extreme values are easy to label $\gamma_t = 1$, but there is considerable gray area. The beauty of the Bayesian model is that other estimates do not depend on making specific decisions about the value of γ_t - the uncertainty is incorporated.

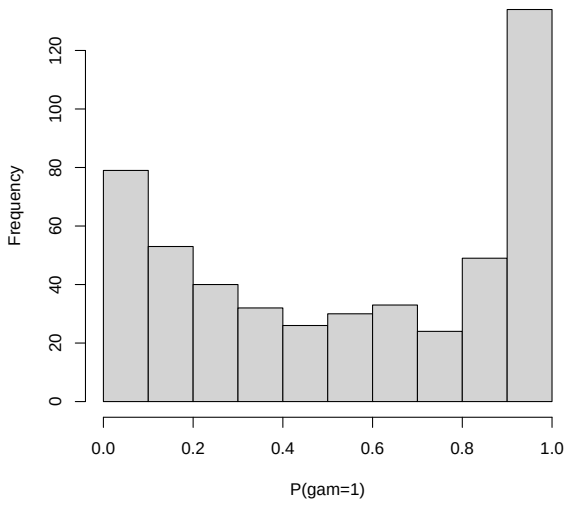

Figure 34: Posterior Prob. of $\gamma_t = 1$

Next we look at $x_{1:T}$, the underlying AR(2) series. The below plot shows the posterior estimated sequence on top of the observed data. Furthermore, the points with $> 99\%$ posterior probability of having $\gamma_t = 1$ are highlighted.

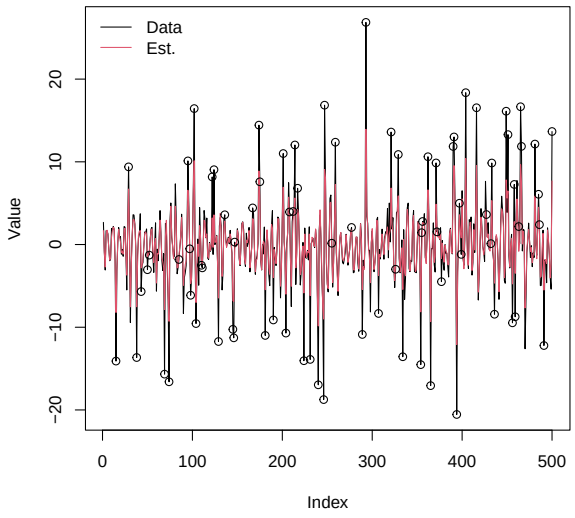

Figure 35: Underlying AR(2) Estimated Series with Outlier Points Highlighted

Overall, the series appears to be a variance-reduced version of the data. Also reassuring is the outlier classified points are predominantly large-discrepancy values. Figure 36 shows a subset of the data, index values 250 to 350, which allows for a cleaner presentation of error bounds and how the outlier classification relates.

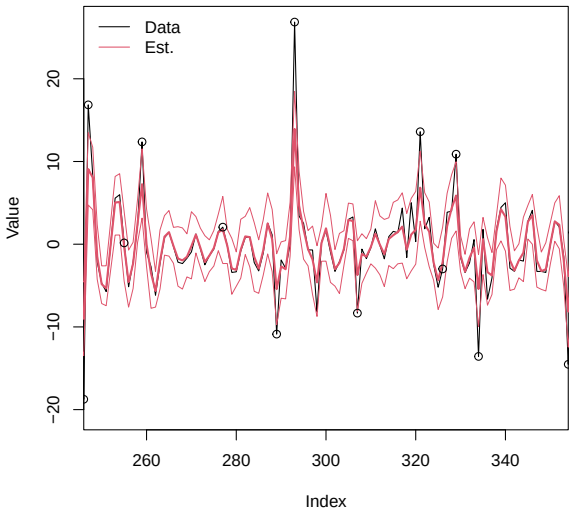

Figure 36: AR(2) Estimate and Error Bounds

To verify that our "filtered" series is indeed AR(2), we consider the posterior estimated ACF and PACF and compare to those of the original data. The ACF plots are displayed below.

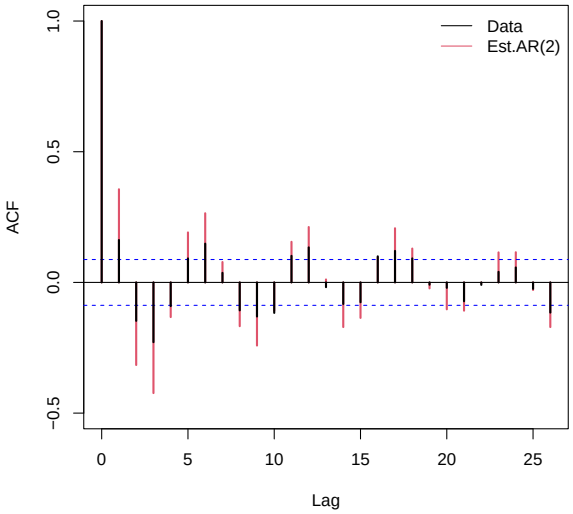

Figure 37: ACF Plots of Data and Est. AR(2)

Notice that the AR(2) estimate inflates the correlations, as is expected because we have reduced marginal variance. Figure 38 displays the corresponding PACF plots.

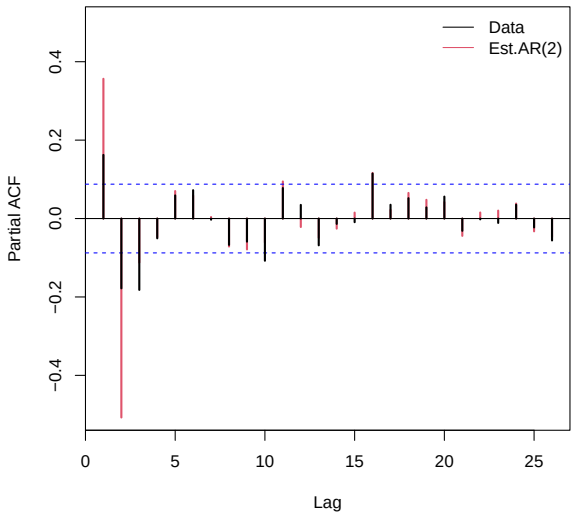

Figure 38: PACF Plots of Data and Est. AR(2)

The PACF plots also suggests the consistency with out estimated series with an AR(2). The estimated PACF plots show two very strong partial correlations, compared to the original series which is less clear.

Just some closing thoughts: I find it interesting that the posterior estimate for ω is about 9, which suggests that noise in the data is mainly driven by the AR(2) process with outliers contributing extra variability at times. I suspect that different choices for the observational variance would yield different results in this regard.

Of course, at the end of the day this was a toy example with simulated data, but it reinforces a theme running under a lot of time-series analysis: unexpected depth. In this case, a simple generalization to a two component mixture led to a useful outlier identification and removal tool. Other examples include discount factors, linear trend models, the super-position principle, and even basic AR(2) theory. Although these may appear simple at first, they prove more nuanced and useful than an initial glance might suggest.